Does language matter? An exploration of differences between high- and low-scoring REF2014 impact case studies

A thesis submitted to the

School of English Literature, Language and Linguistics

Newcastle University

For the degree of Doctor of Philosophy

Bella Ruth Reichard

May 2025

Abstract

Since 2014, UK universities have been required to submit impact case studies (ICS) for assessing research impact beyond the contribution to academic knowledge as part of the national Research Excellence Framework. Using ICS narratives for assessment has been criticised as allowing presentation and style to affect the integrity of the process by taking precedence over the substance of impact claims (e.g. Watermeyer 2019), but such criticisms are not linked to systematic linguistic analysis. This thesis compares a corpus of high- and low-scoring REF2014 ICS (n=217) in a register analysis with the following components:

1. A readability analysis using Coh-Metrix (McNamara *et al.* 2014), showing that high-scoring ICS were marginally easier to read and scored stronger on making explicit causal links;
2. A thematic analysis (Auerbach and Silverstein 2003) to identify common types of content across ICS, showing that low-scoring ICS were more focused on pathways to impact compared to high-scoring ICS which emphasised retrospective impacts;
3. A lexical analysis of n-grams that were compared across sub-corpora using keyword measures (Brezina 2018), showing that the most noticeable differences between high- and low-scoring ICS were tied to specific impact content (where phrasing was fixed), meaning these wordings could not simply be reused in different contexts to depict “outstanding” rather than “modest” impacts;
4. An analysis of evaluative language using the Graduation subsystem of the Appraisal framework (Martin and White 2005), in which for most measures, no significant difference could be detected between high- and low-scoring ICS.

Taken together, these analyses show some language differences between high- and low-scoring ICS, but considering the scale of possible differences in each of the analyses, the statistically significant differences between high- and low-scoring ICS are not enough to assume that language choices had an undue influence on scores to the detriment of substantial impact claims.

To Uli

the love of my life

Acknowledgements

The first mention about this whole journey needs to be Uli – this is all your fault for believing in me.

With that said, I can now write acknowledgements in a more traditional order, thanking the supervisors who were part of the endeavour over the 8+ years:

- Karen Corrigan for taking on this project and for suggesting the topic (me: "What's an impact case study?"), during our first conversation 9 years ago.
- Adam Mearns for providing exactly the right balance of guidance, calm and encouragement, detailed feedback, and frequent confirmations that "this sounds like a sensible option".
- Mark Reed for providing mentorship as well as supervision, helping to keep me calm, treating me as a colleague from the beginning and developing me and my career.
- Andrea Whittle for challenging me from start to finish and for the view from outside the discipline and topic.

Parts of the research are published in "the article" (Reichard *et al.* 2020), which contains the full findings from the thematic analysis included as part of this thesis. This thematic analysis was a collaboration with my supervisor Mark Reed and other colleagues: Ged Hall, Alisha Peart, Lucy Jowett and Jenn Chubb, plus Andrea Whittle for the publication stage – thank you all! In this thesis, the collaborative part is covered in five pages of the methodology section of the thesis (section 4.3.3), where I describe my contribution in relation to that of the other collaborators, and eight pages of the first results and discussion chapter (sections 5.2.2 and 5.3.2), with reference made to the Reichard *et al.* (2020) article throughout the thesis where relevant.

I would like to thank my employers during my studies: first, INTO Newcastle University for contributing to funding the tuition fees and for various bits of research leave, and then Rachel Kendal for providing flexibility and encouragement while I worked on the CES Transformation Fund at Durham University.

For most of the time, I was part of a Linguistics writing group. I made it into a second generation of writing group. Thanks to everyone who provided some sort of companionship over the years. A special mention for Damar – thank you for your ears, your honesty and your calm.

This research has enabled me to create my own job. Thanks to all my clients who have put trust in me and my expertise. I would specifically like to mention Jackie Reynolds, Emma Bond and Louise Rutt, who all played different but crucial roles in developing my career and my business.

With this project spanning so many years, so much has happened outside of it, and I would like to thank people who helped me through that time in many different ways:

- Ed for providing support from the university side at relevant points ("I would like you to have this whiteboard eraser") and for saving my marriage at least once.
- Friends who were there for the ride: Louise, Rosi, Elaine and Gosia who supported me throughout and in some fairly direct ways, and many others who walked part of the way with me.
- Joyce for the various points of coaching and especially for permission to prioritise my thesis – you lived above my desk in the shape of the post-it "Get the PhD DONE".
- Uli for being there all the time. I decidedly do not want to remember all the many, many hours you spent on this – listening to me talk on our walks, suggesting ways to frame things, acting as a second coder during the pilot phase of the Appraisal analysis, and all the time you held me while I was crying, panicking, or failing at other things in life for which I blamed the PhD. Thank you for your patience and, as they say, "unwavering support" with this project and with life.
- Edward for being my writing group at home and safeguarding my phone at relevant points. I will reciprocate when it comes to your exams.
- David for all the hugs and for being lovely amidst all the challenges. You didn't have to spend your pocket money on chocolate for me, but I appreciated it.
- Raf for the co-working sessions to get the final draft together, and for teaching me more about myself, relationships and the wider world. And for feeding me. I love having you in my life, and thank you for making space for me in yours.
- Uli again, because I love you <3

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Definitions

GPA – Grade Point Average

HE – Higher Education

HEFCE - Higher Education Funding Council for England

ICS – Impact Case Study

MP – Main Panel

REF – Research Excellence Framework

UoA – REF Unit of Assessment

The numbering of Units of Assessment refers to that used in REF2014 (not the numbering used in REF2021 or REF2029).

"Submission" in this thesis is defined as all impact case studies that were submitted by a given university in a given unit of assessment, e.g. all submitted by Newcastle University in UoA29.

References to specific impact case studies have the following format, using a short title defined during text preparation based on the official title: UoA29 Newcastle *Poetry.*

An index of impact case studies included in the sample, including the short titles, is provided in Appendix A.

Chapter 1 Introduction

One of the core purposes of universities and academic activity is to generate, or enable the generation of, new knowledge. Asking "why?" is one of the most fundamental parts of this – for example, why does something work in this way? Why did previous generations build a settlement here? And myriad other questions. Perhaps the most overarching question is "why generate new knowledge in the first place", and one possible answer, among others, is: to create an evidence base for change, for addressing societal challenges, for "making a difference". This difference that academic research can make in society is called "research impact", and it is being encouraged in various ways in different university systems.[1]

In the UK, one way in which the pursuit of research impact is incentivised is through a Research Excellence Framework (REF). With this framework, the quality of research in UK higher education institutions is assessed every 5-8 years in order to justify public expenditure on research and universities in general, and differential expenditure for different universities and research disciplines in particular. Since the 2014 iteration, part of this assessment pertains to research impact in order to measure, and enhance, the "return on investment" (Smith *et al.* 2020: 13) of public money. In this exercise, research impact is defined as "an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia" (HEFCE 2011: 26). It is assessed through the peer review of impact case studies (ICS), which are the focus of analysis in this thesis.

1.1 Research background

Some parts of the university system, often in applied disciplines, have welcomed the development of an "impact agenda" because it validates the kind of outreach, engagement and impact work that was already part of the research lifecycle for some projects, and a strong motivator for some researchers. For example, Weinstein et al. (2019: 93) 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

[1] For a discussion of definitions of research impact, see Bayley (2023: 9-13).

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).

At the same time, concerns are often voiced that this increased focus on the societal value of research beyond the generation of new knowledge may indirectly stifle curiosity-driven research (e.g. Smith *et al.* 2020) or critical research (Machen 2020) and therefore be a danger to scientific progress. Watermeyer (2019) fears for the role of academics as credible intellectuals if they turn into instrumentalised solution-providers within a neoliberal landscape.

In addition, the method of assessing research impact contributes to concerns. In the UK REF, research outputs and impact are peer reviewed at disciplinary level in "Units of Assessment" (36 in 2014, 34 in the 2021 and 2029 assessments). Specifically, impact is assessed through narrative case studies that describe the effects of academic research in a short template and are reviewed by both academic peers and expert research users. This peer review approach has been heavily criticised, partly because crafting impact case studies is resource-intensive for universities and partly because it appears to carry a danger of the language and presentation of a case study influencing the assessment outcome. It is this latter concern that is the focus of this thesis.

There are therefore recurring discussions whether narratives are needed and whether the whole assessment process could be simplified by replacing impact case studies with metrics (as described in e.g. Curry *et al.* 2022; Khazragui and Hudson 2014; Smith *et al.* 2020). Various options were explored before designing REF2014, the exercise in which impact was assessed for the first time. The approach to research assessment was revisited between REF 2014 and REF2021, and again between REF2021 and REF2029. In the report of the Future Research Assessment Programme following REF2021, pointedly titled *Harnessing the Metric Tide*, the Chair of its international advisory board, Sir Peter Gluckmann, summarises that "the role and place of research assessment and metrics remains debated and can generate strong opinions" (Curry *et al.* 2022: 11).

Critics of an all-metric approach point to "the richness of the current case study format", which "available indicators or infrastructure cannot approximate" (Curry *et al.* 2022: 7). Specifically, an assessment focused on metrics could disadvantage novel, unanticipated impacts for which no metrics were defined previously, as well as those that cannot be measured quantitatively. Beyond negative effects on the assessment process, a focus on

generating measurable impacts can be a danger to creating meaningful impact that may be less straightforward to report quantitatively (Penfield *et al.* 2014: 29). Furthermore, a metrics-based approach which does not allow for impact claims to be contextualised may exacerbate the potential negative effects of any impact agenda by streamlining "what counts" and therefore contributing further to a restriction of academic freedom.

On the other hand, a central critique brought against the narrative approach is that the assessment cannot be objective if it is based on text, rather than on standardised numbers. Several sources have implied that the presentation of impact case studies may have influenced the score that a case study received (e.g. Chowdhury *et al.* 2016; McKenna 2021; Penfield *et al.* 2014; Watermeyer and Hedgecoe 2016). This assertion assumes that the language and other presentational elements can introduce factors other than content into the assessment, and those universities that use these tools better for persuasion may get higher scores than their actual impacts deserve. From a linguistic perspective, Hyland and Jiang's (2023: 16) study on language used for "hyping" impacts also emphasises the "risks of narrative self-reports in undermining the impartial assessment of research impact". Smith *et al.* (2020: 38) present it as "likely that case studies may perform better than others simply because they are written more persuasively". Similarly, regarding another component of REF (namely, "environment statements"), Thorpe *et al.* (2018: 54) assert that "the submissions were exercises in persuasion directed at assessors". For impact case studies (ICS), perhaps the strongest phrasing of this assertion can be found in Watermeyer (2019: 80-81):

> "Panellists' role in evaluating impact would … appear to have focused more on an assessment of impact *narrative* than an assessment of impact *claims.* In other words, the stylistic rather than substantive achievements of REF2014 ICS assumed precedence. Concordantly, where the narrative flair or 'stylistic virtuosity' of ICS was privileged, panellists were involved in a qualitatively different sort of evaluation that entailed not so much the application of robust criteria in determining a scale of impact 'excellence' as working out the 'truth' of the story being *spun*." (Italics in original)

Indeed, ICS are sometimes treated as a persuasive genre (e.g. Gow and Redwood 2020; Wróblewska 2021). This in itself is not inherently problematic, but it could be if there were a measurable difference in features of persuasion between those texts that received the top score and those that did not achieve a high rating. In other words, regardless of the nature of ICS as inherently a genre of persuasion, if presentation had indeed affected scores, then it should be possible to measure differences in presentation, language and persuasion between high- and low-scoring ICS. It has to be noted, though, that even if such a difference

can be detected, this does not necessarily indicate a causal relationship between a persuasive presentation and the assessment outcome.

Moreover, there may be valid reasons for a difference in language between high- and low-scoring ICS. One central possibility is that lexical differences are related to the kind of content that is reported in high- or low-scoring ICS respectively, and therefore reflective of impacts that were valued more or less, rather than indicating undue language influence. A distinction therefore should be made between lexical differences that a writer or editor of an ICS may have control over, and which could potentially contribute to the "stylistic virtuosity" that Watermeyer (2019: 80) alludes to, and those differences that are determined or at least restricted by the content of an ICS.

In the literature, the observations of presentational differences are usually subjective (e.g. Gow and Redwood 2020), based on self-reported reading experience (e.g. Watermeyer 2019) or framed as advice (e.g. McKenna 2021), rather than being grounded in empirical analysis of the texts themselves. Where empirical analysis of language was conducted, this was not linked to scores (Hyland and Jiang 2023). These contributions will be discussed in more detail at various points in the thesis, but overall this shows that there is a gap in research of the language and presentation of impact case studies in conjunction with assessment scores.

1.2 Research questions

In order to provide empirical data for addressing the question whether narrative assessment should be supported or whether it should be discouraged on the basis of presentation as a potential confounding factor, this thesis asks:

To what extent could presentation have influenced the scoring of impact case studies in REF2014?

This overarching research question is addressed by considering the following two sub-questions:

1. What features related to the *presentation* of the research, pathway and impact, as opposed to the stated criteria of *significance* and *reach* of the impact and the clarity of *attribution* to the research, may be characteristic of high- or low-scoring ICS and therefore may have influenced the score? This question is considered in respect of the following areas:

a. Context and reading experience (situation of writing and reading ICS; cohesion, reading ease and speed) *(chapter 5)*
b. Balance and emphasis of content (balance of material in selected Sections 1; thematic analysis of complete ICS) *(chapter 5)*
c. Words or phrases (themes; content-driven or editorial choice) (*chapter 6)*

2. What linguistic markers of persuasion and evaluation do ICS feature, and does this differ between high-scoring and low-scoring ICS? *(chapter 7)*

Using a combination of quantitative and qualitative methods, I aim to provide evidence towards the evaluation of the reliability, or lack thereof, of narratives as an assessment vehicle for research impact.

1.3 Research significance

A key function of language, particularly relevant for this study, is that it is used for persuasion (that is, to change or influence the other person's point of view) and to show evaluation, especially in an assessment context. However, it is difficult to measure persuasive and evaluative language, because persuasion and evaluation are very context-specific and can sometimes only be noticed by those familiar with the genre or topic (see more detailed discussion in section 3.2). Therefore, linguistic functions in general, and persuasion and evaluation in particular, should be examined with various complementary linguistic methods, to elicit a range of perspectives that are likely to shed light on the function that language plays in these contexts. This study uses a variety of linguistic methods to investigate the functional role of language in the assessment context of REF impact case studies. This provides empirical evidence in respect to the ongoing debate about the role of language in the scoring of ICS, in particular the question of whether the mastery of persuasive linguistic functions can be used to improve ICS ratings. The study therefore also contributes to the ongoing reflection about whether the format of case studies is an appropriate way to assess research impact and ultimately allocate research funding.

1.3.1 Theoretical grounding

This thesis views language as social semiotic. This means that linguistic meaning is constructed between those participating in a linguistic exchange, which is situated in their respective social context, as well as the relationship between those participating, which can involve differences in power and social status (Halliday 1978). I am applying a constructionist

epistemology to impact case studies, which means that I am assuming an interplay of subject and object.

Impact case studies are prepared by groups of people ("writers") with assumptions about how they may be received and interpreted, and they are read by groups of people ("readers") who may have little prior experience of this genre compared to other genres they are asked to base their assessment on in the same exercise (i.e. research outputs typical for their discipline). I refer to these groups as "writers" and "readers" for reading ease, although the social construct is more complex. Writers and readers are not homogeneous groups, as those involved in producing and those tasked with assessing ICS can both have a variety of backgrounds, including researchers, professional service staff and research users. The "writer" is likely to have been an entire team at various stages in the development of the case study, rather than a single person. The "readers", that is, the reviewers were not acting as individuals either, but worked in a team, with calibration and discussions of judgements and ratings. As a result, their view is determined by a social process including various individuals, teams and institutions. In addition, roles often intersect, as those acting as reviewers may also have authored an ICS about their own work. More detail on the situation of preparing ("writing") and assessing ("reading") ICS is provided in sections 5.2.1 and 2.1.3 respectively.

Readers and writers of ICS cannot interact directly with each other to co-construct the reality of the text, given the sequentiality of the assessment process (where writers and readers are separated in time and space), and the meaning that a reader sees in a specific use of language in an ICS may not be the meaning intended by the writer. The reality of the assessment is therefore constructed by readers and writers of ICS who do not interact with each other directly, may have very different backgrounds, and are embedded in different circumstances. This leads to interpretivism as the theoretical perspective of this thesis. While placing importance on reliability, replicability and generalizability in the selection of the sample and approaches to analysis, I reflexively acknowledge that I can only take one perspective and that this is influenced by my own place in the system as a research impact consultant (see section 4.2.3). I interrogate the texts with quantitative methods to create an empirical basis for comparison, but I accept that this does not fully reflect the various subjective and intersubjective realities of others, including the varying perspectives of writers and readers of impact case studies.

The perspectives of constructionism and interpretivism acknowledge that there are several possible realities. So does the perspective of language as a social semiotic, which sees language as both meaningful and context-dependent. In all these perspectives, meaning does not exist outside of people making meaning, in the case of ICS especially through written language. The mix of quantitative and qualitative methods utilised in this study is designed to address the constructionist and interpretivist nature of the subject.

The two main analytical frameworks used are Register Studies and the Appraisal framework of evaluation, which is rooted in Systemic Functional Linguistics. Corpus linguistics, register studies and Appraisal are all compatible with the view of language as social semiotic. Tognini-Bonelli (2004: 18) explains that a corpus shows examples of "social practice", and that corpus linguistic methods can be used to investigate meaning that is shaped by its social context (2004: 19). Register studies assume that language is varied based on the situation of use (including, but not limited to, the participants), and that the language variation that it concerns itself with is functional (Gray and Egbert 2019), that is, it creates meaning within its situational context. Martin and White (2005: 27) also point out that the concept of register is "a connotative semiotic realised through language". Specifically, they state that the Appraisal framework is mostly concerned with the interpersonal aspect of language and assumes language as social semiotic.

With the view of language as social semiotic, that is, as something that is used to encode meaning beyond the factual content of the text, this study contributes to evaluating the reliability, and therefore the value, of using narratives of impact for assessment. Are assessors likely to prioritise this social semiotic at the expense of content? At this point, it is important to note that the claims made in this thesis are not causal, that is, there is no assumption that any differences identified between corpora must inevitably have influenced the ICS score. Rather, such differences may be a result of some universities having the means to produce both research with significant and far-reaching impact *and* polished, standardised impact case studies – that is, potentially any correlation may be due to the same cause. However, a causal link between linguistic features and scoring is assumed by some participants in the debate around ICS (see chapter 2). I aim to provide empirical evidence that is relevant to evaluating such claims.

1.3.2 Original contribution

Whilst the UK was the first country to introduce formal research impact assessment linked to university funding, other countries and funders are watching closely and some have implemented similar policies and evaluation systems, for example Australia, Hong Kong, the United States, The Netherlands, Sweden, Italy, Spain and others (Reed *et al.* 2021). The implications of decisions about the nature of research impact assessment therefore reach far beyond the UK.

In this context, as indicated in section 1.1 and discussed more fully in section 2.1.2, a key debate is about how best to assess research impact. The use of ICS, which has become the preferred method for the assessment of impact in the UK and beyond, has been criticised for relying too much on the ICS writer's ability to "sell" the impact described, suggesting that the language used in ICS may have a significant influence on the score and hence research funding.

This study advances this debate in two main ways. Firstly, I use a range of methods to conduct linguistic analyses that are linked to assessment scores, which can serve as a basis for examining claims about the influence of language differences on the assessment. Previous studies of the language in ICS have not taken scores into account (see section 4.1.2). In this way, this thesis provides empirical evidence that can be used to assess claims about the influence of language on assessment scores that have previously been based only on intuition and experience. Moreover, my study considers the full context and range of possible differences between high- and low-scoring ICS in the Readability, Lexical and Appraisal analyses (introduced later in this section). That is, I discuss the detected differences in the context of *possible* differences, thus also taking into account similarities between high- and low-scoring ICS, rather than taking differences as the norm.

Secondly, I introduce a framework for assessing the extent to which a correlation between linguistic differences in high- and low-scoring ICS could be considered an indicator of undue influence of language on the scoring, and, by contrast, in which cases the linguistic differences are likely to be a result of the difference in the nature or quality of impact. If more persuasive language can be identified in high-scoring ICS, then the social semiotic nature of language may have played a larger role than desired in the assessment. If the differences are mostly thematic, then language differences appear to be based more on content, which is what the assessors are asked to assess.

Overall, I conduct a register analysis of ICS, using different methods. These are based on a corpus of 217 ICS from REF2014 that includes all clearly identifiable 4* ICS (124 texts) and the clearly identifiable 1*/2* ICS in those UoAs where 4*s are available (93 texts). For some analyses, a subset of that corpus was used (see section 4.3 for details of the sample compositions). The analyses include:

- A situational analysis of the writing and reading context explaining the specific circumstances of ICS in contrast to research articles (Biber and Conrad 2019);
- A readability analysis using Coh-Metrix (McNamara *et al.* 2014);
- A thematic analysis (Auerbach and Silverstein 2003) to identify common types of content across ICS;
- A lexical analysis of n-grams that were compared across sub-corpora using keyword measures (Brezina 2018) in order to identify common themes and distinguish between content-led and editorial differences; and
- An analysis of evaluative language using the Graduation subsystem of the Appraisal framework (Martin and White 2005) to explore the use of covert evaluation and persuasion in ICS.

Applying these methods to the genre of ICS has not been attempted before and thus is in itself a contribution to the refinement of the methods, because this can show the implication of the underlying theory and assumptions as the method is applied to a new context. In addition, the final two methods required significant adaptation to be relevant and effective in respect to ICS. For the lexical analysis of n-grams I have applied keyness measures to n-grams, instead of key words (see section 6.1.2). For the Appraisal analysis, I have designed a Graduation coding scheme relevant to ICS based on existing Graduation coding schemes (see section 7.1.2).

1.4 Outline of thesis

The thesis is structured as follows:

Chapter 2 (Assessing Research Impact) describes the situation and controversies around research impact assessment in more detail, drawing on the literature on research impact. First, I outline the political background of introducing research assessment in the UK and the move towards including research impact in the national assessment framework. I describe the debate around different ways of assessing research impact and explain the role of narratives in the assessment. The chapter then summarises a selection of research on impact

case studies that were submitted to REF2014, followed by a closer look at the literature on the content, narrative framing and style in these texts.

Chapter 3 (Impact Case Studies as a Persuasive Register) outlines the linguistic frameworks and theories that underpin my research, namely register studies and the Appraisal framework. I discuss the ways in which ICS can be viewed as an academic register and how it relates to other academic registers and disciplinary differences in those. Based on this overview, I frame the research questions in relation to the different components that constitute the register analysis presented in this thesis. I then move on to various ways of conceptualising persuasion and evaluation in linguistic inquiry, before describing the Appraisal framework in more detail as the system chosen to investigate evaluative language in ICS.

In chapter 4 (Research Methods), I provide a brief overview of previous research on ICS that used relevant methods and then explain my own overarching research approach, followed by a detailed description of the text sample. Different research questions required the use of different methods, and the exact sample used for each of these methods was adapted in various ways depending on the goal of each analysis. In addition to the descriptions of the samples, this research methods chapter includes descriptions of methods of analysis where these are applicable to more than one of the subsequent chapters. Analyses are generally described in the relevant chapter before the findings are presented.

The following three chapters contain further details on the analytical methods applied, the findings that these led to, and discussions of these findings.

Chapter 5 (Context and Content of Impact Case Studies) addresses the parts of research question 1 about the context, writing and reading experience (1a) and about the balance of different types of content (1b). It first provides a comprehensive situational analysis of impact case studies, contrasting them with research articles. I then discuss the circumstances of writing ICS in more detail. This is followed by two analyses on reading ICS: findings of a qualitative analysis conducted with colleagues, reporting on first-hand experience as readers, and a quantitative analysis reporting on measures of readability and cohesion. The final part of this chapter turns to the content of ICS by first describing the types of material that can typically be found in Section 1 of ICS and then providing further findings from the qualitative analysis.

Chapter 6 (Lexical Investigation) focuses on lexical differences in line with research question 1c, first describing the methods of analysis in detail and then providing findings. In a bottom-up analysis of word sequences that appear with a statistically significant higher frequency in either high- or low-scoring ICS, I identify common themes and how they are distributed across high- and low-scoring ICS. I then identify which of these word sequences are available to use in any ICS and which are dependent on certain types of content being present, in order to investigate whether common phrasings that may be typical of high- or low-scoring ICS are predominantly those that writers have control over or not. The final analysis presented in chapter 6 investigates those sequences over which writers have control in light of research question 2, to determine for each sequence whether it could have persuasive meaning in the context of a REF ICS.

The final results chapter, chapter 7 (Evaluative Language: Appraisal), reports on a partial Appraisal analysis, again in order to address research question 2. As with chapter 6, details of the analysis process are followed by findings. In this analysis, I used manual coding to apply feature tags to words or phrases with evaluative meaning within the Graduation framework of the Appraisal system. The first set of results described arise from overall comparisons of various sub-corpora, that is, comparing high- and low-scoring ICS and comparing different disciplinary groupings. This is followed by results that arise from asking more specific questions about how language is used in ICS based on the impact literature and on the other analyses in this thesis, and which can be addressed by interrogating a corpus that is tagged for Appraisal features, rather than relying on searching for lexical expressions.

Chapter 8 (Conclusion) ties together the main points from chapters 5-7 and relates them to the research questions. It summarises the framework of content-driven versus editorial language differences that I develop throughout the thesis and interprets findings from the linguistic analyses in light of the situational analysis to highlight specific characteristics of the ICS register. The chapter then argues that based on all the analyses in this thesis, the differences in language and other presentational features are unlikely to have influenced the assessment process in any substantial way, strengthening the approach of using narratives rather than metrics-based assessment for research impact.

Chapter 2 Assessing Research Impact

To set the context for the thesis, this chapter provides the background to and justification for investigating the language of impact case studies more systematically. It begins with an overview of the political drivers behind the development of research assessment in the UK, followed by a discussion of different purposes that can drive such assessment (section 2.1.1). The stated purposes of research and research impact assessment in REF are "allocation" of research funding, and "accountability" of using public money for such funding in the first place (HEFCE 2011: 4). I argue that these purposes require fundamentally different assumptions about assessment design: Allocation requires a numerical value that can be used for a funding formula, and accountability for such funding would ideally encompass as comprehensive a basis for assessment as possible – powerful arguments exist against both purposes, and especially against the aspiration of an exhaustive basis for assessment.

I discuss various approaches to assessment, including the possible use of metrics or narratives for assessing research impact (section 2.1.2), before focusing on the peer review of narratives as the assessment format chosen for REF (section 2.1.3). The REF context part of the chapter ends with an overview of previous analyses of REF2014 impact case studies to illustrate the research landscape to which this thesis contributes (section 2.1.4). This is followed by a review of the literature more specifically on writing ICS, regarding content, narrative framing and style (section 2.2). Overall, this section shows that, while several publications imply that the presentation of ICS affects the assessment process, no other studies have specifically investigated the relationship between presentation and scores. The present study aims to address that gap.

This chapter summarises a wide range of sources on REF2014 ICS and the wider context of impact assessment in the UK. The starting point was a search in January 2021 of the terms "REF2014 impact", "research impact", "UK research impact", "REF2014 impact case studies", "writing impact case studies", "writing REF2014", "research excellence framework impact", "REF2014 impact case studies analysis" through Web of Science, Scopus and Google Scholar. For each term, searching continued to theoretical saturation, where the articles returned contributed little or no new relevant ideas to the review (typically within the first 50 results). In 2024, following the publication of the REF2021 impact case study database, this search was repeated and relevant additional literature added, since the format and assessment criteria for the impact component were almost identical in REF2014 and 2021. Unless

specified, all mentions of REF (including the numbering of Units of Assessment) refer to 2014.

2.1 REF context

In the UK, the question to what extent public expenditure on research and universities is justified led to a perceived need for research assessment as part of the new public management approach under the Conservative government in the 1980s (for a detailed history, see e.g. Smith *et al.* 2020: 17-25). As a result, a "research assessment exercise" (RAE) was introduced in 1986, and repeated in 1989, 1996, 2001 and 2008. These exercises were then developed into the "Research Excellence Framework" (REF), conducted for the first time in 2014 and then again in 2021 and 2029 (see Table 1 for an overview). As REF, these exercises included an additional component of research assessment, namely to assess research impact (for a policy-maker's perspective on the shift from RAE to REF, see Hill 2016).

Table 1: Timeline of research assessment in the UK

Assessment	RAE					REF		
Year	1986	1989	1996	2001	2008	2014	2021	2029
Party of government	Conservative		Labour			Conservative		Labour
Other information	1980s: Drive to justify public spending		1997: Focus on evidence-based policy	2006: Warry report, recommends impact assessment		2010: RAND report, recommends peer review for impact assessment; pilot of impact component		

Research assessment, and to an even larger extent research impact assessment, involves the reduction of something specific and complex into something generalisable and comparable. As such, they are a part of the wider audit culture described, for example, in Shore and Wright's pointedly titled article "Governing by numbers" (2015), which defines this as "reducing complex processes to simple numerical indicators and rankings for purposes of management and control" (Shore and Wright 2015: 22). The marketisation of research through regular assessment exercises encompassed a culture change within and around academia from a system of trust and professional autonomy towards one of auditing and accountability influenced by neoliberalism, as described for example by Olssen and Peters (2005).

A shift to audit culture also includes the notion that measurement becomes "financialised" (Shore and Wright 2015: 24), which for the UK context is manifested in the link between research assessment and large-scale funding allocations. Some kind of numerical input is a prerequisite for transparent allocation of funding based on formulae, and there was therefore a potentially multifaceted role for metrics in research assessment, with an incentive to increase levels of metricisation. As consecutive research assessment exercises slowly progressed change in academic culture, between the RAEs of 1996 and 2001, "the moment of the metrics occurred", meaning that "academics could no longer avoid the consequences of the developing systems of measure to which they were becoming increasingly subject" (Burrows 2012: 359).

In parallel, the idea of "evidence-based policymaking" was promoted by the incoming Labour government in 1997 (Smith *et al.* 2020: 19). Although this was separate from the metricisation of universities, it also contributed to the development of a desire to articulate the "societal return on investment" (Smith *et al.* 2020: 17) in research. Biri *et al.* (2014: 1) see this codification of the "return on investment" question into UK research life as "the genesis of the 'impact agenda'". This dual climate of increased metricisation and increased justification of research as a public good was the backdrop for a more explicit impact agenda, in which researchers would provide data to help solve policy problems and other societal issues. The linear and positivist view of research use presupposed in this agenda (Smith *et al.* 2020: 20) fosters, at best, what Rickards *et al.* (2020: 4) term "1st generation research impact culture" (a knowledge deficit model in which one-way knowledge transfer is able to generate impact, as long as recommendations from research are adopted, Grant 2023: 1), as opposed to, for example, a more two-way, co-creative research process, which Rickards *et al.* (2000) would call 2nd generation impact culture. The assumptions underpinning 1st generation impact culture seeped into the design of the research assessment system and were then further embedded and enacted in the research system through the implementation of the assessment system.

The development of an explicit impact agenda resulted in the transition from an assessment focused on research outputs (namely, RAE) to a wider assessment of publicly funded research, which included its benefit for society and for the environment in which the research was conducted (namely, REF). A milestone in this transition was the 2006 Warry Report (Warry 2006) with its strong recommendation to increase measurement and

communication of research impacts. The ensuing opportunity to redesign the assessment system so that it extended to research impact again opened up a space for questions about the role of metrics. Should there be more, or less, emphasis on metrics for the assessment of research quality? And could there be a metric that assesses research not for its value to academia (as shown by e.g. citations and associated metrics) but for its value to the wider society?

2.1.1 Purposes of and concerns around research assessment

Having explained the political background and decision to introduce research impact assessment, in this section I discuss the purpose of the assessment itself, that is, the conclusion that one would like to draw from the assessment outcome. Ideally, it should be this purpose that shapes the approach to assessment.

Dotti and Walcyk (2022: 2) list different ways to assess societal impact, with REF as an example of "*ex-post* evaluation" (as opposed to "*ex ante*" anticipatory methods that are applied in funding applications). They highlight the four different major purposes for research impact assessment proposed by Belcher *et al.* (2017: 2): analysis, allocation, advocacy, accountability. The selected instrument needs to have validity for the purpose of the assessment and be robust over the relevant timeframe. An example illustrating how measures may lack such validity is altmetrics (alternative metrics that measure the attention that research publications receive online), which is vulnerable to influences that are completely unrelated to research impact, such as the sale and subsequent transformation and/or decline of individual social media platforms. With reliability of that metric being potentially compromised, it has not enough validity for allocation decisions to be made on its basis.

All four purposes of research assessment suggested by Belcher *et al.* (2017), namely analysis, allocation, advocacy and accountability, can be applied to REF. Based on the account of its history and political context above, the purpose of introducing impact assessment in the UK is accountability, but the implication is allocation of funding. Following the publication of REF2014 ICS, these have been used for analysis, and also for advocacy, especially in disciplines that are sometimes perceived as less important than others and deprioritised in funding decisions (e.g. see the British Academy's recent analysis of REF2021 ICS which aimed to demonstrate the societal value of SHAPE [Social Sciences, Humanities and the Arts for People and the Economy] disciplines, Wagner *et al.* 2024). However, clarity over the purpose

is needed in order to design the exercise in line with that purpose (as also stated by the Harnessing the Metric Tide report, Curry *et al.* 2022: 29), because, as argued below, some purposes need to be carefully balanced with each other to retain an environment that allows for exploratory research.

The official guidance for REF2014 (HEFCE 2011) states three purposes, two of which map onto the four possibilities listed above (highlighted in bold):

> "The primary purpose of REF 2014 is to produce assessment outcomes for each submission made by institutions:
>
> a. The four higher education funding bodies intend to use the assessment outcomes to inform the selective **allocation** of their grant for research to the institutions which they fund, with effect from 2015-16.
> b. The assessment provides **accountability** for public investment in research and produces evidence of the benefits of this investment.
> c. The assessment outcomes provide benchmarking information and establish reputational yardsticks, for use within the higher education (HE) sector and for public information."
>
> (HEFCE 2011: 4, my emphasis)

There is an inherent tension in the question of whether the aim of the assessment is to rank individual universities against each other in a competition, for example to allocate limited funding, or whether the aim is analysis of UK HE research impact as a whole entity, with REF as "a national audit of university and equivalent research in the UK" (Gow and Redwood 2020: 66) and an "accountability exercise" for the funding bodies (Gow and Redwood 2020: 67). The latter purpose, which is linked to the origins of research assessment in the 1980s, is closer to methods that would allow calculations of the (monetary) return on research investment, to provide precise accountability to the taxpayer, and to pinpoint what proportion of UK taxpayer research funding has societal impact in the UK. An approach that uses standardised metrics would facilitate such an analysis that allows for accountability, in addition to assessment for allocation.

Grant *et al.* (2015) appear to have been commissioned by HEFCE (the Higher Education Funding Council for England, which turned into Research England in 2018) to use the dataset of REF2014 ICS to create an overview of UK research impact generally. However, they acknowledged that this was not a valid endeavour, because a sample of impact that is self-selected for assessment purposes cannot be representative for more general activity, as illustrated for example by Meagher and Martin (2017: 19-20). The REF guidelines encouraged universities to submit the "strongest examples" of impacts (HEFCE 2011: 28),

rather than representative examples, which emphasises the competition aspect of the exercise and is consistent with the stated purpose of allocation. This in turn raises the question whether a more representative sampling would be either feasible or desirable.

A sector-wide exercise for analysis and accountability would require a degree of representativeness that cannot be achieved with a self-selected sample that results from universities having some choice over what impacts to submit. It would therefore need to be de-coupled from assessment for allocation where universities can choose which "strong examples" of impact to submit, making the two purposes (sector-wide analysis and university-focused assessment) mutually exclusive. This ability of universities to choose, however, is essential in an environment that allows research that is not designed to have impact. In a view where research is a primary function of universities, and impact is a secondary function that only a subset of that research fulfils (as championed by e.g. Smith 1997), the authority over sampling of research for impact assessment purposes needs to sit with universities in order to preserve a space for "blue skies" research. This links directly into the question of the purpose of universities, which is well beyond the remit of this thesis; see, for example, Śliwa and Kellard for a discussion of the development from the Humboldtian approach of universities as a space that is protected from outside interest to a view of universities as institutions that meet "national needs as defined by government policy" (Śliwa and Kellard 2022: 15).

A foundational contribution to this discussion is Gibbons *et al.*'s (1994) book, titled *The New Production of Knowledge*, where they introduce the terms Mode 1 and Mode 2 research. Mode 1 refers to research where "problems are set and solved in a context governed by the, *largely academic*, interests of a specific community" (1994: 3, my emphasis), whereas Mode 2 research "is carried out in a context of *application* […] characterised […] by heterogeneity" and "*socially accountable*" (1994: 3, my emphasis). In a follow-on publication (Nowotny *et al.* 2003), some of the original authors relate this paradigm shift in the research landscape to research assessment, namely the RAE, under the heading of "accountability of science" (2003: 183). From the descriptions of Mode 1 and Mode 2 research, the latter is more reminiscent of research that is designed, or likely, to have more direct societal impact than exploratory Mode 1 research. The question is therefore whether the introduction of research impact assessment in REF has had the effect of shifting research practice from Mode 1 towards Mode 2.

In their analysis of UoA19 (Business and Management) ICS, Hughes *et al.* (2019) treat Mode 1 and Mode 2 research as two extremes of a spectrum. They find that most research cited in ICS as underpinning the impact is not clearly one or the other (Hughes *et al.* 2019: 636). While some see Mode 2 as the type of research that should underpin impact because of its focus on co-production, interactivity and user involvement, this was not widespread in this UoA, and 25% of the research could be clearly classed as Mode 1. They conclude from this that research questions are not (yet) normally dictated by users (Hughes *et al.* 2019: 631) and that therefore the impact agenda is less of a threat to the independence of research than feared by some. Similarly, Watermeyer and Tomlinson (2021) conducted a survey with academics named in ICS across four Social Science UoAs to determine how inclusion in the REF impact submission has affected their sense of academic self. They conclude that the REF impact agenda did not seem to have changed research practice between its official introduction in 2010 and the survey ten years later (2021: 7).

However, other studies find that the introduction of REF does influence researchers' practice. For example, exploring the writing and publishing behaviours of academics in the age of the REF, Tusting (2018: 483) found that writing was channelled towards research articles and publishing towards high-impact journals (that is, towards publication types and venues highly valued in research assessment), and that even research topics were affected, albeit to a lesser extent. Williams (2020) theorises research impact assessment across different fields (e.g. academia, politics, media) and notes that those within the system "have adjusted their priorities to include the performance of impact" – this includes universities employing dedicated staff, or new technology being developed for tracking impact, as well as academics changing their behaviour by, for example, communicating their research via blog posts (Williams 2020: 196-197). Brauer *et al.* assert that impact assessment turns "contemporary values" into "facts" which shape the research (2019: 66). This is also evident in the interviews with senior academics in Chubb and Reed (2018), some of whom report that the increased saliency of the impact agenda in the early 2010s had influenced their research priorities away from "the cutting edge of a discipline" towards "problem-solving" (Chubb and Reed 2018: 305).

MacDonald (2017), Chubb and Reed (2018) and Weinstein et al. (2019) report concerns from researchers that the impact agenda may be distorting research priorities, "encourag[ing] less discovery-led research" (Weinstein *et al.* 2019: 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" (94), and further questioned by Hill (2016).

Voicing similar concerns, Moran and Browning (2018) see a danger to "blue skies" research from the need to provide a "pathway to impact" statement in funding applications, which might influence the "way research projects are constructed" (2018: 258). However, having linked REF2014 ICS to UKRI grants, Yaqub *et al.* (2023) show that while in most cases the impact submitted to REF was aligned with the impact predicted in the grant application, there were also many cases where the predicted and submitted impacts were not aligned. They highlight the existence of and need for both situations: those where impact area and beneficiaries are identified *ex ante*, and those where there is space for changes to occur in the journey and serendipity to create unanticipated benefits.

Khazragui and Hudson (2014) propose that research impact is inherently linked to innovation, which by their definition needs external partners, at least in order to achieve marketisation. This contrasts with Machen's (2020) claim that most research is not inherently innovative, compared to "critical" research which challenges, rather than confirms, existing assumptions. Machen claims that it is this non-innovative, confirmatory research that is more likely to lead to impact because it is less disruptive and therefore more likely to be heard. Her model of impact from "critical research" includes concerns that such research may be jeopardised by the REF focus on impacts that can be evidenced just a few years after the underpinning research was published; this concern about the tight timeline dictated by the REF cycle is shared by Simpson (2015). Similar discussions about what the goal of research should be and how researchers should relate to the impact agenda are summarised for different disciplines in, for example, Smith and Stewart (2017, Social Policy), Moran and Browning (2018, Politics), Hughes *et al.* (2019, Business) and Brauer *et al.* (2019, Tourism).

Overall, with different possible purposes for research impact assessment, various problems can arise from tensions between different purposes that are officially or unofficially attached to a single exercise. This can include consequences for the freedom and potentially integrity of research, for example through an over-emphasis on problem-solving or research questions and directions influenced by funders or commercial partners. If space for

exploratory, non-applied Mode 1 (Gibbons *et al.* 1994) research is to be preserved, the assessment of research impact cannot be based on comprehensive coverage of all research.

To make a better-informed decision about this fundamental question reaching across the UK university sector, HEFCE commissioned a report from RAND Europe to review international practice of impact assessment and make recommendations for the impact component of the newly planned Research Excellence Framework (REF). Having analysed four different models, the report (Grant *et al.* 2010) recommended the Australian approach (Research Quality Framework, RQF) of using case studies, which had been developed (Department of Education 2006) but never implemented due to a change in the Australian government in 2007 (Williams and Grant 2018: 97). HEFCE defined criteria, namely "significance" and "reach", to operationalise that approach. After a pilot of the impact component, with 29 universities submitting case studies to five Units of Assessments (UoAs) in 2010, peer review, as opposed to metrics, was re-established at the core of the new exercise, as it had been for RAE; in REF, this allowed for impact to be assessed through narratives that could be put forward for assessment of a wide, unanticipated range of impacts and pathways. Crucially, the narrative approach enabled the inclusion of qualitative and contextual information to be taken into account for the assessment.

The inaugural REF (2014), from which the dataset in this study is drawn and whose context therefore is given most weight in this thesis, had the following structure:

1. Assessment was conducted in disciplinary categories, with 36 Units of Assessment grouped into four Main Panels.[2] In the preceding RAE (2008), there had been 67 Units of Assessment grouped into 15 Main Panels, so the number of Main Panels and Sub-Panels (Units of Assessment) was reduced significantly in the transition to REF, simplifying the structure. The Main Panel structure of REF2014 was kept for subsequent REFs, and the number of Units of Assessment was reduced from 36 to 34 in REF2021 (remaining at 34 in REF2029). In this thesis, any references to Units of Assessment (UoAs) follow the numbering of REF2014 because this is the source of the present dataset.

[2] The Main Panels were not labelled beyond A, B, C and D, but they correspond broadly to disciplines as follows:
Main Panel A: Life Sciences
Main Panel B: Physical Sciences
Main Panel C: Social Sciences
Main Panel D: Arts and Humanities

2. Guidance and criteria were published in early 2011. Submissions were made in November 2013, and outcomes were published in November 2014. In 2015, impact case studies (ICS) were published in a publicly available database.[3]
3. There were three components that constituted a submission from a Higher Education Institution (HEI) to a Unit of Assessment: Outputs (worth 65% of the overall score of a submission), Impact (worth 20% of the overall score), Environment (worth 15% of the overall score).
4. The mode of assessment for all three components was peer review by established academics and, in the case of the Impact component, additionally by research users, that is, non-researcher experts. This process is discussed in more detail in sections 2.1.3 and 5.1.
5. 154 HEIs made submissions in one or more UoAs, which included between 2 and 260 ICS per HEI (Hinrichs and Grant 2015: 1).
6. Each element that was included in a submission, that is, each research output, ICS and overview template (for the Impact and Environment components), was assigned a rating. Possible ratings were 1* to 4*, with 4* being the highest rating, plus "unclassified" for ineligible parts, or those that did not meet the threshold for a 1* rating. The descriptors for each star rating differed across the three components. For Impact, they were, from 4* down to 1* respectively: "*outstanding / very considerable / considerable / recognised but modest* impacts in terms of their reach and significance" (HEFCE 2011: 44). Results were published at the level of submission, that is, for each UoA in each university, but not for individual elements within those submissions (such as individual ICS).
7. On the basis of these star ratings, funding was awarded to universities. The process of funding allocation is explained in a blog post by Reed and Kerridge (2017). Funding was only awarded for ratings of 3* and above, but not for 1*, 2* or Unclassified ratings, and 4* ratings were worth more than 3* ratings. A 4* ICS in REF2014 was worth £46,300 per year on average until the next REF several years later (Reed and Kerridge 2017).

[3] The database can be accessed here: https://impact.ref.ac.uk/casestudies/search1.aspx

2.1.2 Approaches to research impact assessment

As shown above, accountability and allocation are the purposes stated explicitly in REF policy documents (HEFCE 2011: 4). Both ideally call for complete coverage, but as argued above, this cannot be achieved in a higher education system that encourages a range of exploratory and applied research. In addition, if the assessment results are to be used for allocation, a formula is needed and therefore one output of the assessment needs to be a number that can be used in such a formula. The assessment design needs to balance these different requirements, namely partial coverage and the production of input for a formula. Different options have been proposed at various times for the design or re-design of the exercise:

a) Assuming that the quality of research impact is associated with the quality of research, a separate assessment of research impact would be redundant if there is assessment of research quality.
b) Existing or new metrics that are proxies for impact could be used.
c) A numerical value could be derived specifically for impact assessment through a peer review system.

These options are now discussed in turn.

a) Are research and impact (scores) correlated?

One of the most comprehensive studies of REF2014 ICS is Gow and Redwood's (2020) analysis of characteristics of top-rated ICS. The juxtaposition of "impact" and "research" in the title of their book, *Impact in International Affairs – The Quest for World-Leading Research*, seems to imply an assumption that world-leading research and excellent impact go hand in hand. Indeed, they suggest that "world leading research" and "world-leading examples of impact" "ought […] to have correlated" in REF scores (Gow and Redwood 2020: 59). However, the REF guidance does not make that assumption; rather, the threshold for research quality permitted for underpinning research was set at 2*, lower than the threshold for funding that was applied to outputs (3*). Following REF2014, Lord Stern suggested in an independent review that even the 2* threshold should be removed (Stern 2016), and while this threshold was maintained in REF2021, it will no longer be a requirement in REF2029 (Research England 2023: 11).

Given the potential problems with comprehensive coverage of research impact assessment, and the suggestion that it may (or should) produce similar results to existing research

assessment, it is worth examining to what extent research scored as "4* - world-leading" and impact scored as "4* - outstanding" are linked. Pinar and Horne's (2022) analysis of grade point averages (GPA) of the three REF components indeed suggested that these are so closely and positively correlated that one element could be dropped, in order to reduce the burden of preparing the various components for REF. Similarly, Terämä *et al.* (2016) compared GPA of impact and output across UoAs and submissions and concluded that a correlation did exist. They interpreted this positively, as showing that the pursuit of impact did not jeopardise excellent research, which was a commonly voiced concern at the time. However, an alternative explanation for Pinar and Horne's (2022) and Terämä *et al.*'s (2016) findings may be that those institutions that had the money and support because they were already recognised as research-intensive also had the best chance to generate, evidence and present impact (a possibility suggested for example by Penfield *et al.* 2014: 30), compared to institutions where teaching loads were higher and support for activity around impact was more sparse. Support for such an explanation may be found in Williams *et al.* (2023), who showed that the submitting institution was the strongest predictor of a high score in REF2014. A third explanation may be found in the belief by some assessors that there was an intrinsic link between excellent research and impact (Derrick and Samuel 2017), which Gow and Redwood (2020) also seem to imply. Dunlop (2018: 288) went further to suggest that a focus on real-world problems could even enhance the quality of social science if those real-world problems are linked to the theory and methods of a discipline.

However, other studies suggest that the quality of research and impact as assessed by REF panels were quite independent of each other. For example, Simpson (2015) showed that Russell Group universities topped the impact rankings in most UoAs, but not everywhere. Especially in smaller disciplines such as his own UoA24 (Anthropology), ICS had relatively more weight in smaller departments due to the minimum requirement of two ICS regardless of FTE submitted for output assessment. With this in mind, Simpson illustrates that the impact submission put some universities higher in the rankings than would have been expected by traditional views of the university landscape, and in divergence from the scores of the respective output submissions. Kellard and Śliwa (2016) come to a similar conclusion for UoA19 (Business and Management). They correlated the impact and output GPA for all submissions to this UoA and detected no statistically significant relationship between the two; in other words, *research* that is assessed as excellent and *impact* that is assessed as excellent do not necessarily happen in the same business schools, and one is possible

without the other being present (or rather, being submitted). Indeed, even Pinar and Horne (2022) point out that the impact element specifically seemed to have affected GPAs more in certain UoAs. They show that some disciplines, especially in Main Panel B (broadly Physical Sciences), were disadvantaged in funding outcomes, that is, these units would have been better off if they had received funding on the basis of only outputs and environment. This suggests that, if funding allocation is to take into account the quality of research impact specifically, it is indeed necessary to assess impact independently of research outputs, as was done in REF2014.

b) Other sources of existing metrics

With mixed evidence around the link of the research and impact components in REF, other studies have explored the possibility of using other existing indicators or metricised methods to arrive at scores for research impact that could be used for accountability and/or for allocation, while avoiding the burden of a separate REF component. For example, Ravenscroft *et al.* (2017) correlated various citation metrics with impact GPAs. After building citation networks from open access sources, extracting academic references from Section 3 of the REF ICS template (for an overview of the template, see section 2.1.3), and matching these to their citation networks, they analysed the five UoAs where they were able to identify the greatest number of matches between ICS and their database. They found no correlation between impact GPA and any of the indicators used to express academic impact, including altmetrics, and consequently they found it impossible to predict impact scores in a regression model with citation metrics as independent variables. Ravenscroft *et al.* (2017) offer two explanations for their finding: the citation metrics and impact scores were measuring different things, or citations and non-academic impact were actually unrelated. The authors conclude that, whichever underlying mechanism was causing the finding, impact assessment could not happen on the basis of citation metrics. Instead, they offer an alternative vision of building a machine-learning model that could predict impact scores from a host of different texts (policy documents, media etc.), while acknowledging that this approach also held many problems (Ravenscroft *et al.* 2017: 18).

In another empirical study, Dunlop (2018) used journal rankings as proxies for the quality of research outputs listed in UoA21 (Politics) ICS and found that the proportion of articles in high-ranking journals was larger in those impact submissions that scored exclusively 3* and 4* than in submissions that included lower ratings. Wooldridge and King (2019) suggested a

correlation between altmetrics and ICS scores in Main Panel B, but they acknowledge that this is crude and can only predict into which tertile an institutional submission falls. Bornmann *et al.* (2019) investigated relationships between several altmetrics measurements and impact GPA and found only weak, if any, correlations; they explicitly backed Ravenscroft *et al.*'s (2017) conclusion that altmetrics of any kind cannot be used for measuring societal impact.

A more theoretical approach was taken by Khazragui and Hudson (2014), who developed an econometric formula to measure impact. Focusing on the purpose of accountability, their aim was to enable the calculation of the economic value of any kind of research impact. Underlying this is the following simplistic view of impact and measurement: Real Situation minus Counterfactual (what if the impact had not happened, the research had not been done, or the pathway to impact not followed) equals Total Impact (this is not to be confused with the measure of Total Research Impact proposed by Pepe and Kurtz 2012, which refers to citation measurements and therefore "academic" impact). For that formula, Khazragui and Hudson suggested that less credit should be given for benefits outside of the UK because British taxpayers' money for research funding should translate into domestic benefits. This latter view, however, is at odds with the high regard of international impacts in the REF assessment, as highlighted by Grant *et al.* (2015).

Having analysed the ICS submitted to the 2010 REF pilot exercise for the ways in which quantifiable, economic benefits were (or were not) claimed, Khazragui and Hudson (2014) found that it was impossible to metricise and calculate the nature and range of impacts submitted even in the pilot submissions of 29 HEIs across 5 UoAs in this rigid way, that is, using an econometric formula. As a middle ground, they called on HEFCE to provide a list of conversion factors for frequent effects (i.e. impacts), such as for human life or traffic congestion. This would enable universities or assessment bodies to convert everything into equivalent monetary values, which could be used for standardised assessment. Indeed, such conversion factors or proxy measures exist in some cases, for example for Disability-Adjusted Life Years in medical research, or ways to calculate compensations for victims of a crime in the legal system. However, the existence of REF and of research impact assessment is not a legal necessity, but a political decision, and the approach suggested by the authors would be highly problematic for several reasons. First, it would risk excluding or devaluing any impacts that HEFCE (or its follow-on organisation Research England) did not anticipate

and provide a metric for (a point also made by Ovseiko *et al.* 2012). Second, it would foster an impression of fair and objective measurement in an area where many aspects simply cannot be measured even with the help of conversion factors (e.g. in UoA4 Stirling *EvoFit*[4]: the feelings of victims when a criminal is captured, enabled by new techniques). Finally, for the purpose of calculating comprehensive return on investment of research funding, universal coverage of all research activities would be required, rather than the selection of impacts assessed in REF – but such selection is necessary if there is to be space for Mode 1 ("blue skies") research, as argued in section 2.1.1.

Another study that used the ICS submitted to the 2010 pilot exercise is Ovseiko *et al.* (2012) who discuss Oxford University's submission in clinical medicine, that is, a discipline where metrics are commonly used and outcomes are often measured in and beyond academia. Having critically reviewed several metric indicators that are often suggested as contenders for research impact assessment (e.g. patents, spin-out companies, research income from the NHS), they conclude that even in an already highly metricised UoA, there are "significant challenges" in using such indicators in a consistent and reliable way, which would be a prerequisite for a fair process of allocation (2012: 18).

The most comprehensive and authoritative publication on the subject is the 2015 *Metric Tide* report (Wilsdon *et al.* 2015). Commissioned by HEFCE, it discusses the increasing role of metrics in research assessment comprehensively and makes recommendations for their responsible use. The report critically evaluates several different contenders for quantitative indicators of research impact, including citations outside of academic publications, and points out that these are often discipline-specific and could therefore not be applied systematically (Wilsdon *et al.* 2015: 45). Regarding the impact component of REF2014, the report expresses the concern that a focus on defined indicators may "constrain thinking around which impact stories have greatest currency" and in effect lead to a narrower research base in UK HE (Wilsdon *et al.* 2015: 139). It also points out the tension between an inclusively broad definition of research impact, as applied in REF, and the fact that "quantitative data and indicators are highly specific to the types of impact" (Wilsdon *et al.* 2015: 132). Perhaps most relevant to the present discussion, it emphasises that context is essential for assessing quantitative indicators, and that this context needs to be provided in the form of "a narrative element" (Wilsdon *et al.* 2015: 132). The report therefore concludes

[4] A list of ICS in my sample with links to the original in the REF database, by UoA, is provided in Appendix A.

that peer review of narratives, despite considerable flaws, should continue to be the mode of research impact assessment in REF, until potentially metrics are better researched and validated – a recommendation reiterated by the follow-on *Harnessing the Metric Tide* report (Curry *et al.* 2022: 28).

c) Creating a metric through peer review

With neither the REF grade point averages nor other existing metrics proving a sufficiently reliable proxy, there is a need to create an impact-specific metric for the purposes of funding allocation which keeps flexibility for varied and unanticipated impacts, as well as maintaining the ability of submitting units to select impacts for submission and thereby protecting space for a wide variety of research. The required flexibility can be provided through the use of narratives, to which peer reviewers attach a score that can be used for funding allocation. This essentially quantifies a qualitative input. However, while this approach satisfies the requirements of flexibility and comparability, it is not without criticism.

In their pointedly titled article "Governing by Narratives", Bandola-Gill and Smith (2022: 1857) discuss research impact assessment as the newest attempt to quantify a higher education domain that, until 2014, was not being measured. They use the term "Commensuration" to describe how, in trying to get the comparable data needed as a basis for funding decisions, something qualitative is being quantified. They compare the attempt to squeeze the messy reality of research impact onto a template of just four pages to the representation of the physical world in a map. In both cases, much of the original information cannot be translated into the format in which it will be used. Rather than framing this as a "metrics versus narratives" dichotomy, Bandola-Gill and Smith (2022) posit that narratives are being turned into a kind of metric through the restrictions of the template and the assumptions of universities of what might be likely to receive high scores in REF (for an example of how a research-and-impact-journey was transformed, or perhaps distorted, into an ICS in this way, see Russell and Lewis 2015).

The peer-review model is regularly criticised as an undue burden on the sector. According to a RAND analysis, the cost of an ICS in 2014 was on average £7,500, or £55m across the sector (Manville *et al.* 2015b). While this seems like a high cost for one piece of writing for assessment, this amount translates to a 3.5% transaction cost, that is, the cost of producing an ICS compared to the financial resource allocated on its basis (Morgan Jones *et al.* 2022: 735). By comparison, however, the transaction cost of research funding allocated through

research grants is nearly 10% (DTZ Consulting and Research 2006), making the impact component of REF a relatively efficient method for allocating funding, despite the impression that it is unduly resource-hungry.

A further concern that is voiced regarding the use of narratives for research impact assessment is a potential for presentation to influence assessors' judgement, compromising the integrity of the assessment. Such concerns around language and narratives are discussed below (section 2.2), and this thesis is a contribution towards examining such claims.

Despite these concerns, the model of generating metrics to assess impact (the 1*-4* rating system) with the help of a narrative (the ICS) was adopted in REF2014 and 2021 and re-confirmed for REF2029. The following section therefore dives deeper into this model of assessment.

2.1.3 Assessing research impact through narratives

With the decision to use narratives instead of metrics-based assessment arises the question about the reliability and accuracy of assessing narratives. This section describes the process of assessment and points towards concerns that are sometimes raised.

The definition of impact used in REF2014 is: "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: 48). It is further specified that "[i]mpact **includes** the reduction or prevention of harm, risk, cost or other negative effects" (2011: 48, bold in original). Such impact is to be assessed against the criteria of "reach and significance" (2011: 27), which the panels were instructed to assess "as a whole, rather than assess 'reach and significance' separately" (2011: 44).

In REF2014, there were two components to an impact submission: a UoA-level template which summarised the approach to impact in the submitting unit (not further discussed in this thesis), and at least two case studies (the number required was dependent on staff numbers, full-time equivalent). The case studies (ICS) were submitted on a 4-page template provided by HEFCE. The template included five sections, with indicative word or other limits (2011: 52):

1. Summary of the impact (100 words)
2. Underpinning research (500 words)
3. References to the research (up to 6 references)
4. Details of the impact (750 words)
5. Sources to corroborate the impact (up to 10 references)

The indicative word limits were often not adhered to, partly because they did not normally fill the four pages that were allowed. Only limits for each section were indicative; page limits and formatting requirements (Arial 11, 2cm on all sides, single spaced) were clearly stated as the smallest format allowed, but while the page limit was generally adhered to, some universities applied more creativity to the formatting. The template for REF2021 allowed five pages but kept the same indicative word and reference limits.

According to Derrick *et al.* (2014), there were as many as 41 different policy and guidance documents on the REF website prior to the 2014 submission. Their analysis based on word frequencies concludes that, despite this proliferation, it was not clear what exactly REF was hoping to see. One of the words most closely connected with "evidence" is "indicators", which they read as showing an implicit bias towards measurable impacts (Derrick *et al.* 2014: 152). The documents provided ample guidance around the presentation and submission, that is, the process of assessment, but were light on descriptions of the expected content of impact. Derrick *et al.* (2014: 152) see this as an opportunity for "the academic literature", rather than government policy, to advance understandings of impact (see for example the contribution by Oancea 2013 on different interpretations of impact across seven disciplines); in practice, this discussion is more likely to have taken place within the assessment panels. These peer reviewers had to navigate the tension between representing both the university sector and the state-dictated assessment system. Concerns have been raised that this dual role may jeopardise the integrity of the exercise, and that the uncertainty of the criteria may have led them to "mark generously" in order to "showcase" research (Derrick and Samuel 2018: 680).

Lim (2020: 108) argues that the use of narratives as impact assessment, as opposed to externally available metrics, hinges on the trust of those affected by the assessment outcomes in the professional judgements made as part of the assessment process. Related to this trust is the question of construct validity (Messick 1989), that is, whether the score *is* awarded for what it *should be* awarded for – something that is near impossible to achieve perfectly in any assessment context, but that it is important to work towards approximating. Gow and Redwood (2020: 69) claim that in REF2014, it was "simply the 'oomph' of a case study being gauged, rather than fine tuning grades". Claims like this risk undermining trust in the assessment process and therefore they need to be critically assessed. Sources on the assessment process include a HEFCE-commissioned report (Manville *et al.* 2015a),

publications by sub-panel Chairs (e.g. Greenhalgh and Fahy 2015; Pidd and Broadbent 2015) and research on the assessment process (e.g. Derrick 2018; Williams *et al.* 2023).

As part of the evaluation of REF as an exercise, HEFCE commissioned RAND Europe to analyse the assessment process and provided them with scores for individual ICS, rather than the results for complete submissions that are available publicly (Manville *et al.* 2015a). This is the most comprehensive report on the process. Assessment panels included academic assessors, most of whom were involved with all parts of the submissions, and impact assessors (research users from outside of academia), who were only involved in assessing the impact component. Based on a survey and interviews with assessors across all panels, Manville *et al.* (2015a: xv) report that 70% of the academic assessors were also involved in preparing their own institution's submission and that they felt the benefit of being part of the assessment process because this increased their understanding of "how to present good case studies".

At the start of the assessment process, extensive calibration took place. According to the RAND report, it was in these sessions that assessors negotiated what did and did not count as impact. This seems problematic because if the boundaries of impact for REF assessment purposes were only defined by the assessors after submission, universities could not have had this information at the time of preparing the ICS – a concern that seems to be corroborated by Derrick *et al.*'s (2014) outline of the multitude of policy documents and the resulting lack of clarity. Following calibration, ICS were allocated to assessors in a process that varied across UoAs and with varying degrees of transparency. Some assessors had the impression that the allocation was random, while most report that it was done according to the academic expertise of an assessor or the type of impact (Manville *et al.* 2015a: 15). In some cases, conflicts of interest were abundant and reduced the number of relevant potential assessors available for a given ICS. This may have contributed to a perceived randomness of allocation or at least to a lack of specific expertise by allocated assessors for a given ICS (personal communication with REF2014 panellist). Each ICS was assessed by at least two, but mostly three and in some cases four assessors (Manville *et al.* 2015a: 15).

One point of discussion was the extent to which prior knowledge or familiarity with the research or impact context matters. Assessors especially in Main Panel D (broadly Arts and Humanities) struggled with determining where to draw the line of taking any known background into account, which could be seen as giving an unfair advantage where an

assessor could essentially add details to the assessment decision compared to other ICS that were assessed at face value on the basis of the information that fit on four pages (Manville *et al.* 2015a: 13). Arguably, this question should not arise where universities present ICS in a way that includes all the information needed to make an informed and comprehensive assessment, which was of course restricted by the 4-page template in 2014. Several assessors surveyed by RAND expressed frustration about not being allowed to penalise for violations of "indicative word and page limits", though (Manville *et al.* 2015a: 29).

Some assessors reported that they found it easier to rank ICS against each other than to apply a grade boundary within those rankings, and they would have found it helpful to have something like an expected distribution (Manville *et al.* 2015a: 25). Such an expectation would essentially turn the exercise into a norm-referenced assessment (Fulcher 2010: 31) where there are cut-off points in relation to the quality of other submissions, rather than the quality of the submission under consideration in its own right. As it stands, REF is overtly a criterion-referenced assessment, where each document is assessed independently of others against the descriptors behind the star ratings. One could argue, however, that the exercise is indirectly norm-referenced in a different way, given that the results inform the distribution of a limited overall amount of funding and that a greater skew towards high scores may reduce the worth of such a score. An alternative solution may be to make the criteria more specific to aid in their application even without expected distribution. For example, Williams (2015: 12) calls for "very clear guidelines on what constitutes impact at each standard of excellence". This, however, may pose problems for the unexpected impacts: the more examples are given, the greater is the danger that universities feel unable to submit unconventional impacts that extend beyond comprehensive guidance. Gow and Redwood (2020: 30) make this point in relation to impact definitions and note that the "and many more" at the end of a long list of examples in the UKRI "Excellence with Impact" statement risks making those that are not named feel second-rate.

It is interesting to note that assessors in Main Panel A found that the criteria met their needs more so than assessors in other Main Panels (Manville *et al.* 2015a: 27). This raises the question whether the guidance was written from a Main Panel A perspective, or whether this is related to the fact that the types of impact expected in Main Panel A are also those that lend themselves to the metricisation called for in parts of the literature. A second way in which Main Panel A differed from the other Main Panels is their partial use of a hypothetical

8-point scale (1*-8*), where all ICS that were assessed as achieving a score of 4*-8* were given the top grade of 4*. The rationale was that the existence of impact that was "off the scale" should not bring down the boundary between 3* and 4* for those ICS that would be a clear 4* if they were compared to a 3* ICS but that look weaker compared to an "off the scale" ICS (Manville *et al.* 2015a: xiii).

Overall, the RAND report on the assessment process concludes that there was no bias and calls the process "fair, reliable and robust" (Manville *et al.* 2015a: xii), although it should be noted that this study was commissioned by HEFCE, who are likely to have had an interest in such an outcome. Pidd and Broadbent (2015) describe the whole REF submission to UoA19 (Business and Management) from the perspective of sub-panel Chair and Deputy Chair, and similar to the assessors surveyed by RAND, they report that the assessment was fair, based on assessors' expertise and discussions. They conclude that there was "no evidence that any particular type of impact was easier to judge or achieved higher scores than any other" (2015: 575). In both cases, this positive evaluation of the assessment process and reliability was self-reported, though. Terämä *et al.* (2016: 3) list sources that took a different view and assumed a certain bias in reviewers relying on their personal experience; likewise, Gow and Redwood (2020: 61) voice the "suspicion" that in some cases reputation was valued over evidence, or at least was allowed to compensate for missing evidence.

This suspicion is further corroborated by Williams *et al.*'s (2023) study that sought to explore whether automated assessment based on machine learning might be a possible alternative to the existing process of peer review of ICS, which they critique as being both open to bias and resource hungry. Unlike the sources discussed in section 2.1.2 (b), their approach did not include a reliance on available metrics or other existing documents, but tested whether REF ICS, once written, could have been scored by a model, and if so, which components were most predictive for high-scoring ICS. Reassuringly, they acknowledge that with any such attempt, past successes would skew the model to the disadvantage of novel cases, and they therefore caution that machine learning should only be used in conjunction with human judgement. For their analysis, the authors selected six features that might have influenced scores:

1. Unit of Assessment/Discipline
2. Submitting institutions
3. Explicit textual features (words and word combinations)
4. Implicit textual features (readability, sentiment)
5. Bibliometrics
6. Policy citations.

Of these, the first two stood out statistically as being most likely to predict scores. The explicit textual features, and the readability component of the implicit textual features, were also highly predictive, but the authors did not discuss the nature of these features in detail, or offer any further explanation for potential mechanisms for why these may have been predictive. Referring to my work (Reichard *et al.* 2020) in conjunction with their own findings, the authors assert that "the presentation of the narratives seems also to have influenced impact assessment" (Williams *et al.* 2023: 13) but there is no evidence in my or their work that "the presentation of the narratives" was the influencing factor, and it is equally possible that a common third factor influenced both the presentation and the score. Indeed, Williams *et al.*'s (2023: 13) finding that the submitting institution had more predictive power than the language used may point to such a common third factor. Moreover, as I will show in chapter 6, much of the language difference is related to the content of an ICS and not open to editorial decisions (e.g. the term *government policy*, as opposed to, for example, *in terms of*), a distinction that is crucial to make before voicing any claims that it was language (or presentation) that had an influence on scores.

2.1.4 Previous analyses of REF2014 impact case studies

In 2015, the impact case studies submitted to REF2014 were published in a database as a resource for a wide variety of purposes. Of an overall 6,975 ICS submitted to REF, 6,679 are included; the remaining ICS were not published because they contained sensitive content that could not be redacted (Grant 2015: 20). This provides a unique opportunity to evaluate the construction of ICS that were perceived by evaluation panels to have successfully demonstrated impact and to compare these to ICS that were judged as less successful, and the texts examined in this thesis are also drawn from this database.

One of the most comprehensive analyses of this database is a HEFCE-commissioned study jointly conducted by King's College London and Digital Science (Grant 2015). This document is widely cited in the REF2014 impact literature and was used by impact professionals in UK

HEIs in preparation of the 2021 submissions (personal communication from several impact professionals). The analysis was conducted over a 6-month period straight after the results were released in November 2014; the time pressure inevitably led to some compromises in the scope of research and presentation of findings. Their comprehensive report includes descriptive information on the overall submission landscape, and I draw on it throughout the thesis. As part of a range of analytical methods, the study also used some text-mining techniques, which will be critiqued below in section 4.1.2.

The database was subsequently used by researchers in many disciplines to explore the impact reported in their UoA or on certain topics across UoAs. Others have analysed the dataset to evaluate the structure of the exercise itself. Kelleher and Zecharia (2020: 3) situate the ICS themselves as "first tier" grey literature according to Adams *et al.*'s (2017: 435) taxonomy because they meet the criteria of accessibility and credibility, and suggest that they can therefore be used as sources for non-REF-related research, too. There are broadly three approaches to sampling in this dataset:

a. Finding impact types or pathways across the database;
b. Focusing on one Unit of Assessment;
c. Exploring a specific topic across the whole database.

For each approach, illustrative examples of studies are discussed in the remainder of this section.

a. Finding impact types or pathways across the database

Terämä *et al.* (2016) used text mining techniques on the whole dataset to determine six impact "classes". They identified six such classes, which they name education, public engagement, environmental and energy solutions, enterprise, policy, and clinical uses. One major limitation of this approach, using hierarchical cluster analysis, is that it only detects classes that appear often enough to be statistically significant. The authors highlight some clear omissions when they compare their classes to the categories outlined in the REF guidelines, including "international development". The fact that many ICS report on impact in this area is evidenced in Hinrichs *et al.* (2015) and discussed at length by Gow and Redwood (2020). These omissions may be in part explained by the fact that the method employed by Terämä *et al.* (2016) requires each submission as a whole to be assigned to one class, despite variations in the ICS included in the submission. On the other hand, assigning

just one class to each submission enables the authors to compare the classes (represented by several submissions, potentially across different UoAs) to the grade point average of the respective submission. Applying this to the submissions by UCL, they find that no class scores generally higher than the others. In particular, they note that there is no preference for economic impact, which contrasts with the assumptions put forward by Khazragui and Hudson (2014), and indeed the finding by Biri *et al.* (2014) who analysed ICS in certain technology-related UoAs from the submissions by UCL and found that "business" is the most frequent "impact type". The authors themselves were surprised by the finding that there seemed to be very little economic impact in certain UoAs (Terämä *et al.* 2016: 12). However, given that their chosen statistical analysis means that each submission was moulded into one class only, a submission that contains, for example, predominantly policy-related impact would be assigned to that class in its entirety, even if other ICS contain economic impact. In fact, the analysis of Grant (2015: 55) found that the word "policy" occurred in more than half of all ICS. The prominent role of public engagement as one of the six impact classes, even a "central component" (Terämä *et al.* 2016: 15), may be surprising because this was treated as a potential pathway to impact, rather than as impact in its own right, by REF2014 guidance (HEFCE 2011: 29-30) and acknowledged as challenging to assess by the RAND assessment report (Manville *et al.* 2015a: 29-30, 33). A possible reason for the prominence of public engagement in Terämä *et al.*'s (2016) analysis is that public engagement may be a frequently used pathway, using a higher percentage of the available words in an ICS compared to other impacts that can be expressed as simpler metrics, and certain words may be used more often when describing public engagement, all of which can cause a submission to gravitate to that class in their automated text analysis, even if the impact claim itself may be different. It is impossible to ascertain this relationship between pathway and impact without inspecting at least the "Summary of the impact" sections of individual ICS. Duncan and Manners (2017: 8) support Terämä *et al.*'s (2016) emphasis on public engagement, as they identified some element of "public engagement" in 47% of REF2014 ICS. Copley contextualises this by specifying that few ICS rested solely on public engagement as impact, and in the vast majority of ICS, this was "included alongside other impacts" (2018: 231).

Bonaccorsi *et al.* (2021) also used the whole database for a text mining approach. On the basis of a pre-determined list of nearly 80,000 people-related terms that they searched for in Sections 1 and 4 of the whole ICS corpus, they defined eight wide-ranging clusters of beneficiaries, including "School", "Health" and "Economy". Applying a more fine-grained

level of analysis, they found 46 distinct groups within those clusters and provided the main words that make up the clusters. As an analysis of who benefits from UK research, this is illuminating; the analysis is not linked to scores, though, and it is not clear from the article how frequent the identified user groups are in relation to each other.

b. Focusing on one Unit of Assessment

In addition to studies that address questions spanning across topics and disciplines, a second group of studies comprises those that focus on impact in one UoA or explore certain questions by using one UoA as a case study, rather than searching for topics in the whole database. Indicative examples across Main Panels (A, B and C respectively) are introduced next.

Greenhalgh, deputy sub-panel Chair of UoA2 (Public Health, Health Services and Primary Care), and Fahy (2015) conducted a manual content analysis of all 162 ICS submitted to UoA2. They acknowledge the role of presentation and that the 4-page template may have disadvantaged some impacts in the assessment. They also note that each ICS on average reports on three different impact areas. This is similar to Reichard *et al.*'s (2020: 8) finding of 2.8 impact types on average in a high-scoring ICS, although their typology is based on Reed (2018: 20-21) which cuts across UoAs and is therefore less fine-grained than Greenhalgh and Fahy's health-related typology. However, the findings from both studies could be interpreted to indicate that related impacts should not be separated out into different ICS.

Koya and Chowdhury (2020) applied a combination of discourse analysis and text mining to the ICS submitted to UoA11 (Computer Science and Informatics) in order to summarise the impact of research in computer science as expressed in certain indicators. These included patents, spin-offs, claims of increased employment and income generation, as well as savings made possible by improved processes across different sectors. All of these can be counted or are already numeric values, and therefore the authors suggest a template specific to their UoA that could capture these metrics for assessment with minimal additional context (Koya and Chowdhury 2020: 82). They acknowledge that the body of ICS is a snapshot of research impact selected for a certain purpose and not necessarily representative of impact across UK HE (2020: 83-84).

In order to describe the journeys from research to impact in UoA25 (Education), Cain and Allan (2017) investigated all ICS from that UoA that could be ascertained as having scored 3*

or 4*. They coded each sentence in these ICS as mostly pertaining to one of three categories: "research", namely the input or starting point of the journey; "practice", which is the term they chose to represent impact, that is, what users or other beneficiaries did or experienced as a result; "intermediaries", that is, for example, policy makers or professional organisations that were instrumental in bringing research to practice through a pathway to impact. From this, they identified four main stages from research to impact (discovering a problem > disseminating knowledge > incorporating it into resources > changes in practice), while acknowledging that the process is "usually messy and iterative" (Cain and Allan 2017: 724).

c. Exploring a specific topic across the whole database

A different approach was taken by studies that explore the complete database in search of impact on certain topics of interest. The research questions and results are often similar to those applied to discrete UoAs, the difference lies mostly in the sampling approach.

Brook (2018) provides results that are especially appliable for impact professionals and academics in the arts (2018). This study used Sections 1, 4 and 5 of 63 ICS across different UoAs to extract the types of evidence used to illustrate the impact of exhibitions. Similar to Cain and Allen (2017), Brook separated ICS that scored guaranteed 3*-4* from others and made inferences on scores on that basis. She found eight main types of evidence and emphasised that specificity and context matter: writers should include enough information to help the readers see the significance of, for example, a certain audience size (Brook 2018: 62).

A second study from the arts, specifically film making, is Kerrigan and Callaghan (2018). In addition to searching for certain terms related to film making in the ICS database, they complement their analysis with information from panel reports. One issue with this study is that it is not clear on what basis they comment on ICS that are "stronger" (which e.g. include testimonial quotes) or "weaker" (which e.g. display problems with linking to research), as they report no attempt to determine the score of ICS they read; their comments may have been based on the panel reports and simply collected for their relevance to the kind of ICS of interest to them. They emphasise that a "compelling and coherent narrative" enabled an ICS to contextualise information, which in turn helps the assessors to see the claims made in it and give it credit accordingly in the rating (Kerrigan and Callaghan 2018: 239).

Other studies that investigated certain topics across UoAs include:

- Hinrichs *et al.* (2015) for International Development,
- Williams (2015) for Sustainable Development,
- Robbins *et al.* (2017) for Engineering and Development,
- Brauer *et al.* (2019) for Tourism,
- Hanna *et al.* (2020) for Cancer Trials,
- Midmore (2017) for Agricultural Science,
- Kelleher and Zecharia (2020) for drug development to see how this process is generally structured in the UK,
- Pullinger and Varley-Winter (2017) for Statistics, who found that impact based on Statistics research was submitted across 20 of the 36 UoAs, and
- Kelly *et al.* (2016) for Nursing, who found that of those ICS that contained "nurs*" (i.e. words such as nurse, nurses, nursing, etc.), only 15% were submitted to UoA3 (Allied Health Professions, Dentistry, Nursing and Pharmacy).

While many of these studies are helpful summaries of impact in the area under investigation, some (notably Brauer *et al.* 2019) occasionally seem to treat the body of REF ICS as representative of impact activity, either in their sector or across all disciplines. Stated concerns that certain aspects of a topic are under-represented and therefore under-valued are not always appropriately contextualised with a mention of the caps on numbers for ICS from a given department, and therefore the limitations placed on ensuring a representative spread of impact activities within and across submissions. This illustrates the tension between the purposes of allocation and analysis, as introduced in section 2.1.1.

2.2 Writing impact case studies

The studies discussed in section 2.1.4 focus on content in the sense that they are looking for information in the database, either about REF or about UK impact. This section introduces publications that identified characteristics of ICS themselves, that is, common and transferable ways in which the discipline- or topic-specific content was packaged. It first addresses the types of content that can commonly be found in REF2014 ICS and points out challenges with this (section 2.2.1), before turning to narrative framing (section 2.2.2) and other features of presentation discussed in the literature about REF2014 ICS (section 2.2.3).

2.2.1 Content

In his textbook on research impact, which is in large part based on his experience as sub-panel Chair of UoA3 (Allied Health), McKenna (2021) reports that ICS that "made the most convincing claims to impact showed **causality**, used **measures** and qualitative and quantitative indicators, addressed issues of **attribution** and **contribution**, emphasised **progression** or spread of impact, showed systematic **capture** of impact information and presented a **tailored account**" (2021: 54, my emphasis). This sets out many of the components that are frequently mentioned in studies about assessment through ICS, as discussed below.

The link between research and impact, variously called "causality", "linkage", "attribution", "contribution", seems to have been especially tricky. In essence, this refers to pathways from research to impact and attribution of effects to research. For example, Watermeyer and Hedgecoe (2016: 655) observed reviewers at a pre-2014 REF mock impact exercise at Cardiff University and set out three key problem areas:

- Attribution – how much of the impact is due to this specific research? Did authors feel that they had to inflate their contribution to give the impression that it was theirs that made the difference to an outcome? This is also reflected in the concept of "sufficient" vs "necessary" causation introduced by Reed *et al.* (2021: 3). Terämä *et al.* also call attribution a "challenge" (2016: 2).
- Timelines – there was sometimes ambiguity around the eligibility of whether underlying research could be claimed (this was subsequently addressed by the modified template for REF2021) and the plausibility of it having made a difference. Watermeyer and Hedgecoe emphasise the importance of showing the sequence of events for a convincing claim in "precise and accurate timeframes" (Watermeyer and Hedgecoe 2016: 655).
- Link to research – it was recommended that this be expressed through a "history of change" (Watermeyer and Hedgecoe 2016: 655). The link to research is also problematised by Penfield *et al.* (2014: 29), who suggested that such claims could be strengthened by adding metrics.

Bonaccorsi *et al.* (2021) provide a helpful description of the difference between "attribution" and "contribution": the former implies a "causal allocation" being possible with clear delineation, whereas the latter emphasises "partial, empirically observed, participation in a

dynamic process" without the ability to control for the individual factors at work in a given impact and therefore no means of measuring them (2021: 5). The terms are also discussed in detail by Morton (2015: 406) in her proposal of a "research contribution framework" as an adaptable tool for impact assessment especially in knowledge exchange contexts.

Testimonials were seen as essential for establishing the research-to-impact link by the assessors surveyed by RAND, who generally reported that they found it difficult to attribute impact to research and commented that sometimes links looked tenuous (Manville *et al.* 2015a: 33). Similarly, Loach *et al.* (2016: 6) showed that this type of evidence was not only the most frequent type, but also associated with higher scores in Main Panel D based on grade point average, although it was associated with lower scores in Main Panel A. As discussed below, Gow and Redwood (2020: 89-94) also include testimonials in their eight characteristics of successful ICS.

There are also various views on using quantitative information within ICS, beyond the wider discussion on whether narratives should be used in the first place. In the HEFCE-commissioned report on the impact component of the 2014 REF, Grant (2015) calls for standardised metrics to be used in the ICS narrative, but rather than suggesting that these should be used to enhance the assessment, this is in the context of facilitating post-assessment analysis of the sector. The report suggests the use of standard definitions in addition to standardised metrics, to avoid situations where different ICS use different abbreviations to mean the same thing (e.g. names of funders written in full or as acronyms). One of the stated aims of this report was to determine whether a metric could replace the narrative approach after all, as discussed in section 2.1.2. The analysis provides support for the view that these narratives could not be replaced by a system built around metrics, because there was an astonishing variety of both impacts reported and metrics used in the ICS (Grant 2015: 72). To help with comparability and post-assessment analysis even of a narrative-based snapshot of impact, that is, an ICS, HEFCE then commissioned RAND to prepare a separate report on using quantitative indicators in REF2021 based on the 2014 submissions (Parks *et al.* 2018). This report explicitly stated that it is not prescriptive. Rather, it was designed to help the preparation of ICS in order to facilitate post-assessment analysis but left it to authors' discretion to use the suggested formats. This is consistent with a previous report (Manville *et al.* 2015b), in which RAND analysts concluded that ICS remained

the most appropriate way for showing impact because they allow submitting units to show a greater range of impacts.

From the perspective of REF assessors, Greenhalgh and Fahy (2015) offer a different view of metrics. They acknowledge that while assessors in UoA2 (Public Health) appreciated metric indicators in the ICS, the panel leadership did not support this to the same extent (Greenhalgh was deputy Chair of sub-panel 2 in REF2014). Rather, they commend the narrative format for allowing authors to express "passion for the topic", explain "why it mattered" and contextualise the "moral" significance, none of which would be possible in a system based on metrics (Greenhalgh and Fahy 2015: 7). Similarly from a different disciplinary background, based on an analysis of ICS reporting on exhibitions, Brook (2018) welcomed the move away from metrics because quantitative audience data cannot account for the qualitative impact that art may have on audiences. Chowdhury *et al.* (2016: 13) call for information to be presented "explicitly" and assert that quantification helps in order to prepare "better quality ICS". It is somewhat unclear what they mean by "better quality"; in the context of their article, it seems to mean that an ICS is easier to score because it is more metricised. However, this can be called into question, as it is unclear by what measure an ICS would be "better quality". For example, this could mean getting a higher score, having better quality impact, or being easier to read and/or to score.

The most comprehensive study of high-scoring ICS from REF2014 is Gow and Redwood's (2020) book-length examination of ICS submissions that achieved a rating of 4* for 100% of the submissions. This is a similar approach to sampling to the one taken in this thesis (see chapter 4), and therefore their conclusions are described here in some detail (for a review of the whole book, see Reichard 2021). The main difference in the sample selection is that Gow and Redwood focus only on high-scoring ICS in order to identify "the elements associated with top-level research impact" (Gow and Redwood 2020: 65), whereas I compare high-scoring and low-scoring ICS in order to pinpoint differences between the two sets (both in this thesis and in the overlapping article Reichard *et al.* 2020, which is reviewed extensively in Chapter 5 of Gow and Redwood 2020).

The eight characteristics of "world-leading, 4* impact" proposed by Gow and Redwood (2020: 76) are:

1. Long-term research and impact context

In many high-scoring ICS, research foundations were established long before the assessment period, then perhaps they influenced policy at the start of the eligibility period and were able to evidence the benefit of implementing it just before the closing date. With many ICS, some of the research was published a decade before the relevant REF evaluation, as has also been highlighted for the ICS dataset of REF2021 (Wagner *et al.* 2024: subfigure c on each summary figure).

2. Research funding

Gow and Redwood assert that "[t]here is a strong correlation between winning grants and conducting research of a type and calibre that might well translate into some wider social, economic, cultural or other human benefit" (2020: 80). As discussed above, this was not necessarily intended or expected by the designers of the research impact assessment exercise (see section 2.1.2a), but is a finding in their analysis: often, there was a "cluster of funding" (2020: 80), such that a research programme had QR funding from their university but also funding from different sources such as government departments, charity funders, or more broadly research users or beneficiaries. At the same time, not all 2014 ICS explicitly mentioned funding, but the authors assume that some of these must have had it, based on the sheer scale of the project (2020: 82).

3. Role of the researcher in the implementation, often embedded

Researchers played different roles in "translating the research into impact", which may be anywhere between "contributing" and "leading", up to being researcher-practitioners, especially in the Arts (Gow and Redwood 2020: 83). The common observation is that in the sample, the link between the research and the impact was clearly stated, and that the involvement of the researcher or researchers (such as positions in committees or consultancy services) was not claimed as impact in its own right – rather, it led to specific benefits that arise from the involvement. Examples provided are:

- UoA23 York *JobCentre*: The whole research project is embedded within relevant government department, providing policy papers.
- UoA25 Sheffield *HE*: University provides consultancy and serves as member of official bodies.
- UoA29 Kingston *Leveson*: Drafting legal text, advising a House of Commons Select Committee.
- UoA35 QMUL *Performance*: Researcher as artist, curator and activist.

4. Commitment to impact (by beneficiaries or third party), e.g. financial or other resource

Resources are committed by the beneficiaries as "indication of the benefit and their welcome for the research" (Gow and Redwood 2020: 86). Unlike research funding (see Characteristic 2), this refers to situations where a beneficiary or other organisation makes available resources such as money, space, equipment or staff time to facilitate the generation or longer-term sustainability of impacts. Examples from various ICS mentioned by Gow and Redwood (2020: 86-89) include:

- Government funding or investment in programmes such as information campaigns, or even the establishment of new official bodies.
- Spin-out companies (also identified as a factor in high-scoring ICS by Reichard *et al.* 2020: 10) which provided new jobs and sales, especially in medicine and applied science. In some ICS, narrating the creation of a spin-out even seemed to take priority over articulating the actual benefit of a product.
- Gow and Redwood also include sales figures in their list of evidence of a third party's "commitment to impact", with the assumption that investments have been made into creating something that can be sold (whether that be products or, in many cases, books or exhibitions).

5. Quotes in the case study text

Referring to the use of corroborating sources in the main text of an ICS, Gow and Redwood assert that "the presence of quotation – or, even, close reference – was a strong indicator of the high-calibre impact being described" (2020: 94). This positive view of quotations especially from testimonials is also echoed in Dunlop (2018: 285), whose study focuses on UoA21 (Politics and International Studies). However, the use of direct quotation from corroborating sources is a feature of the presentation of impact in an ICS, rather than of impact itself. Without comparison to lower-scoring ICS, it is difficult to see how the authors can maintain that quotations were an "indicator" of impact that deserved a high score. Indeed, Reichard *et al.* (2020: 12), which includes such a comparison, found that testimonials were over-used in some low-scoring ICS.

6. Breadth and range of impacts, cumulative effect; "density of the case studies"

Following the finding in Reichard *et al.* (2020: 8) that high-scoring ICS reported on average 2.8 different types of impact (with potentially several discrete benefits of the same type), Gow and Redwood elaborate that "[i]mpact is rarely one single or simple outcome" (2020: 95). Instead, the kinds of impacts described in 4* ICS were often developed through several major steps, or through "a cascade of smaller features" (2020: 96). Either of these structures could also be an account of how several different parties (e.g. research users) have further developed or implemented one development from the research.

7. Something new or transformative for beneficiaries

Gow and Redwood (2020: 98) introduce the "nothing-creation-something-impact trajectory" to emphasise that the essence of research impact is to create something "transformative". This is especially evident in impact from medical research (e.g. UoA1 Bristol *Health Benefits*, UoA2 Bristol *HIV*) but also in Music and Theatre (e.g. UoA35 Goldsmith *Afghan*).

8. News media or public engagement

This final characteristic is introduced as the "most challenging and […] surprising" one (Gow and Redwood 2020: 106). The authors concede that it was not as extensively present as the others: while "media engagement was prominent in many top-level studies", there were 36% "making no evident reference to media" (2020: 107). They speculate that one reason for the frequent mentioning of media engagement in 4* ICS might have been a general assumption that it should be part of impact, despite the official guidance specifying the opposite (2020: 108). They conclude that media engagement was not strictly necessary for a 4* rating, but rather "supplemented otherwise excellent material" (2020: 113).

Having discussed their eight characteristics, the authors express two caveats: not all 4* ICS must share all of these features, and the features are not a template for 4* impact. An additional caveat not explicitly acknowledged in the book is that their aim to analyse the characteristics of "world-leading impact" (Gow and Redwood 2020: 76) is limited by the focus on REF. Relatedly, it is sometimes overlooked, including by this book, that the grade descriptors for impact in REF2014 did not include the term "world-leading", which was used in the 4* descriptor for the output component; rather, the descriptor for a 4* ICS is "outstanding impacts in terms of their reach and significance" (HEFCE 2011: 44), which does not have to correlate to "world-leading". Perhaps the biggest limitation of Gow and

Redwood's (2020) study is the exclusive focus on 4* ICS without comparison to lower-rated ICS. This makes it impossible to ascertain whether the identified elements are typical for the top-level compared to others. Even if they are characteristic of that body of texts, they may not be distinctive. This comparison with a control group is an important part of a register analysis (Biber and Gray 2016: 57) and is therefore included in the present study.

2.2.2 Narrative framing

Having discussed common features of the kind of content that might be found in ICS across disciplines, this section turns to the ways in which such content is arranged within an ICS. This has been researched from the perspectives of Sociology (Bandola-Gill and Smith 2022) and Linguistics (Wróblewska 2021), and the respective studies are now introduced in turn.

Bandola-Gill and Smith (2022) investigated the overall narrative structure of 66 ICS from REF2014, from the three highest- and lowest-scoring submissions from three UoAs across the academic spectrum (UoA2 Public Health, UoA23 Sociology, UoA30 History). They identified three components of ICS narratives: plot, moral, and hero. Of these, "plot" refers to the linear sequence of most ICS and the fact that they focus on specific aspects of the story. "Moral" is the term they chose for a specific outcome, in high-scoring ICS mostly a quantifiable "proof" to add some "objectivity" to a claim of change compared to the original situation (2022: 1865), and "hero(es)" is how the individual researchers are portrayed as they carry the responsibility for the claimed impact.

Separately, they identified four narrative types, that is, the way in which the components were framed: problem-solving, tool-building, public engagement, and reframing. Of these, "problem-solving" was the most frequent type. It presents information in a linear way, leading from an existing problem via a research-based solution to an improvement of the original situation. Bandola-Gill and Smith highlight that this type "often involved specific forms of knowledge production" (2022: 1868) such as consultancy, evaluations or participatory research. The next most common type, "tool-building", is a sub-type of problem solving with a very specific solution, where research activity provides a tool for decision making, such as a model or a database.

The remaining two narrative types were relatively more frequent in the low-scoring ICS in their sample, especially in more applied disciplines (in Becher and Trowler 2001's typology). In the case of "public engagement", characterised by collaboration to produce and/or disseminate knowledge, this was due to low-scoring ICS often missing the "moral"

component, that is, there was no (well-evidenced) claim to a clear, positive outcome. The least common type, "reframing", was found in ICS where the impact was a change in framing or an adaptation of an existing conceptualisation of, for example, policy problems or climate change. Based on interview data, the authors note that this type seems to have been seen as too "risky" (2022: 1868) by universities and was excluded from REF submissions for that reason, not because it was actually an uncommon type of impact in these disciplines.

Bandola-Gill and Smith (2022: 1866) express surprise and imply disappointment at the fact that they were able to match all 66 ICS in their sample to one of only four narrative types, indicating a certain standardisation of how impact was expressed and therefore how it was conceptualised in the first place. However, they do note that this is "standardising *format* rather than *content* within the performance measurement" (2022: 1868, italics in original). In line with suggestions discussed above (section 2.2.1), including by Grant (2015) and Greenhalgh and Fahy (2015), a certain degree of standardisation in format, such as is encouraged by the REF ICS template, could facilitate more variation in content, as assessors may be able to locate the various components of the impact claim more easily.

A linguistic study of narrative structure is Wróblewska (2021), which dives deeper into the genre and narrative structure of ICS in UoA28 (Modern Languages and Linguistics) in the context of the formalisation of the impact agenda. She situates ICS as an emergent genre that is heavily influenced by the five-section template provided by HEFCE and by its context of high-stakes assessment, implications of which are discussed below in section 3.1. Her datasets are 78 ICS and transcripts of 25 hours of interviews with authors of those ICS, including both researchers and impact professionals.

In her analysis of the narrative patterns of ICS, Wróblewska found that they tend to follow the "Situation – Problem – Response – Evaluation" (SPRE) story-telling pattern (2021: 5). To this, she added two elements specific to ICS: "further impact" and "further corroboration". This narrative pattern promotes a linear vision of the research-to-impact journey. While Wróblewska calls "situation" and "problem" "underplayed" (2021: 7), from an impact assessment perspective these two components are a backdrop to evaluating the "response", that is, the impact, and are therefore secondary to the response and its evaluation. This pattern corresponds to the first of Bandola-Gill and Smith's (2022: 1868) narrative types, "problem-solving". This is not surprising since this pattern was the most frequent one found in their sample.

2.2.3 Style

Beyond the overall narrative structure, several sources comment on the writing style of ICS and, in some cases, the extent to which this may have influenced assessment. In doing so, some employ rather conceptual labels, which are difficult to operationalise by those writing the ICS. One study that comments extensively on the "structure and style" is Watermeyer and Hedgecoe's (2016) article, pointedly titled "Selling impact". Their discussion is based on observing reviewers at a pre-2014 REF mock impact exercise at Cardiff University, who in turn referred to the way that ICS were written in their conversations about scores. On the basis of these discussions, the authors recommend a style that is "a hybrid of narrative lyricism, dynamism and informational efficiency" which is "aesthetically pleasing yet sufficiently functional", so writers should "avoid the prosaic yet resist the florid" (2016: 6). In the end, they recommend "a compelling but unfussy style" (2016: 6) and "a dynamic style of narrative writing" (2016: 12). This flurry of style descriptions implies that the language should not be "dumbed down" (wording borrowed from McKenna 2021: 54), while nevertheless being "suitable for a lay audience", making impacts "both obvious and explicit" (Watermeyer and Hedgecoe 2016: 6). Another guiding principle can be found in their repeated reference to "ease-of-evaluation" and the need to facilitate "panellists' fluency in making decisions" (2016: 6).

Similarly, McKenna, sub-panel Chair for UoA3 in both 2014 and 2021, concludes that "the trick is to make it easy for the assessors" (2021: 18). He lists a "lack of coherence and a dense narrative" as hallmarks of a weak ICS, alongside the observations that in many low-scoring ICS a "journalistic style had unnecessarily 'drummed up' or 'dumbed down' the narrative" (McKenna 2021: 54). Conversely, high-scoring ICS had "an articulate, well-written and interesting story". He also recommends an "active authorship style" and "understandable English … with concise sentences, rather than complicated scientific jargon" (2021: 18). This is echoed by Watermeyer and Hedgecoe (2016: 7), who note an overuse of jargon and reliance on disciplinary convention especially in Main Panels A and B, and suggest that authors in these areas should get help from external writers. Neither publication elaborates on what constitutes a "concise" sentence.

Pointing to overview reports from the REF2014 Main Panels, Chowdhury *et al.* (2016: 4) invoke "the lack of academic language and emphasis" as a hindrance for the panel to apply the assessment criteria. This could be interpreted in a similar way to McKenna's observation

about a “journalistic style” being unhelpful (2021: 18), but they leave open what they mean by “academic language”. Arguably, the “complicated scientific jargon” against which McKenna (2021: 54) cautioned could be seen as part of “academic”, as opposed to “journalistic” language. This possible contradiction illustrates the complexity in referring to overarching domains of writing as style recommendations without further elaboration. Only a few studies comment on individual words that could form part of a certain style. For example, Cain and Allan identify “research verbs” that were used frequently in high-scoring ICS in UoA25 (Education): *investigated, explored, tested, measured, tracked* (2017: 724). McKenna (2021: 23) recommends verbs such as *transforming, improving* for ICS titles, but it is unclear on what basis he does this other than being a sub-panel Chair. Having analysed the titles of all 166 ICS in UoA21 (Politics), Dunlop (2018: 273) found that *improving* as the most frequent one accounted for 11% of all verbs used in titles, followed by *informing*, *shaping* and *influencing*.

In the HEFCE-commissioned overview report, Grant (2015: 72) notes that ICS are written in a “style and tone that aims to ‘sell’ the impact to the assessment panels”. Similarly, and based on the same analysis, Hinrichs *et al.* remark on the “universally positive sentiment in the language used” in the ICS (2015: 2), but neither publication elaborates on what this “selling” or “positive sentiment” looks like. They note this positive slant as a problem in the context of using the dataset as a resource for analysing UK research impact as a whole; there is no evidence that they see this as problematic for the assessment itself. In addition to the competitive nature of an assessment context, the “positive sentiment” may be related to documents that precede the ICS: Derrick *et al.* (2014: 153) mention that compared to academic articles, REF2014-related policy documents may contain more “language designed to ‘convince’ the reader”. Similarly, Brauer *et al.* (2019: 66) contrast the ICS, where researchers are required to “boast about their research”, with more traditional research roles and related writing, where academic authors “went through great pains to disassociate themselves from taking sides […] in order to establish their integrity”. The differences between research articles and ICS as two distinct registers are discussed in detail in section 5.1.

Much of this discussion in the literature is subjective and not linked to assessment scores, yet the accusation that the language may have influenced the score shines through. Based on a linguistic analysis, Hyland and Jiang (2023: 2) assert that ICS were “hyped” to an extent

that the "usefulness and reliability" of assessment was in danger. However, assessment scores were not a criterion in their selection of texts, and therefore their article provides no evidence that this "hyping" was actually related to the assessment outcome, whether in a causal way or not. Similarly, Terämä *et al.* (2016: 14) assume that language may have made a difference and suggest an analysis of language at institution level to show "a more nuanced understanding of 'successful' submissions".

Most notably, on this question of whether style and presentation had an influence on score, the RAND report on assessing impact submissions (Manville *et al.* 2015a) notes that in some cases the overall presentation made it difficult to draw out "the substance" of an impact claim which the assessor felt was in there somewhere; assessors noted that they "were aware that presentation affected their assessment of the impact" (2015a: 39). It is therefore important to determine how ICS writers can avoid a situation where recognition cannot be given for impact whose substance is impressive but which is not clearly enough visible.

Similarly, Watermeyer goes as far as to assert that "[p]anellists' role in evaluating impact would […] appear to have focused more on an assessment of impact *narrative* than an assessment of impact *claims.* In other words, the **stylistic** rather than substantive achievements of REF2014 ICS **assumed precedence**" (2019: 80-81, italics in original, bold my emphasis). This thesis contributes to evaluating the claim that "stylistic achievements" took priority in the assessment process by investigating different elements that might contribute to such "stylistic achievement".

2.3 Chapter summary

In this chapter, I have explained the background of REF and examined its stated purpose. I outlined the problems that would arise from a (hypothetical) assessment system geared towards providing complete accountability (section 2.1.1), and therefore argued for a system where a selection of research impacts is put forward for assessment as opposed to complete coverage of impacts of all research. Of the various approaches proposed for obtaining or creating a number that could be used for funding allocation through a formula, while at the same time limiting the negative effects on academic life and freedom of assessing for complete accountability (section 2.1.2), narratives were chosen for REF. This assessment method has been discussed in the literature as broadly successful (section 2.1.3), but concerns have been raised about the role of the presentation of ICS in the assessment process.

In order to test the suggestions that presentation affected the assessment process to the point of potentially having compromised its integrity, a comparison is needed between the presentation of high- and low-scoring ICS respectively. However, while the ICS database has been used for many different analyses (section 2.1.4), very few studies have focused on aspects of presentation, and those that include presentation did not compare specific features of high- and low-scoring ICS (section 2.2). The present study aims to address that gap with the following research questions, first introduced in section 1.2 above:

1. What features related to the *presentation* of the research, pathway and impact, as opposed to the stated criteria of *significance* and *reach* of the impact and the clarity of *attribution* to the research, may be characteristic of high- or low-scoring ICS and therefore may have influenced the score?
2. What linguistic markers of persuasion and evaluation do ICS feature, and does this differ between high-scoring and low-scoring ICS?

It is important to note that my aim is not to posit a causal relationship between presentation and scores, nor to argue that presentation has negatively affected the integrity of the assessment. It is to investigate whether differences in the presentation of ICS, including the role of persuasive and evaluative language, can be detected through a range of methods of linguistic analysis, as a prerequisite for approaching the question about influence. The overarching question of the thesis is therefore:

"To what extent **could** presentation have influenced the scoring of impact case studies in REF2014?"

If little difference can be detected with a range of methods, then the possibility that differences in presentation had an influence on the star rating would seem to be smaller. If differences were uncovered, then such findings might contribute to increased fairness in future assessment rounds because their availability to writers across universities could reduce the gap between those that have the means to invest in more professional support and those that rely on non-specialists for the preparation of ICS.

In the next chapter, I introduce several linguistic frameworks for analysing the presentation of ICS, conceptualising ICS as a separate register within academic writing. This enables the selection of register analysis as the overarching framework, under which I will include ways of investigating different aspects of presentation that are applied in chapters 5 and 6. Different ways of researching the language of persuasion and evaluation are introduced in sections 3.2 and 3.3, which will inform the analyses described in chapters 6 and 7.

Chapter 3 Impact Case Studies as a Persuasive Register

This thesis aims to add to our understanding of REF impact case studies, and specifically the language of those texts. One way to describe the nature of texts is a register analysis. This is a text-linguistic approach, focusing on units above the sentence (Biber 2019). Register analysis is often associated with the multidimensional analysis approach developed by Biber (1988). In that approach, a large number of language features that can be tagged automatically (as outlined in section 4.4.3) are compared across text from different registers in order to define dimensions along which groups of texts vary. This approach is best suited to large amounts of text, so it was not chosen as a principal method for the present study. However, it was used at an exploratory stage to see where ICS sit in relation to previously described registers, which confirmed the applicability of aspects of the research questions and informed the final choice of methods. This initial comparison of ICS to other genres along these dimensions will be referred to where applicable in this chapter.

Part of a register analysis is to systematically check "preconceived notions" of what a given register is like (Biber and Gray 2016: 91). For ICS, some assumptions and assertions that appear in the impact literature were introduced at the end of the previous chapter, where I indicated that these assumptions are contradictory and warrant a systematic analysis. This thesis is therefore a study of the register of ICS, combining various frameworks and methods that are best suited to address the research questions derived from the literature review in chapter 2.

Section 3.1 describes the concept of register in more detail and relates it to ICS, before outlining the various analyses I conducted to address research question 1:

What features related to the presentation of the research, pathway and impact may be characteristic of high- or low-scoring ICS and therefore may have influenced the score?

Section 3.2 then turns to persuasion and evaluation and how these concepts relate to each other, by discussing relevant research approaches that could be chosen for addressing research question 2:

What linguistic markers of persuasion and evaluation do ICS feature, and does this differ between high-scoring and low-scoring ICS?

Finally, section 3.3 introduces the Appraisal framework developed by Martin and White (2005) as the main method for researching evaluative language in ICS as set out in research question 2.

3.1 Register

To start, it is necessary to distinguish the term "register" from related terms, define how I use it and delineate this register analysis from other approaches. The concept of *register* as a characterisation of text is related to the concepts of *genre* and *style*. Each of the three terms is used with different meanings in different linguistic traditions, and I follow the terminology most associated with Douglas Biber. In works in his tradition, a *register* is seen as being defined by linguistic features mostly on the level of lexico-grammar that are (a) frequent and pervasive in a text and (b) related to communicative function (Biber and Conrad 2019: 6). By contrast, a *genre* is defined through textual features that may or may not be pervasive, such as a greeting at the beginning of a letter (a characteristic of the genre but not something that appears throughout the text), which is why research on genre typically uses complete texts, rather than excerpts. Matters of *style*, finally, may be pervasive but are more related to the writer's preference than to the context of the text and its social and communicative function (Biber and Conrad 2019: 2).

Register studies bring together diverse areas of linguistics (Gray and Egbert 2019). One of these is variationist (socio-)linguistics, which can in some cases involve the study of register. However, it differs from the text-linguistic approach of register studies in important ways: Variationists in the tradition of Labov investigate "the way language users choose between semantically and functionally equivalent variants" (Szmrecsanyi 2019: 76), that is, differences in language are of interest because they are based in non-functional choice. Moreover, variationist studies tend to identify particular alternatives (a pair or set of features) and analyse the factors that influence the occurrence of each instance (Szmrecsanyi 2019: 77), whereas register studies that are concerned with functional language differences tend to include a greater number of features (up to ca. 150, Biber 2019: 55) and analyse these at the level of a (sub-)corpus or, preferably, the text (Biber 2019: 54). A further distinction needs to be made between register studies and other sociolinguistic approaches which investigate linguistic variation as "indexical or purely conventional" (Biber 2019: 45).

The terms *register, genre* and *style* are not only used in varying ways within linguistics; they are also used outside of linguistics, often with unclear definitions. In fact, in these broader contexts, *genre* and *style* are more commonly used than *register*, so even though the term *register* is not discussed in the impact literature, we may expect to find more explicit discussion of the *genre* or *style* of impact writing. This is the case for *style*, as described in chapter 2 and contextualised below, but not for *genre*: while the impact literature contains frequent general references to the "narrative", there seems to be little interest in a specific sequence of information that can be generalised across texts ("moves" of a genre, in the terminology of Swales 1990).

The only study of specifically the genre of ICS is Wróblewska (2021), who conducted a Swalesian genre analysis of 78 ICS in UoA28 (Modern Languages and Linguistics) by investigating the social context, narrative pattern and linguistic features. She positions the development of the ICS genre as a component of the wider development of an "impact infrastructure" in and around universities, that is, the creation of processes and human resources to support the generating, evidencing and presenting of impact for REF (Wróblewska 2021: 4). Genre is introduced as a social and functional construct, whereby the function of ICS is clearly persuasive: they are defined as a "performative, persuasive genre – its purpose is to convince the 'ideal readers'" (Wróblewska 2021: 5). This purpose then determines the characteristics of the genre (narrative patterns, grammatical and lexical features) and gives it its "seductive/coercive force" (2021: 7).

Wróblewska (2021: 5) emphasises that the genre was introduced top-down, with a template and guidance documentation provided by HEFCE (2011). From that point on, the preparation for the REF submissions happened across universities in parallel, with simultaneous submission by everyone working on the genre and little opportunity for it to grow organically in response to its function, outside of the rigid REF cycle. The narratives predominantly follow the "Situation – Problem – Response – Evaluation" (SPRE) story-telling pattern, with the two ICS-specific elements "further impact" and "further corroboration" (2021: 5). The grammatical and lexical features of ICS that Wróblewska (2021: 5-7) observes in her dataset include, among others: a focus on numbers and size, including words such as "major"; positive words, especially in testimonials; novelty and a focus on originality; the use of lists. These features also appear in my analyses, as described in chapters 5-7, and will therefore be discussed there.

One explanation for the lack of interest in ICS as a *genre* defined by narrative moves may be that it is heavily constrained by the 5-part HEFCE template, which allows little variation. This top-down origin makes an analysis of the genre less pressing than an analysis of those aspects of the language over which writers do have control. However, there are some ways in which the non-pervasive genre markers, that is in this case the structure of the template, influence the register without being part of the register. This includes the artificial distinction between research and impact, which determines which kind of content is placed where, and the tight space constraints, which are also identified by Biber and Gray (2016: 42) as having an effect on the register of science writing. It is therefore reasonable to assume that the strict page limit also influences language choice in ICS.

Contrasting with the relative lack of interest in ICS as a *genre*, several studies on REF2014 ICS comment on the writing *style* of ICS. As these are not linguistic studies, they employ rather general labels, which are difficult to operationalise by those writing the ICS, as described in section 2.2.3. As explained in that section, the style of ICS is described on the one hand as not academic enough or too simplistic, and on the other hand as too complex and jargon-heavy. This discrepancy between the views of the commentators indicates that there was no clarity or consensus about what style was expected, and since this style is "expected" because it serves a specific communicative purpose, it is in fact the *register* that these sources comment on, rather than subjective and more arbitrary notions of *style*, in Biber's terminology (2019: 2). Therefore, studying the register of ICS would bring some much-needed clarity about the language needed or expected for texts in the situational context of REF ICS.

The situational context of the REF includes both the explicitly stated purposes as discussed in section 2.1.1 (namely, assessment and allocation) and the implicit pressure put on these texts by suggestions that they could be replaced with metrics (see e.g. the discussions in Grant *et al.* 2010; Khazragui and Hudson 2014; Smith *et al.* 2011) or algorithms and machine learning (Ravenscroft *et al.* 2017). In order to examine whether there is support for the view that text-based ICS can be an appropriate assessment mechanism (as discussed in section 2.1.2), it is helpful to describe the register of high-scoring ICS in a way that can be applied by those creating such texts. One way to do this is to analyse the specific linguistic features of the register: what functions do they have that make them particularly suited to this situational context? By understanding the differences as variations in (functional) register,

rather than in (non-functional) style, we can go beyond the description of language as formal or structural grammar and explain why some features may be preferred over others. This may enable empirically and theoretically better justified guidance for writing in this genre, compared to lists of “power words” (as suggested by Van Noorden 2015: 150) or other guidance that appears to have been informed more by the subjective experience of the authors than by empirical evidence (see section 2.2.3). The question thus becomes: what are the defining characteristics of the register of ICS that scored highly in REF2014?

3.1.1 REF impact case studies and “academic writing”

In order to define the register of ICS, it is necessary to situate it within academic writing more generally, because these texts are produced within academia for an academic audience, describing research (at least in part), and they can therefore be understood as one of the many “academic” registers (see Biber 2006; Swales and Feak 2000, for discussions of the variety of academic or university registers). The definition of “academic writing” is “somewhat elusive” (Caplan 2021: 268), but according to Caplan, the features identified and discussed by Biber and Gray (2016: 79-82, Table 3.1 "Grammatical features that are especially common in academic prose") are accepted in the literature as “reliably differentiat[ing] academic writing from other linguistic registers” (Caplan 2021: 274). Research writing acts like an academic “currency” in that it “represents, discusses, and expands the knowledge of a particular field” (Caplan 2021: 268), and journal articles are often presented as the “pinnacle of the register” (Caplan 2021: 274). By contrast, ICS occupy a rather different space in the field of the writing that is done in the academic world: they are UK-specific in this form, and only a fraction of researchers at UK universities are involved in writing one.

The primary purpose of academic writing, understood as research writing, is to “reproduce and extend disciplinary knowledge” (Caplan 2021: 273), but this is not true for ICS, and if the purpose of the text is different, it can be expected that the linguistic aspects of the register are also different. Despite this difference in purpose, there are important contextual similarities between research articles and ICS, such as the large overlap between the two in both authors and readers if it is assumed that researchers write and read these texts. However, this overlap is not complete: ICS are often written at least partly by staff situated in professional services, rather than the researchers themselves. Similarly, readers of ICS include “research users” (Manville *et al.* 2015a: 11), who are not academics. A detailed

contextual comparison between research articles and ICS is provided in the situational analysis in section 5.1 below.

An overarching linguistic similarity is that some grammatical functions may support the purpose or context of both registers. For example, nominalisations are constructions where, for example, abstract concepts such as processes are expressed as nouns, rather than through constructions with a verb at the centre: "nominalization [is] the process of turning a verb (*nominalize*) or clause (*scientists nominalize heavily*) into a noun (*nominalization*) or noun phrase (*the heavy use of nominalization*)" (Caplan 2021: 268). These and similar constructions can help to transform a text to be more concise where strict word or page limits are in place and where the writers can assume a certain degree of shared understanding with the reader, although again it is unclear and variable to what extent this shared understanding is a safe and valid assumption to make (cf. Manville *et al.* 2015a: 15-21, on whether an assessor could take prior knowledge into account, where present). Such space-saving constructions can therefore be found in both research articles and ICS. As with purpose and context, though, not all grammatical functions are equally important in both registers. For example, causal relationships are more important in ICS because of the need to demonstrate and explicitly argue for the link between underpinning research and impact.

As part of situating ICS within and around academic writing, I conducted an initial comparison of ICS to other registers defined in Biber (1988), using Nini's (2015) MAT tagger and all clearly identifiable 4* ICS from REF2014 (Sample A as described in section 4.3.2 below, process described in section 4.4.3). Figure 1, an output from the MAT tagger, plots eight pre-defined registers and the user's input as a further, new, register on the five dimensions from Biber (1988) along which registers vary:

- Dimension 1: Involved vs Informational Discourse
- Dimension 2: Narrative vs Non-Narrative Concerns
- Dimension 3: Explicit vs Situation-Dependent References
- Dimension 4: Overt Expressions of Persuasion
- Dimension 5: Abstract vs Non-Abstract Information

In Figure 1, the five dimensions are represented along the x-axis. The y-axis shows the z-score, with 0.00 represented by a horizontal line half-way up the figure. The distance between points plotted for each dimension shows where different registers, for which the tool includes set data, sit in relation to each other on each dimension. The figure shows that "scientific exposition" (highlighted in dark green rectangles) is close to ICS ("MAT all high full", highlighted in light blue ovals) on most dimensions, but also illustrates how far away the two registers are on two of the dimensions, namely 3 and 5. Dimension 3 places ICS as an outlier at the top of "explicit references" compared to all other registers in the reference corpus used by the analysis tool, with science writing a distant second. Dimension 5 shows that ICS are much less abstract than science writing, but apart from that, it has higher values for "abstract information" than all other registers. These observations of differences support the suggestion that ICS are a separate register within academic writing. The placement of ICS on Dimension 3 justifies a closer look at these mechanisms, as is provided in section 3.1.3 below.

Figure 1: ICS plotted against other registers on Biber 1988 dimensions (figure generated in Nini 2015 MAT tagger)

3.1.2 Disciplinary differences

Research on academic writing in different contexts has documented differences between disciplines (e.g. Gardner *et al.* 2018; Gray 2015), and it is reasonable to assume that this socialisation into different disciplinary conventions in turn influences how academic writers from the various disciplines approach writing an ICS. Indeed, professional services staff have raised this as a concern through a user survey I conducted in early 2021 (see section 4.2.3 below), asking "how can we support people embedded in certain writing cultures?" This section therefore introduces such disciplinary differences in preparation for a discussion of how the function of a register can influence its form (section 3.1.3).

In their comprehensive and authoritative study on complexity in academic writing, Biber and Gray (2016) offer a novel view on differences across academic disciplines. Starting with the oft-repeated assumption, sometimes even framed as accusation, that academic writing is "obscure" because writers use specialist words to impress other people (resulting in "academese", Biber and Gray 2016: 1), they offer an alternative explanation for this perception: language choices in the register are determined by the wish to achieve efficiency in conveying new information. In describing how this wish is pursued, they initially distinguish between "science writing" and "humanities prose". Their analysis collates features that are identified in the *Longman Grammar of Spoken and Written English* (LGSWE, Biber *et al.* 1999) as typical of academic writing and maps these to certain contexts of academic writing. They find (and explain in Biber and Gray 2016: chapter 3) that some features span disciplines across the academic spectrum, namely those that can be generalised as technical vocabulary, nominalisations, and verbs in the passive voice. Others are more associated with the humanities, namely attributive adjectives and clausal complexity. A third group of features is characteristic of science writing and appears much less in the humanities: nouns (rather than adjectives) as noun pre-modifiers (1), noun+participle as noun pre-modifiers (2), and appositive noun phrases (in list form) that serve to explain another noun or noun phrase, usually added in parentheses in science writing (3). For example:

(1) *Heart* surgery

(2) *Lockdown-induced* anxiety

(3) In four cohorts (*Athens, Keio, Mayo, and Florence*), investigators stated that. . ..

All these features typical for science writing contribute to *phrasal complexity*. They are distinct from grammatical constructions that are more common in general language, such as attributive adjectives as modifiers in noun phrases (e.g. "the *red* house"), and that therefore do not create an impression of complexity to the same extent as those in (1)-(3).

A defining feature of science writing in Biber and Gray's (2016) model is therefore that writing is "structurally compressed" (Biber and Gray 2016: 123) through the use of densely packed phrasal structures, which results in writing that is *in*explicit (see section 3.1.3 below). By comparison, humanities prose is more structurally elaborated, that is, related content is expressed by adding dependent clauses into a sentence. For example, the structurally compressed example in (2) above could equally be written in a structurally elaborate way, as (4) or even (5):

(4) anxiety *induced by a lockdown*

(5) anxiety *that was induced by the lockdown in 2020*

These kinds of construction are more readily and traditionally recognised as "grammatically complex" than the phrasal structures typical of science writing, and in using them, writers might be accused of deliberately obscuring meaning by using "academese" style when they could use a seemingly simpler style. Such accusations do not recognise that it is not arbitrary style that creates this impression, but a functionally motivated academic register (Biber and Gray 2016: 1). A second reason why this perception of complexity is exacerbated in relation to humanities prose is that the subject matter is often something that lay readers expect to have a certain understanding of, and therefore they expect to understand the texts, too. By contrast, most non-specialist readers do not expect to be able to pick up a science research article and read it, because they recognise the subject matter as inherently specialist. This incongruence between expectation and actual understanding of the subject matter in humanities disciplines makes people reach for the explanation, or even accusation, of "academese" language, rather than acknowledging the content and therefore the texts as specialist.

With this dichotomy of complex lexis and structural compression in science writing, and grammar that is recognised as complex and structural elaboration in humanities prose, we can ask where ICS sit generally in the range of different kinds of academic writing. In the broad context of research question 1, we can further ask whether those disciplinary

differences in presentation are also found in ICS from different Main Panels with different writing traditions, or whether ICS are more homogenous than research writing. The disciplinary differences and lack of agreement on a general, trans-disciplinary, "academic style" may also partly explain the discrepancies found in the impact literature which describes ICS as either too academic or not academic enough, depending on perspective and disciplinary expectations (see section 2.2.3 above).

3.1.3 Degrees of explicitness

Biber and Gray (2016: chapter 4) illustrate how the function of a register can influence its form, and therefore how the disciplinary differences described in the previous section arose historically. The increasing specialisation of disciplines and sub-disciplines in academia changed grammar norms in science writing away from clausal complexity (structural elaboration, i.e. more words, as in (5) above) towards increasing use of phrasal structures (structural compression, i.e. fewer words, as in (2) above). They explain the "need for greater 'economy'" (Biber and Gray: 129) in writing through increased specialisation fuelled by an information explosion: centuries ago, a generalist scientist would read what any other scientist wrote, and their writings would therefore contain enough explicit information to be understood by people in other fields. By the end of the 20th century, a scientist would read ever more specialist writing that is encoded in a kind of shorthand specialist terminology. In this new environment, efficiency takes precedence over explicitness because for that specialist reader, the level of clarity of that shorthand is sufficient – they can decode the meaning. (Of course this may cause problems for the novice reader in the discipline, as well as for many other potential readers.) The assessors of ICS are specialists to a certain degree, but see section 5.1 below for a closer discussion of the extent to which shared knowledge can be assumed.

In explaining the nature of 20th-century science writing, Biber and Gray (2016: chapter 4) link different types of grammatical complexity with different degrees of explicitness. While *clausal* complexity as in (5) above can make meanings explicit through the use of lexical linking expressions, *phrasal* complexity as in (2) above (a less-recognised type of grammatical complexity) is maximally *in*explicit because the dense packing of information in noun phrases or similar phrasal constructs does not use overt grammatical information in the same measure that clausal structures use. This loss of explicitness is hypothesised as an "unintended consequence" of maximally efficient use of word counts and space constraints

(Biber and Gray 2016: 42). Such phrasal discourse is therefore distinctive for texts "with highly informational purposes and specialised audiences" (Biber and Gray 2016: 40). The purpose and audience of ICS, and the extent to which these are "informational" and "specialised" respectively, will be assessed in detail in the situational analysis presented in chapter 5.

This difference in explicitness across disciplines is especially pertinent to ICS, which are often written or at least supported by impact officers working across different units of assessment and who therefore have to adapt to different disciplinary preferences, while operating under tight space constraints. They have a range of choices for encoding relationships between claims. The scale as illustrated in examples (6) to (9) ranges from finite dependent clauses which encode the most explicit information, including tense (6), via non-finite clauses with less information encoded in the verb (7), to dependent phrases that include almost no overt information on the relationship between the phrase and its referent (8) - (9) (examples (8) and (9) from Biber and Gray 2016: 238):

(6) She works hard *because she wants this qualification.* [*wants* = finite verb]
(7) She works hard *in order to obtain this qualification*. [*obtain* = non-finite verb]
(8) Farms *in Malaysia* [*in* = the location of the farms]
(9) Experiments *in agricultural chemistry* [*in* = the academic sub-discipline]

The example in Table 2 illustrates the difference between two options for connecting clauses, namely "linking adverbials" and "punctuation", with different degrees of explicit encoding of information (Biber and Gray 2016: 121).

Table 2: Example of explicit and inexplicit choices in academic writing: Linking adverbials vs Punctuation (adapted from Biber and Gray 2016: 121)

Linking adverbials	**Punctuation**
maximally explicit	maximally *in*explicit
Examples: *however, therefore*	Examples: *colon : semicolon ;*
specify *what* the logical relationship between the two clauses is	indicates *that* a logical relationship exists but does not say what that relationship is
frequent in Humanities and Social Sciences but not Science	frequent in Social Science and Science but not in the Humanities

Note that the linking adverbials listed as examples in Table 2 have contradictory meanings, and a mere semicolon does not unambiguously distinguish which of these or other meanings a reader should infer in a given instance. Therefore, linking adverbials have an important

discourse function, even though syntactically they are often optional and could be replaced by punctuation in many contexts (Biber and Gray 2016: 240).

Given the disciplinary distribution of these features in academic writing, it is possible that writers from a more science-based writing tradition are less likely to make frequent use of linking adverbials in ICS, even though this register has a different situational context than their research writing. It is therefore important to have a clearer understanding of the linguistic features of high-scoring ICS (research questions 1a and 1c), in part to help ICS writers to avoid a situation where recognition cannot be given for impact whose substance is impressive but which is not clearly enough visible because logical links are not made sufficiently unambiguously explicit.

3.1.4 Components of the register analysis in this thesis

An analysis of register starts with the situational context of a text, before moving on to the linguistic analysis (Biber and Gray 2016: 247). This is because, as illustrated in the previous section, the situation in which a register is used has an effect on the linguistic features, and its analysis therefore takes precedence. However, the process of analysis (situational vs linguistic) is often iterative, such that findings from a linguistic analysis can be used to refine the situational analysis, for example to highlight a certain sub-purpose of a register that may be hidden to the observer or some participants of the register (such as writers or readers), but is uncovered by the otherwise unexplained prevalence of certain linguistic features (Biber and Conrad 2019: 71-72). The final component of a register analysis is then to compare the main two analyses in order to offer explanations from the situation for the prevalence of language features in registers with those circumstances.

Bridging situational descriptions and quantitative linguistic analyses, the framework of register analysis is well suited to addressing research question 1: *What features related to the presentation of the research, pathway and impact may be characteristic of high- or low-scoring ICS and therefore may have influenced the score?* The register analysis in this thesis has the following components:

Section 5.1 reports on a detailed situational analysis that compares ICS to research articles by juxtaposing the following characteristics, as suggested by Biber and Conrad (2019: 40):

a. Participants, including authors, readers and third parties

b. Relations between participants, including the degree of interactiveness, social roles and power differentials, and the degree of shared knowledge
c. Processing circumstances, including production/writing and comprehension/reading
d. Communicative purpose, including general and specific purposes and the purported factuality of the register, as well as expressions of stance
e. Channel, Setting and Topic are characteristics that are included in Biber and Conrad's (2019) framework but are only discussed briefly because they are seen as less relevant for defining ICS.

The situational analysis is followed by a number of other analyses, most of which use linguistic methods. Section 5.2 zooms into the processing circumstances, as well as introducing findings from a quantitative analysis of grammatical features of ICS. This is complemented in section 5.3 by two analyses relating to the content of ICS. Chapter 6 presents findings from a lexical analysis that identifies themes that are mentioned frequently in high- and low-scoring ICS respectively. It then moves on to separating lexical differences into those over which writers and editors have control and those which are pre-determined or restricted by the subject matter and therefore do not contribute to a description of language choice in this register.

The analyses that follow in the final part of chapter 6 and in chapter 7 are related to the communicative purpose of ICS, as defined in section 5.1 based on the literature reviewed in chapter 2, in that they examine the degree and nature of persuasive and evaluative language in ICS. These chapters address research question 2: *What linguistic markers of persuasion and evaluation do ICS feature, and does this differ between high-scoring and low-scoring ICS?* The frameworks for these analyses are outlined in the remainder of this chapter.

3.2 Persuasion in impact case studies

In the context of high-stakes assessment, Penfield et al. (2014) present it as a limitation of the case study approach that "a persuasive, well-evidenced case study" may increase the chances of a higher rating. In a HEFCE-commissioned report, Grant (2015: 72) notes that ICS are written in a "style and tone that aims to 'sell' the impact to the assessment panels". And in her comparison of the UK REF (2014 dataset) and the Norwegian evaluation exercise Humeval (2015-2017), Wróblewska notes that compared to their Norwegian counterparts, the British REF ICS are "highly persuasive in their nature [… and] tone" (Wróblewska 2019: 32). Clearly persuasion is a main implicit function of ICS. However, persuasion may be

expressed in very different ways, and its relationship to the cognate concepts of evaluation and stance will be clarified in this section.

3.2.1 Approaches to persuasion

In a recent study on how persuasion is realised linguistically across specialist genres, Dontcheva-Navratilova (2020) explores four genres (research articles, corporate reports, religious sermons and technical user manuals) across two languages (Czech and English). Starting with the claim that "all communication can be regarded as inevitably persuasive" (2020: 1), she defines persuasion as a "communicative act […] between a persuader […] and a persuadee" which has as its goal a change in the persuadee's beliefs (2020: 15). The aim of her study is to define discourse strategies and associated linguistic means that convey persuasion within each of the genres, "establishing genre-specific repertoires" (2020: 28), which is close to the secondary aim of the present study of identifying a repertoire of evaluative language in ICS.

Dontcheva-Navratilova's study focuses on the persuader's intention, as opposed to studies of "perlocutionary effect" which assess the success of attempts at persuasion (Dontcheva-Navratilova 2020: 15). In ICS, there could be some evidence of effectiveness of persuasion on the reader if there was an assumption that ICS with high scores were successful in persuading the assessor; however, as long as there is a reasonable possibility that language was not a factor (or only a minor factor) in awarding scores, this outcome measure for impact should not be used for measuring the communicative act of persuasion, and even less so the linguistic means through which this persuasion was realised.

Persuasion can be construed in various ways. For example, the Modern Rhetoric framework (e.g. Hogan 2012) emphasises the importance of adapting any message to its audience and construes persuasion as dialogic and dynamic. This approach is less applicable to ICS because the message that they convey (i.e., impact claims) can be sent in only one version to the primary audience, namely the assessment panel of each UoA. There is therefore little variability in the audience, and the message does not have to be adapted further for other audiences. Arguably, the genre itself is already highly adapted to its audience of REF assessors, but this is done in a top-down way through extensive guidance from the REF team and the emphasis on the two criteria of "significance" and "reach". This top-down perspective is also apparent in Wróblewska's (2021) concept of the "infrastructure", where the genre is shaped through the REF guidance and the university impact managers employed

to implement it. There is, however, an opportunity to refine the message for the primary audience through pre-submission reviews by mock assessment panels.

Communication scholars such as Perloff (2014: 17) view persuasion as an attempt to trigger a change of attitude or behaviour. Applied to ICS, it is the assessors' attitude towards the quality of the impact that is subject to this attempted change, in this case ideally from a neutral stance towards one that is aligned with the quality of the impact being claimed. A central point here is the recipient's free choice regarding whether or not they are convinced by the persuasion attempt. Where this free choice is not given, persuasion ends and propaganda and manipulation start (Perloff 2014: 22). In the REF ICS context, there is little danger for persuasion to be sufficiently covert to tip into manipulation, and the power differential between the persuaders and the assessors also mitigates any threat to the free choice of the audience.

With these general, more rhetorically focused, approaches to persuasion in the background, Dontcheva-Navratilova situates "the analysis of linguistic indicators of evaluation" (2020: 5) as key to studying persuasion. She associates evaluation with more covert and less explicit means, as opposed to more explicit markers of stance that are also a way of attempting persuasion. Section 3.2.2 outlines approaches to defining and studying evaluation, which constitute possible frameworks for analysing REF ICS.

3.2.2 Approaches to evaluation

The term "evaluation" can be understood theoretically in different ways (Hunston 2011: 11), ranging from most to least involvement of a person as agent:

i. something which *a person does*, e.g. a teacher evaluating the work of a student
ii. the *language* that expresses evaluative meaning, e.g. the grammatical or lexical choices made by a teacher when commenting on a student's performance (the view applied in research by Biber, Hyland and others in their respective traditions and captured in the notions of Stance and Metadiscourse, discussed below)
iii. a *set of meanings* expressed in different ways through text, e.g. the degree of specificity a teacher expresses in such comments (as suggested by Martin and White, 2005, in their Appraisal framework, discussed below)
iv. something which *a text does*, e.g. a teacher's end-of-year report about a student (this is closest to the model developed by Hunston introduced below)

These different views of evaluation, especially ii-iv, map loosely onto different theoretical models of evaluative language and their associated research approaches. One model of evaluative language (ii) is Stance, used by many researchers (see e.g. the collection edited by Hyland and Sancho Guinda 2012). An example is the work by Conrad and Biber (2000), who quantify markers or expressions of stance (e.g. the word *important)*, rather than the content of stance (e.g. positive). In order to conduct such studies, a large amount of data is needed on the basis of which expressions of stance can be classified into certain types and then compared across corpora.

A second model that understands evaluation as language (ii) is Metadiscourse. Here, a distinction is made between "primary" (informative) and "secondary" (interactional) discourse, following Thompson (2001) who suggested the division of metadiscourse into interactive and interactional language. With its focus on lexical items (e.g. *however, you, according to*), it complements the grammar focus of Conrad and Biber's (2002) approach. However, it lacks a taxonomy that is transferable across texts beyond discrete lexical items and is therefore less suited for defining functions of evaluation (with language examples) in newly studied text groups (e.g. impact writing).

Such a taxonomy is provided in the Appraisal system by Martin and White (2005), which is consequently described by Hunston as "probably the most theory-grounded study of the functions and forms of evaluative meaning in English" (2011: 2-3). Appraisal is not about the "linguistic features" but about the "meanings" (iii in Hunston's four ways of understanding evaluation, listed above) expressed through them (Hunston 2011: 20). The framework allows for debate or at least uncertainty regarding the classification of certain instances; therefore, an appraisal analysis is "'a reading' (open to debate) rather than 'an analysis' (a definitive account)" (Hunston 2011: 21). As discussed in section 4.3.4 below, with the approaches to standardisation introduced by Fuoli (2015), the balance can be shifted towards an analysis being at least a shared account, rather than just one reading.

Finally, Hunston's three-part model conceptualises evaluation as including an epistemic object ("Status"), the value given to that object ("Value") and the relevance of parts of the text ("Relevance"). It allows, or even assumes, for every expression to be evaluative in a way, and therefore her analyses are concerned with the types or functions of evaluation in a given text or register, rather than asking whether evaluation is present at all (iv). These types of evaluation are implicit and not consolidated into a general taxonomy, as is done in the

Appraisal system (Hunston 2011: 22). This implicitness makes Hunston's model less applicable in contexts where an aim is to create transferable knowledge that can be used by future writers, and where a system with a generalisable taxonomy is likely to yield clearer results. However, the "Status" element of that model, that is, the kind of proposition or content to which evaluation is ascribed, matters greatly in ICS, where for example negative evaluation of a problem needs to be distinguished from positive evaluation of its solution based on research, namely the claimed impact. In this study, I addressed this by introducing an additional layer of analysis when coding for evaluation (see sections 5.3.1 and 7.2.2).

For ICS, the understanding of "evaluation" that is furthest removed from a human agent, namely construing the text as the agent (iv above), may be most appropriate, but it is difficult to generalise data from applying this approach. The "set of meanings" approach (iii above) developed as the Appraisal system by Martin and White (2005) is still text-based, but due to the fine-grained nature of their framework which sets out the different meanings (introduced in 3.3 below), it is better suited for making generalisations that apply to a set of texts, based on analysing a principled sample of those texts.

Across all these approaches to evaluation, there are six points of consensus according to Hunston (2011: 12-19). The first, and most fundamental one, is that evaluation is both "subjective", in that there is no objectively identifiable truth value, and "intersubjective", in that it includes interaction and a degree of shared understanding "with a social other" (2011: 12).

Second, and related to the point of intersubjectivity, evaluation "construes an ideology that is shared by writer and reader" (2011: 12), an assumption that is especially important for implicit or "invoked" evaluation. This is only recognisable if both interactants operate in the same value system (for example, "academic rigour" in Western research cultures), even if they do not share all individual values to the same extent. For ICS, this is part of the question about the extent to which an author can make assumptions about the readers' values, and what the role of the REF guidance is in creating preferences for certain values. For example, this includes an assumption that impact is "positive" by default, even if this positive view includes economic benefit for private firms without explicit societal benefits such as job creation.

A third element that is commonly recognised across different approaches to evaluation is that there is a huge range of language indicators (lexical and other) that could encode evaluative information, such that Hunston says it is "pointless" (2011: 13) to list them. Despite the difference in what the various models include in the overall label of "evaluation", there is consensus about the wide variability of language that can be used and therefore the difficulty in researching it with automated methods. Another layer of variability is the difference in explicitness, which is accounted for in several systems; for example, Martin and White (2005: 67) distinguish between "inscribed" (explicit) and "afforded" (inexplicit) evaluation in the Appraisal system, later called "inscribed" and "invoked" by Hood (2010).

The fourth element that is common across evaluation studies is that context matters: in many cases, without context, a reader cannot know whether evaluation is present, nor whether it is positive or negative ("polarity"). A distinction is made between "prior polarity" which some items have, partly corroborated through large corpus studies, and "contextual polarity", which does not always align with an item's prior polarity. For example, in science writing, where contextual polarity assumes alignment with a hypothesis, the mere act of contradicting a hypothesis can be construed as evaluation without any attitudinal lexis that has prior polarity. Another way of evaluation being created through context is cumulative evaluation, where similar evidence can either lead to an evaluative conclusion directly, or can prepare the ground for a more explicit expression of evaluation. In such a case, the cumulation "affords" the contexts for an "inscribed" evaluation.

Fifth, evaluation has a a "target" or object, and a "source". In several models, the type of target entity (a thing, concept or person) determines the type of evaluation; for example, Martin and White (2005: 45) use the complementary categories of "appreciation" and "judgement" to denote the evaluation of inanimate objects or events and of people respectively in the "Attitude" subsystem of their "Appraisal" system. The source of evaluation can be tricky to pinpoint; for example, it can be the author of a text or a quoted speaker in a text, which can create one or more layers of evaluation.

Finally, Hunston (2011: 19) acknowledges that for researchers of evaluative language with any of the above models, almost everything can be seen as "evaluative", and one can find evaluation even in the most objective-looking policy texts (cf. Tupala 2019, who applied the Appraisal system to EU policy documents on migration). Therefore, a researcher needs to

define what they wish to include in the search for evaluation in their texts, and what should be excluded. This is a more realistic endeavour when using a defined system such as Appraisal, especially when focusing on one of its three main branches.

Turning towards different linguistic methods (rather than theoretical models) for researching evaluation in texts, Sentiment Analysis is a type of automated study of evaluation, as opposed to manual analysis of texts (Lei 2021: 1). It can detect mostly superficial, visible evaluation and is applied to large numbers of texts in order to generalise across a corpus. In the case of REF ICS, this is less helpful because they are written for assessment purposes and are therefore expected to present their topics in a positive light (cf. Grant 2015, who point to the "positive sentiment" in ICS without elaborating what this is based on). Moreover, the broad brush approach of sentiment analysis, where the large volume of data is relied on to even out differences in individual texts, yields results that are "broadly accurate" (Hunston 2011: 55) rather than detailed, and therefore this approach is less useful for making recommendations for text production on how to achieve a certain effect. Indeed, Williams *et al.* (2023) conducted a sentiment analysis of the top 20% and bottom 20% ICS and found that it was not predictive of either high- or low-scoring ICS.

Beyond a fully automated approach, some methodical considerations can be introduced with a closer look at the relationship between stance and evaluation. Despite stance being presented as a model of evaluation, these two terms are often juxtaposed (e.g. Biber and Zhang 2018; Dontcheva-Navratilova 2020). Hunston (2011: 51) describes the difference along the following lines, while allowing for some overlap: Evaluation is the value ascribed to an entity, and stance is the positioning of the interactants (source and recipient) towards this value. While stance is often expressed more formally, that is, with certain linguistic features that may be searchable and therefore more easily quantifiable, evaluation is often more implicit and therefore can be studied mostly through qualitative analysis.

The analysis of evaluative language is difficult because it is not clearly associated with certain lexico-grammatical features. For example, adjectives and adverbs often have evaluative meaning, but not always, and they are not the only means for conveying evaluation. Among other features, "patterns of use" can carry evaluative meanings, but a given pattern does not always carry the same meaning (Hunston 2011: 3). Hunston (2011: 4) exemplifies this with science writing, where lexical repetition of, for example, research goals in experiment descriptions and results sections may be construed as positive evaluation of the research as

successful, even without words or sentence-level constructions that would make this visible to readers unfamiliar with the genre. In this case, the evaluation is hidden in the reading of the text and not reliably encoded in searchable language items. From this, Hunston concludes that for researching such covert evaluation, a qualitative text-based approach is more suitable than a quantitative corpus-based one (Hunston 2011: 4).

In further explaining this distinction, Hunston (2011: 50) suggests that "corpus-based" analyses are quantitative and may include significance testing when comparing findings across corpora, and therefore any qualitative analysis of texts is "text based". However, arguably the boundaries are less clear-cut. A text-based analysis can be done on a principled collection of texts and therefore a corpus; and if text-based analysis includes (manual) tagging of certain features, these tags can still be quantified across the corpus and any sub-corpora. In this way, any textual feature (not restricted to certain linguistic structures but also including Appraisal resources as introduced in section 3.3 below) can be analysed in a corpus-based way, where manual tagging provides a basis for some form of quantitative analysis. Like Biber and Conrad (2019), Hunston (2011: 65) acknowledges that corpus studies should combine quantitative and qualitative research, and the difference in approach is more one of starting point and emphasis: quantitative research for example in register variation can be a starting point for qualitative analysis, and findings from qualitative corpus research, which is more conducive for finding covert evaluation, can also be quantified.

3.2.3 Stance versus evaluation in persuasion

One aim of studying ICS is to uncover hidden evaluation in a way that is generalisable (see research question 2 in section 1.2). The previous section implies that this is all but impossible. It has nonetheless been attempted in other registers, for example in a study by Biber and Zhang (2018) on how persuasion and evaluation are conveyed in certain internet registers, where they juxtapose *explicit* grammatical "stance" and *implicit* lexical "evaluation". This study is described in detail in this section. Biber and Zhang (2018)contrast the two frameworks such that stance is expressed in overt lexico-grammatical devices showing "epistemic or attitudinal meaning" about a proposition that can be directly attributed to the speaker, as in (10), or more indirect, as in (11) (Biber and Zhang 2018: 98, their examples):

(10) I'm sure you're right.

(11) It seems that we may be experiencing a perfect storm.

Studies researching stance are usually corpus-based and quantitative, tagging and analysing a finite set of linguistic features that are identified *a priori* as explicit stance markers (see e.g. the discussions in Hyland and Sancho Guinda 2012). By contrast, and in line with the discussion in Hunston (2011), evaluation is introduced by Biber and Zhang as being broader and often expressed through a wide variety of lexical items, and therefore "less clearly delimited in its linguistic definition" (Biber and Zhang 2018: 99). This makes it impossible to have a similarly defined set of markers, either grammatical or lexical, especially since evaluative meaning is often "implied rather than stated explicitly" (Hunston 2011: 19).

A "clearly delimited […] definition" would be needed in order to apply Biber's flagship method of investigating register, which relies on automatic tagging of large amounts of text, and the view of grammatical stance employed by Biber and Zhang (2018) facilitates that method. Where corpora are used for studying evaluation, however, this is often based on manual analysis of concordance lines, an approach which Biber and Zhang (2018: 112) critique as not allowing generalisations about the registers present in the corpus. According to Larsson (2019: 246), their 2018 study bridges the "paradigm gap" between quantitative and qualitative approaches to register analysis.

After describing the opposing methodological approaches associated with the theoretical concepts of stance and evaluation, Biber and Zhang (2018: 100) point out that researchers from different traditions in applied linguistics come to "opposing conclusions regarding the prevalence of attitudinal language in particular registers". Academic research writing is a case in point, as it is a register "marked for the absence of lexico-grammatical stance features" (Biber and Zhang, 2018: 100, referring to earlier work including Biber 2006), while the work of discourse analysts such as Hunston (e.g. 1994) and Hyland (e.g. 1998, 2005b) consistently shows the evaluative nature of academic prose. In fact, both Hunston and Hyland assert that evaluation, and persuading the disciplinary community of new knowledge claims, is the main purpose of research writing.

Indeed, Biber and Zhang's (2018) own findings show that grammatical stance cannot explain the nature of persuasion in all texts, and other frameworks are therefore needed for describing persuasion in registers that are perceived as persuasive but are not marked for grammatically visible stance. Their study uses this well-described discrepancy in assessments of whether a text is persuasive or not in order to explain the nature of a newly-defined register of internet texts, which they call "informational persuasion" (2018: 101). These

texts, chiefly product descriptions in online shops, are characterised by the absence of grammatical stance, similar to academic prose, but they were clearly identified as examples of a persuasive register by users. They were further specified as "informational description with an intent to sell" (Biber and Zhang 2018: 102). The "informational persuasion" register has two primary communicative functions, namely to present descriptive information and to persuade the reader. In order to resolve the apparent discrepancy between user categorisation and the findings of their previous multidimensional analysis, Biber and Zhang (2018) set out to determine what other language features these texts have that mark them as persuasive in the users' eyes.

Starting with the theoretical principle of register studies that "communicative functions are realised through linguistic devices, in ways that can be generalised for a register across a corpus of texts" (Biber and Zhang 2018: 112), they designed a novel kind of analysis to pinpoint the linguistic devices that are used in these texts to realise the evaluative function under consideration. They combined several corpus-linguistic methods to contrast the target register with a parallel internet register ("Opinion", i.e. opinion blogs or product reviews, as opposed to product descriptions) which is marked for overt lexico-grammatical stance markers in a way that "informational persuasion" is not. After identifying keywords that distinguish these two specialist corpora from each other, they coded those keywords that were deemed to have an evaluative function into different categories. The framework they apply is reminiscent of the Attitude sub-system of Martin and White's (2005) Appraisal system, both through the use of categories called "attitude" and "judgement" and because of the focus on overt lexical markings, although with a much broader focus than the defined grammatical markers of "stance". The next step was to quantify the occurrence of the words in the respective corpora, which showed that certain functions were used more in the "informational persuasion" register than in the "opinion" register, namely those of "evaluation of other objects" (e.g. *free, fascinating, perfect*), "evaluative manner" (e.g. *highly*) and "indirect value judgement" (e.g. *success, offer, recommend*). Words in these three categories do not usually occur in the grammatical structures that constitute stance markers, such as examples (10) and (11) given at the beginning of this sub-section, and they were therefore missed in their previous analysis (Biber and Zhang 2018: 116). The caveat of this more specific approach is that these evaluative items cannot be generalised to other registers beyond the one from which they were extracted through keyword analysis, due to

the specific situational context. For a register with similar communicative purposes (description *and* persuasion) with an otherwise different situational analysis, a new analysis will have to be conducted. The authors recommend further similar studies for academic research writing, which also has this double function of description and persuasion (Biber and Zhang 2018: 120).

Having been described for the non-academic "informational persuasion" internet register and for general academic writing, this discrepancy between a perceived persuasive purpose and a lack of grammatical features associated with persuasion is also found in impact case studies. My comparison of ICS with other registers along the Biber (1988) dimensions introduced in 3.1.1 shows that on the one dimension that relates to persuasion (Dimension 4 *Overt expressions of persuasion*), ICS have negative loadings (Figure 2, blue oval). This means that compared to other registers, such as "involved persuasion" or "imaginative narrative" which have positive loadings (above the middle line), these texts display fewer of the features typical for persuasion, such as modal verbs.

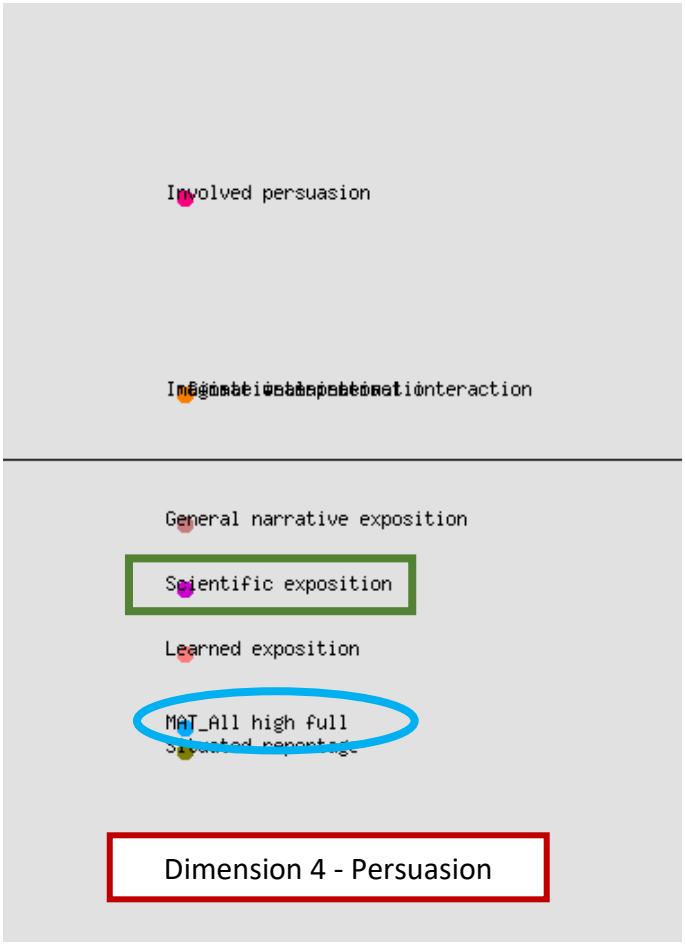


Figure 2: ICS have a negative loading of explicit persuasion (Dimension 4 in Biber 1988); extract of Figure 1

The literature on ICS expects a register with high levels of persuasion (see e.g. Watermeyer and Hedgecoe 2016). As in Biber and Zhang's (2018) study, it is therefore necessary to tease out the linguistic features that are used to realise this discursive function. If an explicit awareness of these features could be made available to both writers and assessors of ICS, the assessment process could be more transparent because writers would have the same information, and assessors may be in a better position to recognise covert evaluation more consciously. However, Biber and Zhang's (2018) solution of finding specific lexical items that distinguish their target register from another specific reference register through keyword analysis, followed by examination of concordance lines, is less suitable for my analysis of ICS. One reason is that there is no obvious reference register: while a corpus of academic research articles could potentially serve this function, building such a corpus is beyond the scope of this project. The second, more important, reason is that evaluation is also conveyed in non-lexical, non-grammatical ways that would not be possible to quantify using their method, such as intensification through repetition or metaphor. These, however, can be categorised and quantified as part of the Graduation sub-system of the Appraisal framework (Martin and White 2005) first introduced in section 3.2.2 above and discussed in detail in section 3.3 below. The Appraisal framework may therefore be more suitable for an analysis of evaluative language in ICS. A further reason for choosing the Appraisal framework of different kinds of meaning is that some commentators criticise "marketing" language in ICS (e.g. McKenna 2021), and it would therefore be problematic to simply provide a list of words that are then at risk of being read as recommendation, being overused and losing their meaning, potentially making the assessment process more difficult. Such an approach would also invite accusations that the process was being turned into a kind of "bingo", which could validly be levelled against the table of words provided by Biber and Zhang (2018) for "informational persuasion" registers.

3.3 Appraisal framework

Register analysis includes functional interpretation of language features in specific situational contexts, explaining why these features are associated with that situation. In the Biber tradition, this is done with the help of comparative tables of situational and linguistic analyses of at least two registers, to enable links to be made between the two types of analysis and describe differences between the registers (Biber and Conrad 2019: 69). However, Hood (2010: 4) argues that this functional interpretation in register analyses is still

based on intuition, rather than the data itself, and posits that when a functional perspective is applied from the start, the interpretation is more reliably grounded in the data. The present register analysis of ICS has a functional focus of describing evaluation within a register, rather than comparing different registers; therefore, a functionally grounded analysis may be more suitable. While register analysis can describe the formal properties of language and their distribution across texts (linguistic analysis), and relate this to the context and purpose of the text (situational analysis), it is less suited to making claims about the effects that writers may have intended the texts to have on the reader, by using these linguistic features (see Finegan 2019: 201 who emphasises the "powerful descriptive reach" of register studies as a main asset). This aim of explaining the effect of language choices, that is, their function, is related to research question 2.

As established in section 3.2.2, the Appraisal framework offers the tools for categorising varied features of evaluation. Moreover, Appraisal can work well with a social theory of discourse because both assume that the context (social practices and conditions) "shape the linguistic features of texts" (Tavassoli *et al.* 2019: 76). While this is similar to the situational analysis in Register studies, the sequence of an Appraisal analysis is *situation > function > linguistic features*, rather than *situation > linguistic features > explain function by matching features to situation* as applied in register analyses. The study of evaluative language in ICS in chapter 7 follows this pattern, complementary to the elements of the register analysis presented in chapters 5 and 6.

Many studies apply the Appraisal framework to overtly evaluative texts, such as certain media texts or reviews (e.g. Hommerberg and Don 2015; Tavassoli *et al.* 2019). For the study of REF ICS with their dual function of description and covert persuasion, the most relevant recent study is Tupala (2019), who proposes a framework for applying quantitative appraisal analysis to official institutional documents that are assumed to be factual and to exhibit no evaluation. However, as Tupala shows, underlying assumptions are everywhere and they certainly exist in institutions, and the challenge is therefore to uncover these and the ways in which they are expressed. This is similar to ICS, which are variously described as factual and persuasive. Although the presence of attitudinal positioning is less surprising in a text written for assessment rather than as a policy document, the observation from Tupala about policy documents that "the values and attitudes expressed in them have to be deep-rooted in order to penetrate the seemingly neutral" text (2019: 3) can also be true for ICS, where

some wordings may be understood as evaluative by those familiar with the UK HE landscape and the REF criteria, but not by others.

The Appraisal framework was developed in the tradition of Systemic Functional Linguistics (SFL). It is beyond the scope of this thesis to introduce this theory beyond those aspects that are relevant for Appraisal. SFL is a holistic theory of language offering a framework of analysis that can reach from the smallest unit of language (phonology or graphology) up to the broad level of discourse semantics (level of "realisation"), and from the language system as a whole and its social context down to the individual utterance (level of "instantiation", Martin and White 2005: 7-25). According to Matthiessen (2019), studying register in the SFL framework sits between the extremes in both dimensions: the focus of SFL-informed register studies are the semantics as realised in lexicogrammar, and the features of related texts (rather than instantiation in a single text, or random language examples).

At the heart of SFL is the model of metafunctions that are at play in a given text: ideational, textual and interpersonal. The ideational metafunction is related to the informational content of a text, whereas the textual metafunction refers to the way in which this informational content is represented. This distinction will become relevant again in chapter 6 where, in addressing research question 1c, language segments are categorised depending on the extent to which they are content-driven, that is, linked to the ideational metafunction, or an editorial language choice, that is, connected to the textual metafunction. Outside of making this distinction, however, these two metafunctions are not addressed in this thesis: ideational content varies too much across ICS to be meaningfully analysed, and textual observations are discussed with the help of other frameworks (see sections 5.2.3 and 6.2.1). The final metafunction, interpersonal meaning, is most relevant to this thesis, specifically for discussing evaluation to address research question 2.

Appraisal theory is part of this interpersonal metafunction of SFL. Unlike register inquiry more generally, Appraisal works on the "discourse-semantic" level of realisation and can therefore be studied in many different lexico-grammatical forms. It is a way to *systematically* uncover the "evaluative and persuasive workings" of texts (Tavassoli *et al.* 2019: 66). Rather than relying on impressionistic description, it offers a way "to operationalize [certain] characterizations in a linguistically principled manner" (Tavassoli *et al.* 2019: 68). This happens along three dimensions, which constitute the three subsystems of Appraisal theory:

- The expression of values as positive or negative assessment: **ATTITUDE**[5]
- Manipulation of the intensity of these values: **GRADUATION**
- The introduction and management of voices to whom values are attributed: **ENGAGEMENT**

ATTITUDE can be conveyed explicitly (INSCRIBED) or implicitly (INVOKED). INSCRIBED attitude can be identified by the overt presence of positive or negative *value* and the possibility to *grade* that value up or down (Hood 2010: 75). INVOKED attitude is less overt and is often implied through a resource of GRADUATION but can also happen through juxtaposition of a positive and an apparently neutral proposition, which turns the neutral proposition into a negative appraisal (Tavassoli *et al.* 2019: 69). The general ATTITUDE in ICS is expected to be positive throughout (e.g. Brauer *et al.* 2019: 66; Grant 2015: 72; Hinrichs and Grant 2015: 2). Therefore, this subsystem was not chosen for analysing ICS.

GRADUATION is the element of the Appraisal system most meaningful for studying ICS. Questions include: how, and how far, is the assumed positive attitude pushed? What linguistic devices are most common for this function in high- and low-scoring ICS respectively? Does this vary across Main Panels?

ENGAGEMENT is widely studied in academic writing because a large component of academic discourse is positioning research in relation to previous research (e.g. Humphrey and Economou 2015; Liardét and Black 2019; Xu and Nesi 2019). In REF ICS, however, this is less relevant – the texts are descriptive and persuasive, but less discursive and argumentative. Other voices are included in the text, for example through testimonial quotes or other references to sources, but they are all assumed to support the same value: a claim to significant and far-reaching impact. Therefore, again, this subsystem was not chosen for analysing ICS.

In this study, I therefore apply the GRADUATION subsystem to a balanced sample of Section 1 texts from 76 ICS (see section 4.3.4 for a description of the sample) in order to describe the use of covert evaluation, which may help to explain the perceived discrepancy between the characterisation of ICS as description and persuasion respectively.

[5] Components of the Appraisal framework are set as SMALL CAPS in this thesis.

In academic writing, instances of evaluative language are "invitations to align and to build relationships of solidarity" (Hood 2010: 37). The reader and reading context of a text varies, and therefore the degree of alignment with the writer's proposition can vary, too. A writer has various resources ("discourse strategies") available to maximise this alignment, and Appraisal theory "provides a framework for the analysis of these evaluative strategies" (Hood 2010: 75). More specifically within the Appraisal framework, GRADUATION enables researchers to explore "the multiple ways in which academic writers grade meanings strategically in their writing" (Hood 2010: 16). An essential part of this strategic grading is "the play of inscribed and invoked attitude" (Hood 2019: 390) because invoking attitude, rather than inscribing it, makes the stance of a text clear but not obvious. In that way, the text or genre could still be seen as descriptive, rather than evaluative. This invoking of attitude through the use of GRADUATION resources can take the shape of, for example, downtoning, vague language or hedging especially in academic writing, but it can also represent positive evaluation and boosting, which matches the persuasive function in ICS.

The system of GRADUATION is described in detail in the Coding Manual created for this thesis (Appendix F). It is in principle divided into FORCE and FOCUS. FORCE refers to the intensity of a value, which can be dialled up or down in various ways. These broadly fall into the two categories of INTENSIFICATION (e.g. "VERY/SLIGHTLY/SOMEWHAT important") and QUANTIFICATION (e.g. "greater competence", "long-lasting"). FOCUS refers to the prototypicality of an entity, that is, how "real" or "complete" it is. Most of the categories can be further subdivided, as shown in Figure 3. The various levels of division are termed "levels of delicacy", and I will refer to different levels of delicacy when describing findings in the analysis in chapter 7.

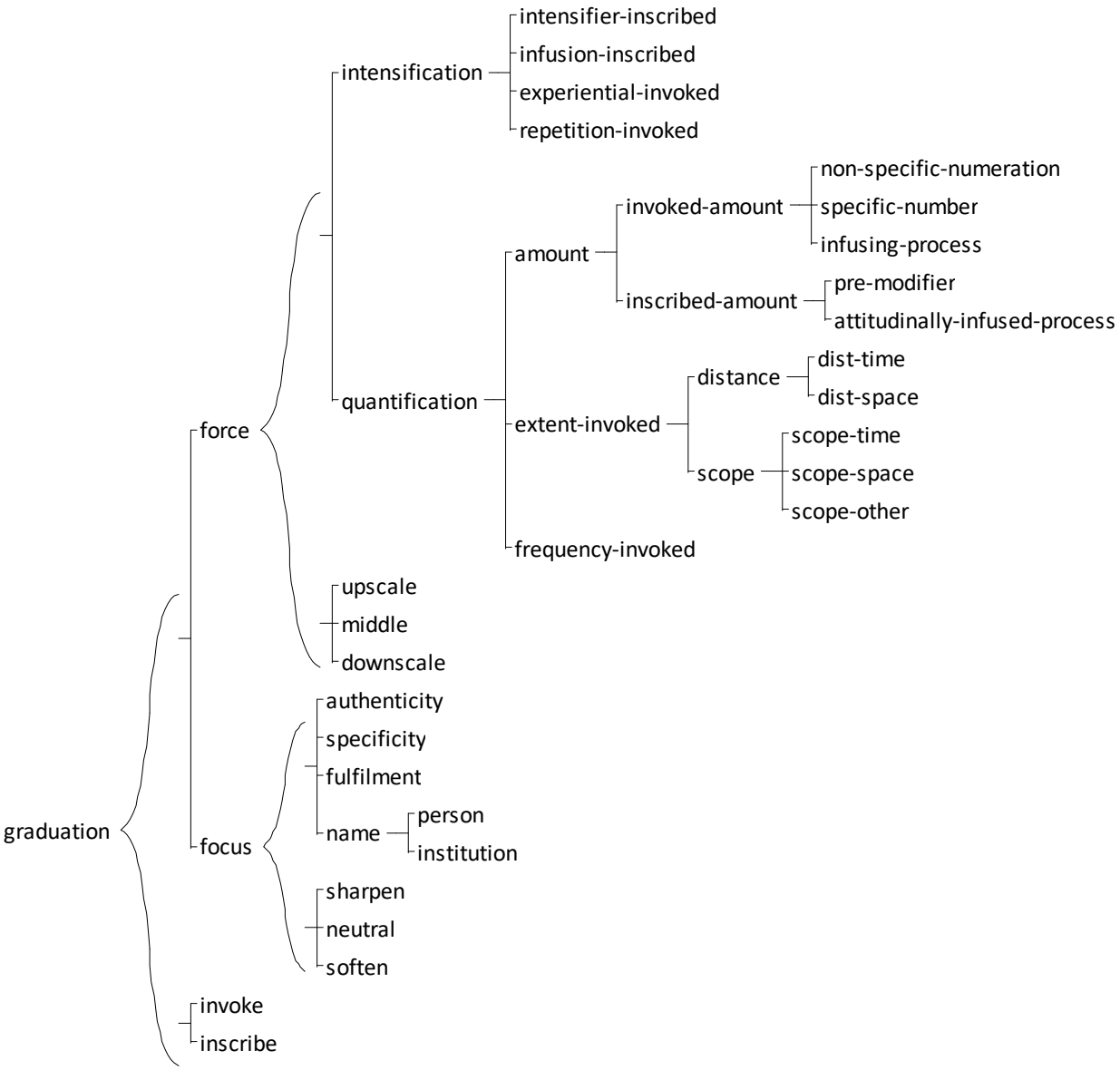


Figure 3: Coding scheme for the GRADUATION analysis used in this thesis

All the end points on the right in Figure 3 can be scaled up or down (e.g. for FORCE:QUANTIFICATION:AMOUNT:INVOKED-AMOUNT:NON-SPECIFIC NUMERATION: "few studies" or "many studies"). For resources of FORCE, this can be UPSCALING or DOWNSCALING, and for resources of FOCUS, the directions are termed SHARPEN and SOFTEN, akin to a camera lens. In both cases, an intermediate category was added for analysing ICS, as explained in chapter 7.

Often, GRADUATION resources are applied to attitudes, but they can also be applied to "experiential", rather than attitudinal, meanings. In this case, they can still be read as invoking attitude by their sheer presence, especially if they grade a verb meaning which is not normally graded. Similarly, where quantity is graded, this opens a space for evaluative reading, because it is marked and therefore flags to the reader that something can be interpreted as invoking attitude. If this is found frequently in ICS, it could contribute to

characterising the nature of ICS as evaluative texts despite the relative absence of overt stance markers.

A GRADUATION analysis is therefore suitable for exploring the question of how the "selling" (Watermeyer and Hedgecoe 2016) of impact is expressed in ICS. This approach is especially promising because of the potential to use GRADUATION resources as a way of covert persuasion (Hood 2010: 89), that is, by using a framework that includes a way to record covert or invoked meaning. It allows, for example, questions such as: is the positive Attitude ascribed to ICS by the Impact literature INVOKED more in some corpus parts than in others, where it may be more INSCRIBED? Does this result in a different kind of "selling"? Similarly, Tupala (2019: 3) suggests that the "distinction between explicit and implicit appraisal […] is an important part of decoding attitudes in official texts", especially because the instances of implicit evaluation can be invoked through word combinations, rather than individual words, and the interpretation of these words or combinations as "attitudinal" is contingent on context, such as cultural or social reference.

3.4 Chapter summary

In this chapter, I first described the overall framework of register analysis and its relevance to impact case studies, with reference especially to differences between academic disciplines in research writing more widely and the role that this may play in writing ICS. I then introduced several ways of conceptualising persuasion and evaluation. With this background, I argued that the Appraisal framework is the most appropriate approach for investigating evaluative language in this particular register, as it provides both a framework to enable comparison of findings across sub-corpora and a large degree of flexibility to code specific language items as evaluative even if these may not carry such meaning in other contexts. In the next chapter, specifically section 4.2.2, I explain the structure of this register study, and in chapter 7 I apply the Appraisal framework to a sample of ICS.

Chapter 4 Research Methods

Following the discussion in chapter 3 of linguistic frameworks that relate to issues raised by the research questions, this chapter provides an overview of relevant approaches to researching REF impact case studies (section 4.1), followed by an outline of my own overall research design (section 4.2). It then describes and justifies the various samples and sub-samples used in the different analyses performed for subsequent chapters (section 4.3) and the technical steps involved in preparing the corpus (section 4.4). Due to the variety of analyses, detailed descriptions of the analysis methods are provided in the respective results chapters (5-7) before presenting results, rather than in this overarching chapter, where possible.

4.1 Methods in previous research on REF impact case studies

As indicated in section 2.1.4, the REF ICS database has been used as a research resource to answer many different questions. In this section, I will discuss two aspects of research design that are particularly relevant to my own questions and critically assess their use in other studies, before proposing the unique combination of approaches used in this thesis in section 4.2. First, I will discuss methods that have been employed to identify differences, whether textual or other, between ICS in different scoring brackets, because this dimension has so far been missing as an empirical variable in discussions of the presentation of ICS. Then I will discuss examples of studies that used automated text analysis for addressing questions related to the content, rather than the language, of ICS.

4.1.1 Analysing the influence of independent variables on REF scores

Several approaches have been used to interrogate differences between high- and low-scoring ICS. Individual scores for ICS are not reported in the official REF results;[6] the official spreadsheet does, however, show the percentage of a submission (i.e. all ICS submitted by one university to one UoA) that received a certain score (see Figure 4 below). From these percentages, the grade point average (GPA) for the respective impact submissions can be derived. This measure has been used by a number of studies that applied it to the whole dataset of published ICS, using descriptive statistics to infer generalisations. For example, it was applied by Ginsparg (published in Van Noorden 2015: 150) to distinguish "power words" in each UoA (see section 4.1.2 for more on their approach), and by Loach (2016: 5-6) to

[6] The official results spreadsheet is downloadable at http://results.ref.ac.uk/DownloadFile/AllResults/xlsx.

examine the types of evidence that were associated with high and low GPAs in each Main Panel. However, this is a rather crude measure which rests on the power of using statistics on a large sample, while masking a great deal of variation and nuance within submissions. For example, Figure 4 shows how some submissions are more homogenous (e.g. St Andrews, Open University, University of Chester) and some are heterogeneous in the scores they received, making GPA a potentially misleading measure.

	A	B	C	D	E	F
1						
2		Percentage of the submission meeting the standard for:				
3	University	4*	3*	2*	1*	unclassif
4	University of St Andrews	76.0	24.0	0.0	0.0	0.0
5	Open University	20.0	80.0	0.0	0.0	0.0
6	University of Chester	0.0	0.0	0.0	90.0	10.0
7	University College London	48.6	28.5	17.2	5.7	0.0
8	Queen's University Belfast	40.0	10.0	40.0	10.0	0.0

Figure 4: Screenshot showing an extract of the official REF results spreadsheet (selection of universities, UoA17: Geography, Environmental Studies and Archaeology)

GPAs are also used by several studies focused on one discipline, rather than the whole dataset. One example is Baguley (2017), who attempted to predict GPA based on the results of his own categorisation of UoA4 (Psychology, Psychiatry and Neuroscience) ICS into impact types. Williams *et al.* (2023) recognised that using GPA as a measure to rank all ICS is problematic. While they also initially ranked ICS by GPA, they attached a GPA to individual ICS, rather than whole submissions, for this ranking. Crucially, they then only included the resulting top and bottom 20% of ICS in their subsequent analyses of characteristics, noting that "widening this threshold would make those characteristics less prominent" in their dataset (2023: 5).

A second approach used GPAs as a starting point and then applied a statistical technique called *k*-means clustering, to categorise the submissions into quality profiles (top – middle – bottom), in analysing the impact submissions in UoA19 (Business and Management, Kellard and Śliwa 2016: 698). Their aim was to determine the prevalence of certain factors relating to the researchers named in ICS which they hypothesised may have influenced the score, such as gender, career stage and whether ICS included teams, rather than centring on single researchers. Phillips *et al.* (2020) extended this work in their article on UoA26 (Sport and Exercise Sciences, Leisure and Tourism). They first correlated the GPA of each impact

submission with the scores for the respective Environment component in order to test the validity of using submission-level GPA as a proxy score for ICS. They then applied Qualitative Comparative Analysis to eight factors related to the research environment based on those used by Kellard and Śliwa (2016) and Chowdhury *et al.* (2016). This analysis is similar to a regression analysis but with a focus on the qualitative clustering of factors and their combinations, rather than the quantitative effect of each factor. Of the eight factors, their analysis identified four that they interpreted as being predictive of a higher score: either a large grant, or a combination of (1) public interaction, (2) a high proportion of journal articles in the outputs listed in Section 3, (3) a key researcher who had been in their current position for a long time. As Phillips *et al.*'s (2020) study is restricted to one UoA only, the role of these factors cannot be generalised to other UoAs.

A third approach to determine factors that influence scores was employed by Chowdhury *et al.* (2016), who randomly selected five UoAs across the Main Panels for their sample. Rather than using grade point averages, they ranked all submissions in the chosen UoAs by the percentage of 4* ICS and chose the top and bottom five submissions in each (top and bottom 10 in UoAs where submissions were very small). They started with six parameters which they hypothesised may be related to scores, such as grant income. Two of these proposed parameters were related to presentation (of income and of esteem factors) but for neither of these two factors could a relationship with scores be established in any of the selected UoAs (Chowdhury *et al.* 2016: 12), which seems to indicate that presentational factors did not affect assessment outcomes in those UoAs. Some of their measures especially around funding were unclear. For example, they seem to have combined different kinds of income that was either presented as impact generated for a partner or as income won for research, and they present a measure of "total research income" which again cannot be accurate for all ICS because in the 2014 template only four grants could be listed. The article states that universities "have to consider [these factors] in order to get a good impact score" (Chowdhury *et al.* 2016: 10). This is a bold claim for a questionable level of validity, even though their aim was to tease out what factors played a role beyond the official criteria of "significance" and "reach", neither of which can be metricised in the way their factors can.

The study that goes furthest in isolating top-scoring ICS is the analysis by Gow and Redwood (2020) discussed in detail in section 2.1.4. They focus exclusively on ICS that are part of a

submission that achieved 100% 4*, without acknowledging that these submissions represent only a tiny fraction of 4* ICS (111 in their sample, of nearly 3,000 4* overall). While this is a small proportion, it allows them to tease out characteristics that are common across 4* ICS. However, unlike Williams *et al.* (2023), their analysis does not include a comparison with lower-scoring ICS, which makes any claims about the influence of their characteristics on scores difficult to substantiate.

4.1.2 Textual analysis of impact case studies

In addition to considerations of REF scores, a second relevant aspect of research is the use of linguistic methods in studies with other aims. For example, the first official analysis of ICS, by Grant (2015: 42), included the use of concordance lines to identify potential stakeholder groups. Searches for "stakeholders", "beneficiaries" and "users" helped the team to uncover frequencies of the nouns appearing around these in all ICS. However, they did not examine the individual concordance lines of the nouns they found to be most frequent, which would have been necessary in order to ascertain whether they were included in a role as beneficiaries or rather mentioned in the vicinity of the node for other reasons.

Another study that used text mining in a similar way to Grant (2015) is Terämä *et al.* (2016), whose aim was to "build a sharper overview of the impact landscape" (Terämä *et al.* 2016: 4). The advantage of using the full dataset comes with the disadvantage of being unable to read the individual texts; without having read any ICS as part of the reported analysis, Terämä *et al.* (2016: 14) claim that they can identify how researchers "interpreted" impact. Their claims are based on the relative frequencies of single words, but again without inspecting the details of the contexts reflected in the concordance lines. Where Grant (2015) used Section 4 only, Terämä *et al.* included Sections 1 and 4 in their analysis, as well as the available metadata. Their aim of finding impact "classes" through this text mining analysis, on which submissions can then be mapped, is somewhat hampered by the fact that they analysed texts at submission level. The problem with that approach is that it misses variation of ICS within these submissions; for example, the same submission might include one ICS with impact mainly on education and another one with impact mainly on policy (both are classes that emerge from their analysis). This makes it difficult to assign a whole submission to just one class. In fact, universities were indirectly encouraged to submit a range of different impacts based on different research: the impact template submitted alongside ICS

in REF2014 was designed to illustrate the breadth of the unit's impact strategy and should be illustrated by ICS, thereby encouraging variety in the ICS as well, where possible.

A study that used text mining in a different way to Grant (2015) is Bonaccorsi *et al.* (2021). This group of data scientists combined two approaches, text mining and lexicon, to generate a comprehensive ("semantically saturated") list of users and user groups that could appear in ICS and then tagged these in the whole database. They essentially drilled down into one aspect of Grant (2015), but rather than identifying collocations of "stakeholder", "beneficiary" and "user" directly from the dataset, they used a top-down, pre-determined set of close to 80,000 terms that they tagged. Their article includes tables of "user groups" (e.g. "cultural heritage volunteer", "startup investor", "civil servant", "primary school", "fish farmer" – Bonaccorsi *et al.* 2021: 19-20) and the words associated with those, but no connection is made to scores. Moreover, no context is provided for the user groups, that is, no indication of whether and how they are impacted: they are assumed to be user groups because they are (groups of) people that are mentioned in Sections 1 or 4 of an ICS, even though there may be different reasons for including a group in the text (for example, they may have been negatively impacted or opposed the impact, or may have been research partners rather than beneficiaries of the research).

Further studies that include textual analysis are Derrick *et al.* (2014), who used an application called VOSviewer to visualise noun phrases around certain words in policy documents linked to the preparation of the 2014 submission, and the aforementioned Chowdhury *et al.* (2016). In addition to the analysis of factors that may be related to scores, Chowdhury *et al.* (2016) reports on a study that employed frequency counts to extract "research themes" from their sample of all ICS in five randomly selected UoAs. The label of "research themes" is debatable for something that in the context of their article can be clearly identified as "impact themes". Unlike their other analysis, this textual analysis is not linked to scores, and therefore their conclusion based on both studies that universities "should" submit ICS on those themes is perhaps misleading.

The studies introduced in this section have used text mining or other corpus analysis methods to uncover overarching themes, such as an understanding of impact (based on policy documents in Derrick *et al.* 2014 and on ICS in Terämä *et al.* 2016), or to extract impact types (Chowdhury *et al.* 2016). However, they do not use these language-based methods to interrogate the language of ICS itself, and certainly not to explore why certain

language choices may have been made. One of the few publications that focus explicitly on language is Van Noorden (2015), who reports on "power words" that appear more frequently in high-scoring than in low-scoring ICS. In this approach, all available ICS were included and categorised by UoA, then raw word frequencies were determined and stopwords (i.e. those that are frequent in texts across genres, such as "the", "it", "for") from conventional stopword lists were removed. The GPA for each submission was then correlated with the most frequent words, resulting in an overview of words that were relatively over- or under-used in high- or low-scoring submissions within each UoA. This approach masks variation of scores within submissions (as illustrated in Figure 4 above) and does not clearly identify high-scoring texts and therefore the language used in those. By contrast, my sample includes a clear distinction into high- and low-scoring ICS, rather than GPA, as explained in 4.3.1 below. Moreover, the smaller sample in my study enables a closer analysis of the use of words in context, as described in sections 6.1 and 7.1.

4.2 Research design

When deciding on appropriate research methods, a researcher first needs to ask what is possible to know about the subject of their study, what they can claim, and on what basis. The overview of this thesis' research design therefore starts with the underlying epistemology (constructionism) and theoretical perspective (interpretivism) following the terminology and definitions of Moon and Blackman (2014). I will then outline my approach to register analysis (section 4.2.2) and provide reflections on the role that my positionality as a research impact consultant plays in the research (section 4.2.3).

4.2.1 Epistemology

A constructionist epistemology assumes an interplay between subject and object, that is, the researcher and the researched (Moon and Blackman 2014: 6). This applies in the current study of the writing and reading of ICS and their possible association with a score. Under a constructionist view, the researcher acknowledges that the exact sample and analytical methods influence the possible outcomes of the analyses. Other researchers applying other methods, such as interviews, to the same context of REF impact assessment may come to different conclusions about the role of presentation in this assessment (e.g. Gow and Redwood 2020; Watermeyer 2019). For the overall research question, there cannot be one single definitive answer, but it is possible to take measures that increase the robustness of findings, including:

1) use of natural data from as wide and principled a sample as possible
2) use of descriptive and inferential statistics, which add a layer of safeguarding against too much reliance on the researcher's intuition
3) use of established frameworks for the analysis of evaluative language
4) thorough documentation and the use of second coders for manual coding in order to make the analysis as principled as possible.

Interpretivism assumes that reality is culturally and historically situated (Moon and Blackman 2014). Therefore, it can only be generalised within the same context. Section 5.1 provides a comprehensive situational analysis, which complements the historical and political background of assessing research impact in the UK presented in section 2.1. Any findings from this study can be applicable to this context and situation at best.

It is important to stress at this point that no causal claims are made in this thesis, certainly not between language and score, and readers are advised to take these limitations into account and not treat the findings as recommendations. The aim is first and foremost to find, or exclude, a relationship of any kind between presentation and score. Others have suspected a causal relationship where scores were influenced by presentation to the detriment of content (Watermeyer 2019: 80), and my contribution is to uncover whether a quantifiable association exists in the first place. If this cannot be detected with a range of research methods, then it would be more tenuous for others to infer a causal nature of this purported relationship between presentation and assessment outcomes.

4.2.2 Approach to register analysis

Based on the frameworks of constructionism and interpretivism, this study is designed as a register study using a range of quantitative and qualitative corpus-linguistic methods. Register studies ask questions about form ("What do we find?") and function ("Why do we find this?"). The former is the fundamental question asked in any corpus study. The latter is distinctive to register studies: it relates to the context external to the text, specifically how that informs the choice of the form (Biber and Conrad 2019: 9) and what implications this choice may have, in the case of REF ICS on the assessment process and potentially its outcome. This context, which forms the basis for a functional analysis, is provided in the situational analysis in section 5.1.

As a register study, this thesis aims to identify whether certain linguistic features are "typical" for certain parts of the register. For any such claims, three considerations need to be met (Biber and Conrad 2019: 53-59):

1. A comparative approach to determine respective frequencies and distribution, in this case within-register comparison across scoring brackets (unlike e.g. Gow and Redwood 2020, who only discuss high-scoring ICS);
2. Quantitative analysis to ensure that conclusions are not based mainly on intuition (unlike e.g. McKenna 2021, whose recommendations are based on his personal experience as assessor and sub-panel chair, rather than on structured empirical research);
3. A representative sample to ensure that findings are not skewed by higher- or lower-than-expected prevalence in certain texts. If the register under consideration is very specific, as is the case in this thesis, a smaller sample is acceptable (Biber and Conrad 2019: 33).

One of the defining characteristics of corpus linguistics is that researchers can use large datasets and apply quantitative, computer-based methods to analyse these (Friginal and Hardy 2014: 19). This can make findings more generalisable to the type of texts from which the corpus sample is taken, but only if the sample is sufficiently representative (McEnery *et al.* 2006: 13). It is therefore important to carefully consider and explain the sampling principles and sample size (see section 4.3.1 below). On the basis of a representative sample, a linguist can then use quantitative methods to identify differences in frequencies of words or word patterns (Desagulier 2017: 7) or of grammatical features such as tense markers or the passive voice (e.g. Biber 1988: 72).

However, with the use of purely quantitative methods, only very limited claims can be made about the texts under investigation. In the words of Friginal and Hardy, for example, "there is little importance to knowing […] that one gender group uses more passive voice constructions than another without being able to explore the functional reasons behind that difference in a particular context" (Friginal and Hardy 2014: 20). Qualitative analysis is therefore an essential step to interpret quantitative results and to establish whether they are meaningful for answering a specific research question (Desagulier 2017: 8). This often takes the form of inspecting and coding concordance lines, which show all or a defined

number of instances of a search term in context. Examining a corpus in this way can help to validate and add meaning to quantitative findings.

The main methods used in this research are therefore, from most quantitative to most qualitative:

1) Quantitative analysis of grammatical features linked to cohesion and readability using the Coh-Metrix tool (McNamara *et al.* 2014);
2) Quantitative lexical analysis with standard corpus tools (Keywords, frequency lists, n-grams), supplemented with manual checks on how certain expressions are used (e.g. Durrant 2017; Scott 1997);
3) Manual tagging of textual features based on adapted frameworks, and subsequent statistical analysis (Target and Appraisal analyses) (Fuoli 2015; Martin and White 2005);
4) Qualitative thematic analysis based on close reading of the texts (e.g. Braun and Clarke 2006).

Findings from these analyses are presented in chapters 5-7 not primarily by type of analysis but in a thematic order, zooming in from the writing and reading context of ICS and content-related analysis (chapter 5) to lexical choices (chapter 6) and culminating in notes on evaluative language (chapter 7). An overview of which analyses are conducted on which sample and presented in which chapter is provided in Table 3 below.

Table 3: Overview of research methods used in the analyses, which research question they address and in which chapter they are further described

Analysis (no)	**Analysis**	**Method (from list above)**	**Research Question**	**Chapter**
1	Coh-Metrix: Cohesion and Readability (quantitative)	1)	1a	5.2
2	Thematic analysis (qualitative)	4)	1a	5.2
			1b	5.3
3	Type of material (e.g. research, impact) (quantitative)	3)	1b	5.3
4	Lexical analysis through 2-, 3- and 4-grams (quantitative)	2)	1c	6
5	Appraisal analysis of Graduation resources (qualitative)	3)	2	7

Overall, this study makes empirical comparisons between different sections of the register based on quantitative analysis. These are contextualised and interpreted within a constructionist framework, using my perspective as participant in the impact infrastructure, as explained next.

4.2.3 The role of my professional experience

In addition to the formal research methods described in this chapter, this thesis draws on my activity as research impact consultant. Early findings from the quantitative and thematic analyses (1, 2, 4 in the list above, published in Reichard *et al.* 2020) allowed me to build a reputation as an expert on writing REF ICS. This led to consultancy activity starting in 2019, with a heavy focus on shaping the narrative, language and presentation of ICS being prepared for submission to REF2021 and later to REF2029. Through this activity, I gained first-hand insights into the process of writing ICS, from decisions on content and narrative, down to word choice. This gave me access to knowledge of how the texts themselves were assembled, constructed and revised, including insights into who was involved in this process and what considerations were made during writing and editing.

Being embedded in the community of professionals preparing ICS has influenced and enhanced my subsequent analyses based on manual tagging (item 3 in the list above, presented in chapter 7) and the situational analysis presented in section 5.1, as well as the interpretation of results from all analyses and the overall framing of the thesis.

As consultant, I also developed contacts with impact professionals and academics, through which I gained a sense of the questions that they were grappling with related to ICS. In early 2021, I formalised this anecdotal input with a short survey sent to my network and promoted via social media. The two questions asked in the survey, listed here, invited comments on what the corpus could be used for:

1. If you could ask this dataset anything, what would it be? E.g. certain kinds of words or grammar structures that you think may have been used differently in high- or low-scoring case studies, or specific differences between Main Panels
2. What kind of insights would you be interested in from this dataset? This will inform the direction of the next stage of my PhD research.

Twenty responses were received from a range of impact professionals and academic UoA leads. The results of this user survey did not play a determining role in decisions about research design, but some relevant insights are included in section 5.1.

4.3 Corpus sample and analyses

The register analysed in this study, and therefore the text basis from which the sample is drawn, is clearly defined: ICS submitted to REF2014. A total of 6,679 ICS from 154 universities are publicly available. An overview of how this corresponds to university size is provided in Grant (2015: 20).

ICS are freely available on the REF2014 Impact Case Studies database.[7] The HEFCE licence makes clear that this database is freely available for use in research: "researchers can […] make copies of any copyright material under fair dealing provisions, for example to conduct text and data mining. They can do this without having to obtain additional permission to make these copies from the rights holder" (HEFCE 2015a). The Intellectual Property Office of the UK Government also states that "an exception to copyright exists which allows researchers to make copies of any copyright material for the purpose of computational analysis if they already have the right to read the work (that is, they have 'lawful access' to the work). This exception only permits the making of copies for the purpose of text and data mining for non-commercial research" (Intellectual Property Office 2014).

As described in sections 2.1.4 and 4.1, the database has been used extensively for research, either as a whole or for drawing text samples.

4.3.1 Sampling approach

The composition of the corpus is one of the most important decisions in any corpus-based study. While any claims about extending findings beyond the texts included in the sample to other texts from the same register, language or genre have to be viewed with caution, applying principles and criteria to the selection of texts for inclusion increases the likelihood that findings may be more widely appliable. One of the main selection criteria often cited is representativeness (Biber and Conrad 2019: 58; McEnery *et al.* 2006: 13). Others are balance (McEnery *et al.* 2006: 16) and size (McEnery *et al.* 2006: 20), where larger corpora allow for more statistical power in quantitative analyses, while smaller corpora are more suitable for manual qualitative analyses. All three criteria were considered for each of the present

[7] The database can be accessed here: http://impact.ref.ac.uk/CaseStudies/

study's research questions and related analyses. Because the criteria have different weight for the various questions and methods, I created three different samples. Sample A (described in section 4.3.2 below) includes all identifiable ICS that are unambiguously known to have received the top score, and low-scoring ICS from those UoAs where top-scoring ICS can be identified. The aim was to achieve the largest possible sample for statistical analysis, prioritising the criterion of size as far as possible. Sample B (section 4.3.3) largely overlaps with Sample A, but prioritises balance in order to allow for a reading and comparison of whole ICS texts. Sample C (section 4.3.4), a balanced subset of Sample A, is smaller to enable manual textual analysis and tagging. For representativeness, all three samples include material drawn from all four Main Panels, and they are divided into high- and low-scoring sub-corpora in order to help distinguish features of high-scoring ICS as compared to low-scoring ones. Table 4 extends Table 3 by providing an overview of which of the three samples was used in which analyses, ordered by the chapter and section in which the analysis is described.

Table 4: Overview of which analyses use which sample

Analysis (no)	**Analysis**	**Sample**	**Chapter**
1	Coh-Metrix: Cohesion and Readability (quantitative)	A – all 4*	5.2
2	Thematic analysis (qualitative)	B – balanced full ICS	5.2
			5.3
3	Type of material (e.g. research, impact) (quantitative)	C – Section 1 selection	5.3
4	Lexical analysis through 2-, 3- and 4-grams (quantitative)	A – all 4*	6
5	Appraisal analysis of Graduation resources (qualitative)	C – Section 1 selection	7

The maximum possible sample for including ICS whose scores can be known, and therefore the application of the above criteria, was restricted by the way that REF results are reported. As stated in section 2.1.1, ratings are assigned on a scale from 4* to 1*. In order to explore the language of high-scoring REF2014 ICS and compare this to low-scoring ones, it was important to include in the corpus as many ICS as possible where the score can be clearly identified as either high or low. This was done on the basis of the official REF results spreadsheet introduced above (section 4.1.1), which shows the percentage of a submission that received a certain score (see Figure 4 in that section). For this corpus, ICS were defined

as high-scoring if they were part of a submission that had "100%" in the 4* column, that is, all individual ICS in the submission must have been rated 4*. ICS were considered low-scoring if the submission had 0% in the 4*, 3* and "unclassified" columns, thus identifying submissions where 100% of the ICS received either a 1* or a 2* rating. "Unclassified" ICS were not included because this rating was given for a range of reasons, not all of which indicate a lower quality of impact; the descriptor for a rating of "unclassified" is:

> "The impact is of little or no reach and significance; or the impact was not eligible; or the impact was not underpinned by excellent research produced by the submitted unit." (HEFCE 2011: 44)

The analysis of ICS in UoA22 (Social Policy) by Smith and Stewart (2017), briefly described above (section 2.1.1), uses a very similar sampling approach to the one outlined here. However, one key difference is that their exact approach (high = 100% 4*, low = 0% 4* and 0% 3*) does not exclude "unclassified", which allows them to include a further three ICS in their low-scoring sample. This UoA is one of very few where a meaningful sample can be obtained in this way because it has sufficient numbers in each of these categories.

4.3.2 Sample A: Quantitative linguistic analysis

The quantitative linguistic analyses (1 and 4 in Table 4) were based on a sample of all identifiable high-scoring ICS in any UoA (n=124) and all identifiable low-scoring ICS in those UoAs where clear-cut high-scoring ICS were available (n=93). The list of ICS included in Sample A is provided in Appendix A. Table 5 summarises the number of words included in this sample.

Table 5: Number of words in ICS included in Sample A, by Main Panel (MP)

Main Panel	High-scoring ICS – number of words	Low-scoring ICS – 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**

While 350,000 words may seem modest compared to multi-million word corpora such as the British National Corpus (BNC), Sinclair (2004: section 8) explains that specialised corpora can be much smaller. This is because less material is needed to ascertain the characteristics of

those texts than for large corpora that are designed to be representative of a whole language variety, as the characteristic language items are likely to have greater prominence in word frequency lists of specialised texts. Furthermore, as Koester (2010: 67) points out, small corpora allow the researcher to explore in detail the context of a search term in each of its occurrences and therefore to study its specific usage in the targeted type of text. Flowerdew (2004: 19) notes that "a corpus of up to 250,000 words can be considered as *small*", and therefore a corpus of the size used in this thesis can be of an appropriate size even though much larger corpora exist.

The specialised corpus compiled for the present study is designed to describe a specific academic register, namely "high-scoring REF2014 ICS", and to compare this to a related register, namely "low-scoring REF2014 ICS". To achieve what Biber (1993: 245) calls "situational representativeness", it includes all texts that can definitively be classified as part of the target register and therefore constitutes the largest possible dataset for making claims about the nature of this register. The sampling in this study is unusual in that in many other cases, it is clear or can be determined unequivocally whether texts are part of a certain register or genre.[8] For this study, while it is clear that all texts are ICS, it is simply not possible to ascertain which of the 6,679 ICS available are the ones that have received 4*-ratings – apart from the 124 texts that belong to submissions that were 100% 4*, which are therefore used. Biber's criterion of linguistic representativeness (namely at least 10 texts with at least 1,000 words per text, 1990: 159) is met: the parts of the REF2014 ICS included in the analysis (see section 4.4.1 on corpus preparation) usually contain between 900 and 1800 words (cf. McEnery *et al.* 2006: 21, who emphasize that the research question should drive the corpus composition, size and sampling strategy). Finally, this sample size allows a more qualitative component to be included in the predominantly quantitative approach of linguistic analysis: unlike Grant (2015), who used the whole database of over 6,000 ICS, the principled selection used in this thesis means that concordance lines could be consulted and examined in detail during the lexical analysis.

An alternative approach to including all identifiable 4* ICS is to aim for more balance across Main Panels, which could enable a clearer comparison between academic disciplines. With the strict criteria described above (section 4.3.1), the percentage spread across Main Panels

[8] Where this is not the case, more than one interpretation may be valid, e.g. a single entry in a lifestyle blog might be validly categorised as an example of health advice, a baking recipe, or a self-promotional text.

ends up being biased towards Main Panel C, with very low numbers in Main Panel B (see Table 6 below). However, in order to balance this, it would be necessary to relax the criteria for inclusion in a way that is discussed in 4.3.3 below for Sample B. Since the main axis for comparison in this study is to distinguish high-scoring from low-scoring ICS, with observations on disciplinary differences included where relevant, this balance across Main Panels was less of a concern for linguistic analysis than achieving a larger corpus size. Moreover, the academic fields within each Main Panel and even some UoAs are rather diverse, as can be seen in the provision for multiple submissions in one UoA by the same institution (HEFCE 2011: 11). Making claims about the linguistic characteristics of ICS from different Main Panels may therefore be misleading (see e.g. Gray 2015; Mauranen 2006; Thompson *et al.* 2017, on factors other than discipline that contribute to variation in academic language).

Similarly, it could be argued that the basis for this quantitative analysis should be a corpus which consists of UoAs where both clear-cut high-scoring and clear-cut low-scoring submissions are identifiable on the results spreadsheet and the respective ICS are available on the database, in order to avoid misleading results that may arise from certain UoAs that are represented in the high-scoring corpus but not in the low-scoring corpus. However, simply excluding those UoAs where no low-scoring ICS are available would have reduced the number of high-scoring ICS available by 35 (28%) from 124 to 89. Any quantitative results which were suspected to stem from a disciplinary skew were identified as such during a manual analysis of these results, as described in section 6.1.3 below. Moreover, the scores of Main Panel A are skewed notably towards the higher end, with 60.9% of all ICS at 4* (Manville *et al.* 2015a: 46).[9] It is therefore representative of the overall scores if my sample is skewed towards high-scoring ICS in Main Panel A. The composition of Sample A is therefore as set out in Table 6.

[9] This skew may potentially have been partly influenced by the scoring method as explained in Manville (2015: xiii), which allowed assessors to award on an 8-point scale to avoid truly excellent impact to be an upper benchmark for other outstanding research. This may have had the effect of setting an expectation of 4* as the middle mark.

Table 6: Distribution of ICS across Main Panels (MP) – Sample A

	MP A	MP B	MP C	MP D	Overall
High no.	44	6	37	37	124
High %	35%	5%	30%	30%	100%
Low no.	12	2	53	26	93
Low %	13%	2%	57%	28%	100%
High and low no.	**56**	**8**	**90**	**63**	**217**
High and low %	**26%**	**4%**	**41%**	**29%**	**100%**

One issue arising was that some institutions had made multiple submissions in the same UoA (cf. HEFCE 2011: 11). For example, in UoA13 (Electrical and Electronic Engineering, Metallurgy and Materials), Imperial College London made separate submissions for Electrical and Electronic Engineering and for Materials. Only one of these submissions (Electrical and Electronic Engineering) met the criterion of 100% 4* according to the results spreadsheet. However, the database makes no distinction between these submissions: it is searchable by several parameters including institution and UoA, but no further level of distinction is available to identify multiple submissions from the same institution in that UoA. It was therefore necessary to determine manually which submission each ICS in these combinations of institution and UoA that was returned by the database belonged to. For example, ICS from Imperial College UoA13 usually stated that they were part of the Electrical Engineering department somewhere near the beginning of the ICS, and those from Goldsmiths UoA35 made explicit in the Section 1 summary, if not the title, whether they were part of Music or Drama.

Sample A was used for various analyses that are described in section 5.2.3 (Quantifying readability) and chapter 6 (Lexical Investigation). The methods of analysis are described in detail in those chapters.

4.3.3 Sample B: Thematic analysis

In order to detect patterns of content and common characteristics of presentation in high- and low-scoring ICS across all four Main Panels, a sub-sample of ICS was selected from the full sample described as Sample A in section 4.3.2 for a qualitative thematic analysis (analysis 2 in Table 4). A more structured and balanced sample was used for this thematic analysis because the qualitative approach required a smaller sample than the quantitative approach, and a selection was therefore both possible and appropriate. This set, Sample B, included 60% of high-scoring ICS and 97% of low-scoring ICS from Sample A, such that only UoAs were

included where both high- and low-scoring ICS are clearly identifiable. Further selection criteria were then designed to create a greater balance in the number of high- and low-scoring ICS across Main Panels. Main Panels A (high) and C (low) were particularly over-represented, so a smaller number of those ICS was selected respectively. In addition, ten further high-scoring ICS were identified in Main Panel B. This was achieved by slightly relaxing the criteria for inclusion of submissions in Main Panel B, to include institutions where at least 85% of the ICS scored 4* and the remaining scores were 3*. As this added a further UoA (namely, UoA11 Computer Science), 14 more low-scoring ICS could also be included from that additional UoA. This resulted in a total of 85 high-scoring and 90 low-scoring ICS in Sample B, as set out in Table 7.

Table 7: Distribution of ICS across Main Panels (MP) – Sample B

	MP A	MP B	MP C	MP D	Overall
High no.	27	16	20	22	85
High %	32%	19%	24%	26%	100%
Low no.	12	16	36	26	90
Low %	13%	18%	40%	29%	100%
High and low no.	**39**	**32**	**56**	**48**	**175**
High and low %	**22%**	**18%**	**32%**	**27%**	**100%**

Appendix B lists all ICS included in Sample B. Table 8 provides an overview of the number of high- and low-scoring ICS from each UoA that is included in Samples A and B respectively.

Table 8: Overview of units of assessment (UoA) and ratings included in Samples A and B

Main Panel	Unit of Assessment Name	UoA Number	Sample A			Sample B		
			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**

Sample B was used to explore common features in high- and low-scoring ICS respectively that cannot be detected with quantitative methods. Thematic analysis was chosen to identify such patterns and infer meaning from qualitative data (Auerbach and Silverstein 2003; Braun and Clarke 2006; Saldana 2009). This process started as a collaborative parallel project with one of my supervisors (Reed) and four other researchers and was then integrated into the PhD project. The thematic analysis included the complete ICS, including Sections 3 "References to the research" and 5 "Sources to corroborate the impact" which were excluded from the analyses based on Sample A (see section 4.4.1 below). The process of analysis is described here because findings are referred to throughout the thesis.

To familiarise ourselves with the data and for inter-coder reliability, two research team members read a selection of REF2014 ICS from different Main Panels, before generating initial codes for each of the five sections of the ICS template. These were discussed with the full research team and piloted prior to defining a final set of themes and questions (reproduced in Table 9) against which the data was coded (based on the six-step process outlined by Braun and Clarke 2006). An additional category ("Structure and style") was used to code for features of presentation, to triangulate elements of the quantitative analysis (e.g. readability) and to include additional presentational features that are difficult to assess in quantitative terms (e.g. effective use of testimonials). In addition to this, ten different types of impact were coded for, based on Reed's (2018: 20-21) typology. There was room for coders to include additional insights arising in each section of the ICS 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.

Table 9: Themes and questions that guided the qualitative analysis of ICS, structured around the main sections of the REF2014 ICS template (titles in grey rows); those included in the thesis (chapter 5) are in italics

Theme or question
General information
Unit of Assessment
Impact types: understanding and awareness; attitudinal; economic; environmental; health and wellbeing; policy; other forms of decision-making and behaviour change; cultural; other social; 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 ICS?
Underpinning research and references to the research (Sections 2 and 3):
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 (Sections 1 and 4):
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 ICS?
Is public engagement a major component of this ICS?
To what extent does the ICS 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 (Section 5):
Examples of high- or low-quality corroborating evidence with justification for why it is considered 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. are they 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?

Coders summarised ICS content pertaining to each code, for example by listing examples of effective or poor use of structure and formatting as they were encountered in each ICS. Coders also quoted the original material next to their summaries so that their interpretation could be assessed by others during subsequent analysis. This initial coding of ICS text was conducted by six coders, with intercoder reliability assessed at over 90%, based on 10% of the sample. Subsequent thematic analysis within the codes was conducted by two of the co-authors (myself and Reed). This involved categorising coded material into themes as a way of assigning meaning to features that occurred across multiple ICS (e.g. categorising types of corroborating evidence typically used in high- versus low-scoring ICS).

Full results of the thematic analysis are published in Reichard *et al.* (2020) for an audience of those preparing ICS for the 2021 REF. This thesis includes selected findings from that analysis that are relevant to its research questions in chapter 5 (especially sections 5.2.2 and 5.3.2).

4.3.4 Sample C: Qualitative linguistic analysis (Section 1 of impact case studies)

As discussed in section 3.3 above, I use the Appraisal framework to research evaluative language in ICS (chapter 7). This section describes how the sample for that analysis was designed and selected.

Appraisal studies can have many different shapes and sizes: they can focus on just a small number of texts or a larger amount of data, they might be based on concordance lines and they can focus on one subsystem of Appraisal or include all three systems: ATTITUDE, ENGAGEMENT and GRADUATION (for an overview, see Tupala 2019: 4).[10] The system and the methodological approaches that can be employed to study Appraisal in texts are very flexible and can be adapted to fit the texts and research questions. The following examples illustrate how corpora of different sizes can be used in different ways for analyses under the Appraisal umbrella:

1. At one end of the scale, Tupala's (2019) corpus is just over 200,000 words; her whole PhD research consisted in developing a method for applying all three components of the Appraisal system to public policy texts and then coding the corpus using the UAM Corpus Tool (O'Donnell 2008). This is a fairly large corpus for an Appraisal analysis by comparison, so a smaller dataset can be appropriate for a project such as the present one where Appraisal analysis is one method among others. As Tupala highlights

[10] The components of the Appraisal system are represented in SMALL CAPS.

(2019: 2), the important factors for determining the quality of an Appraisal study are the level of detail in the annotation, which should be appropriate to the research question, and the representativeness of the corpus, rather than its size.

2. Yang *et al.* (2015) include 25 research articles, five each from five different journals, at an average length of just over 3,000 words. This means that their corpus comprises approximately 75,000-80,000 words (no exact figure is given). In their study, the unit of analysis is the clause, which means that one coding decision needs to be made for each verbal group (Yang *et al.* 2015: 4), and these decisions are also tied to certain linguistic forms (e.g. "modal auxiliary"). This speeds up the analysis because unitisation is pre-determined, rather than the coder having to decide which items in a text to include in the analysis, and tags need to be applied to a smaller number of larger units, compared to a completely open analysis.
3. Tavassoli *et al.* (2019) include 20 newspaper editorials, 10 each from two newspapers and sampled across two distinct timeframes (before and after a certain event). Their article does not include any information on the number of words in a typical text or in the corpus, just the number and distribution of texts. A quick check on the Opinion sections in both newspapers (14/06/2021) suggested that the average length of such an article might have been 1,000 words, translating into a corpus of potentially up to 20,000 words. The analysis in their study compared the stance of opposing sections of the British press towards refugees at two points in time (before and after a specific political event) and was focused on the ATTITUDE subsystem in order to determine a "welcoming" or "unwelcoming" stance in each editorial.
4. Ho and Crosthwaite (2018) investigate electoral manifestos of three candidates, a small corpus of less than 3,000 words overall. They apply all three Appraisal subsystems.
5. At the other end of the scale, Hommerberg and Don's (2015) analysis of wine reviews includes 20 texts with an average length of 87 words, resulting in a corpus of less than 2,000 words. They focus on one component of one of the subsystems of Appraisal only: ATTITUDE:APPRECIATION.

Manual coding of text in context creates richer qualitative data. Although this can be a labour-intensive process, the richer data means that the kind of sample needed for an Appraisal analysis can be smaller than for quantitative corpus studies, such as the keyword

analysis conducted on the larger Sample A in this study (chapter 6). In principle, such a smaller sample can be created either through selecting a smaller number of texts, or through focusing on parts of a text, which reduces the text length for analysis and thereby allows a larger number of texts to be coded and therefore included. For this study, a combination of both techniques was applied.

When choosing sections of a text for inclusion in a corpus, it is important to ensure that these are comparable to each other. The REF2014 ICS template has three text-based sections, and choosing one of these could provide for a straightforward delineation of text to include. Section 2 of the template describes the "underpinning research", rather than the impact, and is therefore of less interest for a study of impact writing. The options were therefore to focus on:

a) Section 1 (ca. 100 words "Summary of the impact");
b) a much smaller number of Section 4 texts (750-1000 words "Details of the impact");
c) extracts of Section 4 (e.g. the first 500 words).

The first of these options, Section 1, was chosen over partial or complete Section 4 texts for several reasons. First, this selection allows analysis of a complete unit of text, rather than navigating an arbitrary cut-off point if only part of Section 4 was used. Second, shorter texts enable the inclusion of a greater number of texts representing a greater number of ICS, which in turn increases confidence in the conclusions drawn. And finally, many studies of academic writing focus on introductions to research articles (e.g. Swales 1990 and studies following his model). A leading study in applying the Appraisal system to academic writing is Hood (2010), which also focuses on research article introductions. Her study shows that the attitude expressed at the beginning of the texts serves to establish the "value or significance" of the subject matter (2010: 52), a function that is also needed in ICS. Therefore, although the reading context and purpose of ICS Section 1 and research article introductions are different, they both constitute the start of the text and serve as a hook.

In order to assess feasibility, an initial analysis of one Section 1 was conducted in the UAM Corpus Tool (O'Donnell 2008), using the Appraisal coding scheme available on the Appraisal website (White 2003). This indicated that 15-30 minutes was likely to be a realistic time allocation for one Section 1, and this information was taken into account when designing the sample for the Appraisal part of this thesis.

The full sample for this project (Sample A) includes all available 4* ICS from REF2014, plus low-scoring ICS in corresponding UoAs, as described above in section 4.3.2. From this, a purposeful sub-sample for the Appraisal analysis was drawn to achieve a greater balance between high- and low-scoring ICS and between Main Panels. The following principles were used to define the Appraisal dataset:

1. Similar to Sample B used in the Thematic analysis (section 4.3.3), the Appraisal sample only includes UoAs where both high- and low-scoring ICS were available.
2. To ensure a balanced basis for comparison, the sample includes equal numbers of high- and low-scoring ICS within each of the chosen UoAs.
3. Often, the texts in a submission within a UoA or even within a university are homogenised; the sample therefore includes UoAs where it was possible to choose ICS from different submissions in order to include a varied spread of styles. This can be seen as equivalent to including one text per author, as applied in other studies (e.g. Dontcheva-Navratilova 2020: 14). It was not possible in all UoAs to include only one ICS from each submission, but this guiding principle was applied to select UoAs where this could be best achieved.

Applied to the 87 submissions in the full sample (Sample A), these principles result in Sample C consisting of three sub-corpora of broad disciplines, broadly corresponding to the REF Main Panels: Science, Social Science, and Arts and Humanities, with three UoAs included in each. UoAs, and ICS from these UoAs, were selected as follows:

Science: There are only four UoAs with both high- and low-scoring ICS (Principle 1) across all of Main Panel A and Main Panel B: UoAs 3 (Allied Health), 4 (Psychology), 6 (Agriculture) and 13 (EEE). Of these, UoA4 and UoA6 had the greatest number of separate submissions with 4* ICS, so these were chosen in order to enable a greater spread of ICS in the sample across different submissions (Principle 3). This is important to minimise skewing the results towards the particular style of one institution, as many high-scoring submissions were standardised at institution level. In choosing the third UoA between Allied Health and Engineering, I decided to include the only UoA from Main Panel B over a second health-related UoA (i.e. UoA3, given that UoA4 was clearly going to be included), which led to the inclusion of UoA13. The final sample includes four high-scoring ICS from each UoA, selected randomly from within each submission. There are only 12 low-scoring ICS overall in these UoAs and these were all included. Although this means that the balance across UoAs is not as even as

for high-scoring ICS, this was the best possible spread (Principle 2) given the restrictions of how REF results were reported and therefore the restricted pool of ICS to draw the sample from.

Social Science: In Main Panel C, there were five UoAs that met Principle 1: 20 (Law), 22 (Social Work and Social Policy), 23 (Sociology), 25 (Education) and 26 (Sport and Exercise Sciences). Principle 2 led to the exclusion of UoA26 with 2 high-scoring and 16 low-scoring ICS. Principle 3 pointed to UoAs 22 and 25, which both had several submissions in each scoring bracket, so they were both included in the sample to provide texts from a variety of institutions. In choosing between UoA20 (Law) and UoA23 (Sociology), disciplinary diversity of UoAs was taken into account, similarly to the Science sub-corpus. UoA20 was selected because it represents a broader disciplinary spread than if a UoA had been chosen that is closer to UoAs 22 and 25.

Arts and Humanities: Similar to the Science selection, there was not much choice. The ideal sample would include four ICS per UoA per scoring bracket (Principles 1 and 2), which was possible for only three UoAs in Main Panel D: 29 (English), 35 (Music), 36 (Communication). This selection also constitutes a representative spread across the varied subjects combined in Main Panel D.

As described, the Science sub-corpus includes UoAs from both Main Panel A (Life Sciences) and Main Panel B (Physical Sciences) because only one UoA in Main Panel B fulfils Principle 1 above, so it would have been impossible to have a separate Main Panel B sub-corpus. This UoA was therefore integrated into the Main Panel A corpus. Rather than being a limitation, however, this ties in with Biber and Gray's (2016) general distinction between "science writing" and "humanities prose". In their study, they combined subjects that would fall broadly into Main Panels A and B in their "science" corpus and those that are grouped into Main Panel D in their "humanities" corpus. They then added a separate "social science" corpus when they found that social science research articles did not fit neatly into one of the other two categories. This corpus by Biber and Gray (2016) marks an authoritative precedent for making this three-way disciplinary distinction (Science / Social Science / Humanities) for applied linguistics inquiry, independent of the epistemological or practical principles that may have informed the REF team's decision for panel combinations.

From each chosen UoA, four high- and four low-scoring ICS were included, resulting in 24 ICS for each sub-corpus. Due to the breadth of submissions in Main Panel C, an additional two ICS were included in both the high- and low-scoring sub-corpora. The Appraisal corpus therefore consists of 76 texts and a total of around 9,200 words (5,299 in the high-scoring sub-corpus and 3,941 in the low-scoring one). The full selection of ICS in Sample C is listed in Appendix C.

Additional ICS were needed for pilot coding and refining the coding manual. These were chosen from the UoAs included in the final sample but were randomly selected from the remaining ICS within submissions where there were more texts than needed for the analysis. For example, there are four high-scoring submissions in UoA4, so the final sample includes just one ICS from each submission, but the pilot sample includes ICS drawn from the remaining ICS in those submissions. It was decided to keep the pilot texts separate from the texts for analysis so as not to risk influencing the final coding through decisions that were previously made and then remembered, rather than being based on the coding manual at the time of coding.

This sample was used for the analyses described in section 5.3.1 (Type of material) and chapter 7 (Appraisal). Overall principles of analysis that are applicable to both parts are explained in the remainder of this section.

Reliability and replicability are not always a goal in Appraisal analyses, partly because of a belief that insights gained from the text(s) under consideration should not be applied to other texts. By contrast, the present study of language in REF ICS is intended for providing insights that *can* be applied to other, similar texts, either for future REF submissions or for other impact assessments. Fuoli (2015: 12) specifically commends the Appraisal system as a framework for decision making that can improve reliability not only between coders, but also within the same coder over time and across texts. To aid with this, he developed a seven-step procedure for annotating texts for Appraisal. At the centre of this procedure is the creation of a coding manual for a maximally reliable analysis, which can form the basis for maximally valid claims. For reliability, Fuoli lists three key questions to ask:

a. Would I make the same decision if I coded this text next month?
b. Would I make the same decision later today?
c. Would someone else make the same decision?

Replicability is especially difficult in evaluation research, but it is enhanced through the use of the Appraisal framework as a fine-grained guiding structure. It can be further increased through the use of explicit criteria and systematic recording of decisions where several options were plausible, to enable other researchers to interpret the annotation and findings in the same way. These steps were taken in the present analysis, while acknowledging that complete reliability and replicability cannot be achieved, especially in an analysis that is mostly qualitative.

The remainder of this section describes Fuoli's (2015) seven-step process and how it was applied in this study.

1. **Define the scope (first draft of coding scheme)**

   The first step is to define the goals of the project, and specifically which part of Appraisal is used in the analysis. For this project, the scope was defined through a literature review (section 2.2) and a situational analysis (section 5.1), which point towards GRADUATION being the most interesting aspect to explore in ICS (see above section 3.3).

   Similarly, the degree of delicacy is determined by the project goals and the text features. For example, those features that appear more often may be coded to a more delicate level, especially in GRADUATION:FORCE:QUANTIFICATION. A smaller text base also allows a more delicate analysis. With the ICS text base being small enough for the most delicate coding that is helpful for the purpose of the analysis, the exact level of delicacy in each branch can be determined during refinement of the coding manual.

   Fuoli recommends an informal exploratory analysis as a first step, in which instances are underlined manually in some texts and an initial coding scheme devised on that basis (Fuoli 2015: 16). As this was the final analysis conducted on the same texts (after readability, thematic and lexical analysis) and I was already familiar with the texts, the initial coding scheme for ICS was created by consulting three different coding schemes that applied the Appraisal framework to academic texts (Hood 2010; Martin and White 2005; Xu 2017), combining and adapting them while considering how they could be applied to ICS (details in section 7.1.2).

2. **Decide on the tool to use**

   Fuoli introduces two software options, of which the UAM Corpus Tool was chosen for this study. One advantage is that it is desktop-based, rather than web-based, which was preferable for the practical situation of this study. Other factors in the decision in favour of this tool were its widespread use, the availability of tutorials and training material, and the pre-existing Appraisal coding scheme which can be adapted to my coding manual, rather than having to be created from scratch.

3. **Create a coding manual**

   The coding manual is at the heart of the endeavour to make a study reliable, replicable and therefore its results as generalisable as possible. It needs to be context-specific, that is, shaped around the texts in the study, and it needs to make the assumptions of the texts' audience and the communicative purpose explicit. Fuoli and Hommerberg (2015) provide an example coding manual, the structure of which was followed for this study. Once drafted, the robustness of the manual needs to be tested on another random sample and adapted before moving to the next stage, leading to several repetitions of Steps 4 and 5. See Appendix F for the coding manual created for this study and section 7.1.3 for a description of how this process was conducted.

   The manual should contain:

   - Explicit rules for applying the definition to the texts in the study
   - Outline of all coding schemes that are used in an analysis
   - Definitions of each category label to help the coder recognise evaluative instances especially where these are invoked, or where an otherwise neutral term is likely to be seen as positive or negative in the specific context (here: UK REF assessment)
   - Illustrative examples accompanying the definitions – here, Martin and White (2005) and Hood (2010) served as starting points but were supplemented with context-specific examples from the study corpus and, where helpful, explicit discussion of any background assumptions that the original writer and the target reader are likely to share but which may not be obvious to a coder less familiar with the context.

4. **Assess reliability**

   Two basic types of reliability need to be established: intra-coder (reliability over time) and inter-coder (reliability across people). Intra-coder reliability is the first step, ensuring stability of decision making. Once the coding scheme and other details in

the manual are refined enough for a coder to make the same decision on a separate day, the focus can shift to inter-coder reliability. This is important for replicability as it establishes whether the coding manual is explicit and detailed enough for another researcher to come to similar-enough conclusions to the writer of the manual. This in turn is important for the validity of claims that are made on the basis of the analysis. For this step, stratified random samples from the full corpus should be used, where possible. Ideally, these should not be part of the corpus that is ultimately used for analysis. In this study, three small pilot corpora were created as explained earlier in this section; see below section 7.1.3 for details of how they were used.

5. **Refine manual**

   The pilot corpora were used for these steps recommended by Fuoli:

   - Refine definitions
   - Modify or add rules or categories
   - Add more examples, both to the glosses in the coding scheme in the UAM Corpus Tool itself and to the coding manual
   - Discuss with other coder

   Informal assessment of reliability was done throughout this process, creating a loop between Steps 4 and 5. The process was completed with a formal assessment of inter-coder reliability on the basis of Pilot Corpus 3, as described in section 7.1.3.

6. **Annotation**

   All the steps in the process up to this point are important preparation for this main step. At this stage, it is important to constantly be aware of ensuring internal consistency, especially if the corpus is coded by one coder only and therefore little external accountability exists. This means following the manual closely and taking regular breaks to minimise fatigue and cognitive load. Fuoli suggests considering separate coding rounds for identifying and classifying instances, which would provide a built-in opportunity to re-evaluate the first decision. However, identification is aided by thinking through possible classification options, and therefore these steps are combined here. To maximise reliability when doing both steps in the same coding round, a different review mechanism was used, where the coder completed the annotations over the space of five days and regularly searched for newly-coded words in previously-coded text to review all instances. The annotation was further refined at the exploratory stage of the quantitative analysis when annotations of a

certain tag were viewed side by side and very occasionally adjusted. Any statistical analysis was done after completing this post-coding check.

7. **Analysis**

   The process of statistical analysis of coded features is described in section 7.1.4.

Further details of how the Appraisal analysis was conducted are provided in section 7.1, and the features of the coding scheme are detailed in the coding manual (Appendix F).

4.4 Corpus preparation

Once the most comprehensive sample, Sample A, was determined, files had to be prepared for analysis. This included three main steps: preparing the text files, manual tagging of presentational features, and automatic tagging of grammatical features.

4.4.1 Preparing the text files

The ICS were downloaded in PDF format from the database by searching for ICS from a given UoA and institution, as described for Sample A. The database of REF2014 ICS offers fewer features than the one that was built for REF2021 ICS, and at the time of downloading the files, PDF downloads were the only option.

PDF files of ICS were converted to plain text (TXT) using AntFileConverter (Anthony 2017). Notepad++ (Ho 2016) was then used to prepare the TXT files for textual processing. The following steps of manual editing were applied:

1. Sections that had been displaced during the file conversion were moved to their correct location. These could be any length from single words to whole paragraphs.
2. Bullet points were transformed into numbered lists ending in full stops, because this punctuation signal is used by the MAT tagger (Nini 2015, see below section 4.4.3) to assign several tags for which the beginning of a sentence needs to be clearly detectable.
3. Superscript footnotes in the PDFs that appeared as normal text after conversion to plain text, effectively adding a random number or letter to a word, were put in square brackets to avoid a situation where they create hapax legomena and the words to which they were attached would not appear in searches when they should.
4. **[text removed for publication]** notes, which were used in ICS in the publicly available database to indicate text that was redacted before publication, were replaced with **XXXX**, following the practice of Grant *et al.* (2015).

5. All material that was part of the REF template document, for example the instruction “indicative maximum 100 words”, was removed.
6. Text included in figures and tables that was preserved in the converted text was removed, as it appears for example in Figure 5. Captions were kept because they were seen as prose where more language choices were possible than inside tables, as illustrated in Figure 6. These were treated as part of the text of the ICS.

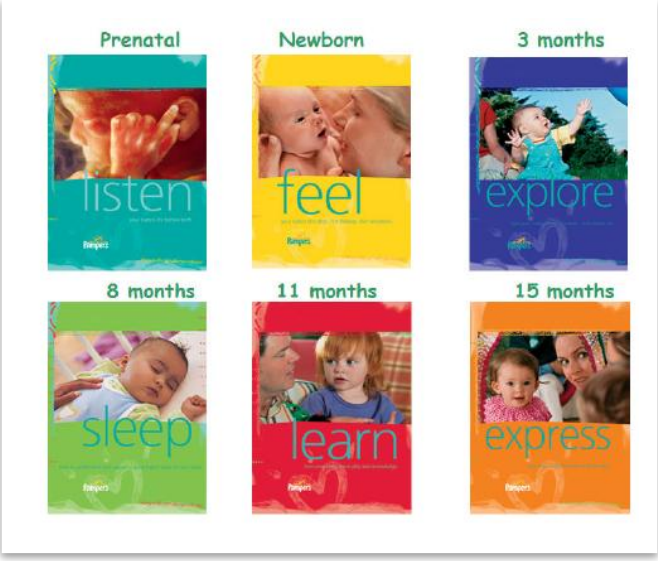


Figure 5: Example of ICS figure where text was considered part of the figure and therefore removed

Figure 6: Example of ICS figure where text was seen as caption written for the ICS and therefore kept

Sections 3 (“References to the research”) and 5 (“Sources to corroborate the impact”) were removed completely. Although some ICS included additional material in these sections, this was not consistent. Deciding what constitutes prose where language choices could be made, and what is simply a listing of a research grant, could have become complex, and therefore all Sections 3 and 5 were deleted entirely. The files used for analysis therefore only contained the main text of sections 1, 2 and 4.

4.4.2 Manual tagging

From the initially cleaned files, multiple versions were created to allow for different types of tagging to be applied, alongside a cleaned but untagged version. In one set of files, the text structure was annotated using Notepad++ (Ho 2016). This annotation enabled searches for specific tags as reported in section 5.2.2, but also facilitated the qualitative analysis reported in section 6.2.3 by indicating the position of a term in the text.

The following items indicating text structure were tagged with the XML tags shown in bold:

- Sections of the ICS template **<s1> <s2> <s4>**.
- Paragraphs **<p>**: This tag was applied to paragraphs. It was also used to wrap around lists if the list items were sentence fragments that were grammatically dependent on text in the paragraph.
- Lists **<list>**: This included bulleted or numbered lists where each item starts on a new line, rather than lists in-line within a paragraph. Sometimes in ICS sections 2 or 4, the whole section was formatted as an indented list and therefore the numbered headings were treated as the headings of the list items. If the sections were structured with numbered headings but these were not formatted as lists, they were treated as headings in their own right, rather than treating the whole section as a list.
- List items **<item>**: This tag was applied to each item in a <list>, with <p> nested inside in cases where the item was at least one full paragraph.
- Lists in a list **<ul2>**, usually part of an <item>.
- Headings **<heading>**. If the heading was in-line, rather than having the paragraph start on the next line, the <p> tag was wrapped around the heading, e.g. <item><p><heading>. Where there were noun phrases followed by a colon at the start of a paragraph, this was treated as an in-line heading.
- Captions **<caption>**.

4.4.3 Automatic tagging

Beyond lexical frequency, which can be usefully extracted from untagged files, other textual features can be studied in a corpus with part-of-speech (POS) tags or other functional or grammatical tags. The MAT tagger (Multi-dimensional Analysis Tagger, Nini 2015) was used here to automatically tag the TXT files for 67 features that were part of the multi-dimensional analysis originally proposed by Biber (1988). It performs two consecutive steps. First, the integrated Stanford tagger,[11] a log-linear tagger that tags the TXT files according to the Penn Treebank tagset (Toutanova 2003), creates a set of POS-tagged files. The MAT-component of the program then adds further tags to these POS-tagged text files according to the features used by Biber's (1988) study, with some amendments described in the tagger manual (Nini 2015).

Although these steps are separate inside the program and separate sets of tagged TXT files are provided, it is not possible for the user to edit the POS-tagged files before the MAT tags are added. Both tagger components, Stanford POS and MAT, were checked to determine whether they are sufficiently reliable when applied to the kinds of text in this corpus. The accuracy of the POS tagger was manually checked in parts of three ICS (31, 32 and 33 – part of UoA 4, which was used in the pilot phase), together comprising around 2500 running words. There were two issues that arose as a result of the tagging process:

1. "that" or "which" introducing a relative clause was tagged as "WH-Determiner", where it should have been tagged as "WH-Pronoun"; this can easily be solved by searching for "that_WDT" and manually checking which of these instances should be replaced by "that_WP" if a search on these POS-tagged texts is done. This was ultimately not necessary in this study because the POS-tagged files were not used directly for any further analyses.
2. In some cases, verb tags were assigned where adjective tags should have been used. Past simple, past participle and adjectives, that is, words ending in "-ed", were sometimes not distinguished correctly from each other, and some nouns ending in "-ing" (e.g. "eyetracking") were wrongly tagged as present participles. These issues cannot be rectified semi-automatically, and the small number of incorrect tags compared to the vast number of correct tags does not warrant manual checking,

[11] Available at https://nlp.stanford.edu/software/tagger.shtml

especially as these POS-texts are not designed for analysis and the MAT tags are assigned before any post-tag editing can take place.

In addition, miscellaneous mistakes occurred, for example the noun “viewer” was tagged as a comparative adjective and the noun “gaze” was tagged as a verb. Overall, 65 errors were marked in 2500 words across three texts, which is an accuracy of 97%. Results are consistent within and across texts. The MAT component of the tagger was checked on ICS 31, which contained 49 errors in a total of 1350 words (96% accurate).

After tagging, the program plots the input corpus, that is, the register under investigation, against other registers on the five dimensions defined in Biber (1988). The result is provided in Figure 1 in section 3.1.1 above, to illustrate the similarities and differences in these dimensions between ICS and other research writing. Specifically, as shown in Figure 2 in section 3.2.3, the placement of ICS on Dimension 4 *Overt expression of persuasion* shows the low rate of persuasive language that can be detected in ICS using the Biber framework of multi-dimensional register analysis. As discussed in section 3.2.3 above, this is at odds with the perception of users that ICS is an example of a persuasive register, and this discrepancy calls for other research methods to explain this. These are described in section 6.1.6 (findings in section 6.2.3) and chapter 7.

4.5 Chapter summary

In this chapter, I first described previous research on ICS to consider how the challenge of researching differences across scoring brackets without direct access to the scores of individual ICS had been approached, and what kind of textual analyses had been attempted. I then set out my overall research design, before explaining the composition of the corpus for this thesis. Analysis processes that are relevant for more than one Results chapter were also described, but most details of analyses are included in the relevant chapters (5-7). Finally, I provided technical information on preparing and processing the text files for analysis.

Chapter 5 Context and Content of Impact Case Studies

The comprehensive analysis of ICS starts with a close look at the situation and circumstances of these texts. This begins with a broad situational analysis (section 5.1) as described in the Register literature (see above section 3.1), followed by a more in-depth discussion of the "processing circumstances" (Biber and Conrad 2019: 43): the writing and reading processes respectively (section 5.2). The situational analysis (section 5.1) and discussion of production circumstances (section 5.2.1) are predominantly based on the literature around research impact and REF introduced in chapter 2, while the discussion of reading circumstances is introduced with observations from a thematic analysis (Sample B, section 5.2.2) and then centred around a quantitative textual analysis (Sample A, section 5.2.3). Additionally, all parts of this analysis are informed by my own experience as a reader and writer of ICS, as described in section 4.2.3 above.

The chapter then moves to the types of content and themes that can be found across ICS from different topical domains (section 5.3). It first reports on the type of material that can be found in Sections 1 of ICS (Sample C; section 5.3.1), and then provides evidence from the Thematic analysis on how claims were made and evidenced in the complete ICS (Sample B; section 5.3.2). Discussions of findings are integrated throughout the chapter.

5.1 Situational analysis

A systematic analysis of the situational context of a register can provide insights into the factors that may play a part in shaping the distinctive linguistic characteristics of its texts. Biber and Conrad (2019: 40) suggest a list of situational characteristics that should be examined in such an analysis, including variables such as the circumstances of production, mode of communication (written/spoken), audience (size, degree of shared background knowledge), and purpose (e.g. information, persuasion, entertainment). Registers can be analysed in relation to these characteristics at different levels of specialisation (Biber and Conrad 2019: 32); for example, for the present study it could be increasingly specialised contexts from (1) "Academic writing" as the most general level, to (2) "Impact case studies" as a specific subset of writing in the academic community, and then (3) "Humanities ICS submitted to REF2014" as an even more specific subset of impact case studies. Since the situational context of ICS is fairly homogenous across the register, but differs from other academic writing, the level "REF impact case studies" (as opposed to focusing on a subset, or

including other kinds of impact case studies) is chosen here as the level of analysis. Regardless of any potential differences in the linguistic analysis between certain sub-corpora of ICS, the only substantive differences between those in the situational analysis may be found in the "Addressor" and "Production" categories (see section 5.2.1), but both are differences where it is not known where individual texts fall. Therefore, the different possibilities are described when outlining the situational characteristics of ICS, but then treated as elements of a single "situational context", representing the whole register. No distinction is made between high- and low-scoring ICS in this section for this reason.

The following overview juxtaposes the situational context of research articles offered by Biber and Gray (2016: chapter 3) with my own, more specific analysis of the situational context of REF2014 ICS. Following Biber and Conrad's (2019: 38) suggested sources of information (namely, own experience, expert informants, previous research), this analysis is based on the review of the literature on ICS in chapter 2, on insights from the user survey conducted in February 2021 (see section 4.2.3) and on knowledge gleaned from my own consultancy work and conversations with colleagues responsible for their universities' REF submissions. The overall comparison is followed by a more in-depth discussion of the "addressor" category in the next section of this chapter (5.2.1), that is, the question of who exactly writes ICS.

Out of the characteristics that constitute a situational analysis according to Biber and Conrad (2019: 40), those that are most relevant to ICS fall under the following main headings: (a) Participants, (b) Relations between participants, (c) Processing circumstances, and (d) Communicative purpose. The three additional characteristics of Channel, Setting and Topic, also set out in Biber and Conrad (2019: 40), are less relevant to ICS but are briefly addressed here as "additional variables" for completeness. All variables are first described briefly for the research article context discussed by Biber and Gray (2016: chapter 3). Their characteristics and relevance are then considered in relation to ICS in order to illustrate how they are a different register compared to more research(er)-focused academic writing. At the end of this section, I will place ICS in a matrix of registers and highlight the implications of this analysis for the level of persuasive language that can be expected.

a. Participants

Addressor(s): Research articles are generally written by academics, but depending on the discipline, the writing process is often a collaboration with other academics at the same (or other) universities, and potentially with others who provide further external help. By contrast, the entity that addresses the ICS is technically the university submitting to the REF, or at a more fine-grained level, the submitting unit of assessment at that university, and only in exceptional circumstances would collaboration with other universities be involved, given the competition element of REF as explained in section 2.1.1. The stated addressor of a given ICS may be the researcher(s) or research team(s), or an academic department. The actual addressor, that is, the person(s) writing and otherwise preparing the ICS, may be a combination of the researchers themselves in conjunction with a professional services colleague or team within a university, or internal or external professional writers. The stated addressor and the writer may therefore differ, and there are likely to be several writers and additional editors involved in the process of drafting and revision (see 5.2.1 below).

Addressee (audience): Research articles have a potentially large audience within the specific academic discipline, although this varies hugely (see e.g. Congleton *et al.* 2022: 104). ICS have a very small primary audience (2-4 assessors, see Manville *et al.* 2015a: 15) composed of usually senior university researchers, as well as professionals from outside of academia called "research users" in the REF documentation, that is, "members from the private, public or third sectors with expertise in commissioning, applying or making use of research" (HEFCE 2010: 5).

Onlookers: In addition to the primary audience, there may be secondary audiences that are not the stated addressee but can influence the editorial decisions of the addressors. For research articles, these "onlookers" are not stated explicitly, but they might include university press offices and future hiring committees, as well as the journal editors and reviewers acting as gatekeepers for publication. In the UK, this can also include internal REF assessment panels, and then the actual REF panels, through the pressure to think about future REF scores when writing articles (Tusting 2018: 490). There is thus an opportunity for the secondary audience to influence the writing directly (in the case of reviewers) or indirectly, as there is a power differential between some of the onlookers and the addressors. For ICS, the onlookers include users of the published ICS database, such as other researchers, universities or funders. There are also likely to be university-internal onlookers

such as marketing departments looking for impacts to feature in communications, or career progression panels assessing the contributions of individual researchers. None of these onlookers are likely to influence the language of ICS; the primary focus is on the official addressees because the stakes are by far the highest with this audience. However, as ICS consultant, I encountered one example in early 2021 where a university wanted to take account of the "onlookers", in this case the non-academic, public readers of the ICS after assessment. The text in question was written in a style with many short sentences that interrupted the reading flow, and I had advised combining some of them, partly in order to avoid repetition in the grammatical/syntactical subject of consecutive sentences. I made this recommendation based on comments in the impact literature where panel members were reported as not appreciating text in a style that looked oversimplified (e.g. McKenna 2021: 18). The university explained that "We've been trying to improve the Flesch readability scores… *more for the benefit of the general public who will read the case studies than the assessors*." (direct email communication, my emphasis).

b. Relations between participants

Interactiveness: Neither ICS nor research articles invite much interaction between addressors and addressees, although for articles, peer review before publication is part of the established process, and communication about the content outside of the publication is common (e.g. through presentations and informal conversations at conferences, or on social media platforms). With ICS, there is no formal interaction established or required, and the relationship is even more one-directional. The addressor may have a rough idea who the reader will be through lists of assessors in each UoA that are published in advance, but there is no guarantee that their assumed assessor will be part of the actual audience (see Manville *et al.* 2015a: 15). The overt addressor is known, but the writer may be different, especially if the "addressor" is understood as the "university". The addressee must treat the pages of the ICS as the complete set of information, without taking into account any additional information they may otherwise have. This includes information that is signposted within the ICS: while Section 5 of the template lists "Sources to corroborate the impact" and assessors may access this material, as well as the "Underpinning research" referenced in Section 3, this is for audit and verification purposes only and they are not allowed to take any of this information into account when forming their rating decision

Social roles: Research articles are addressed to the addressor's peers (Hyland 2002: 219). In principle, this is also true for ICS, but there is more variation on both sides: the addressor may partly be a non-researcher-writer, and the addressees include non-researcher-readers, who are not peers of each other nor of the academic researchers. An addressor may sometimes be a junior or mid-career researcher, but this is (even) less likely for the addressee. The addressee's role as assessor creates a significant power differential.

Specialism: Writers and intended readers of a given research article have a high degree of shared background knowledge, but the same cannot be assumed for ICS. Some shared background knowledge may be assumed for part of the text, namely Section 2 "Underpinning Research", as there is at least a rough disciplinary proximity between the addressor and the addressee. However, the assessors are unlikely to be specialists in the exact same subject area (Manville *et al.* 2015a: 15), and the impact sections may vary widely, to the point of being rather independent from the research area. For example, one ICS in UoA4 (Psychology, Psychiatry and Neuroscience) reported on impact on the use of colour in grassroot football games (Chichester, *The influence of colour in the appraisal of visual information by professionals and others*).

c. Processing circumstances

Production: Research articles are written mostly by academics and often over a long period of time, often without a firm external deadline. ICS are written by a mixture of academics and professional writers, as described in the discussion of professional writers in the literature (section 5.2.1 below); for example, I was involved directly in the writing of six ICS for 2021 as impact consultant. Like research articles, they are heavily planned, edited and revised, often over the span of several years.

Reading: Research articles can in principle be read anywhere and at any time by many readers with a variety of reading purposes. Some will read the text very carefully, while others will only skim it, and individual readers may focus only on certain sections (e.g. methods or results). The primary addressees of ICS, however, have only limited access for assessment purposes. This includes very careful reading of the text and discussion with fellow panellists. While the vast majority of ICS from the 2014 and 2021 REFs have been made available to onlookers who can now read the texts for other purposes and with fewer restrictions (Hinrichs and Grant 2015), it is unlikely that this post-assessment publication had

much influence on the writing of ICS given the high stakes of the assessment itself (but see the example above under "onlookers").

d. Communicative purpose

General purpose: Research articles are mostly informational, based on specific topics. ICS in their most fundamental formula describe past events and their effects (research and impact).

Specific purpose: Research articles communicate research findings to peers. The goal of this is the creation and negotiation of knowledge, and the role of persuasion in this context is to convince peers of a proposition for advancing the knowledge base, establishing a consensus and perhaps improving the world through impact. The specific communicative purpose of ICS is to describe links between research and its effect outside of academia, and to persuade (Wróblewska 2021: 5) the reader that these links are strong and the impact is "significant" for a narrow purpose of being successful in an assessment (see section 2.1.1 above for a discussion of the explicitly stated purposes of REF). The persuasive element is much more directly part of the purpose. Note that the other REF criterion, "reach", is more descriptive than "significance", hence less persuasion is needed. Some persuasion can be added to claims of "reach" by providing context to explain why a certain reach claim is impressive.

Purported factuality: Research articles are generally presented and perceived as factual, although depending on the discipline, they may include arguments or an explicitly well-founded opinion. ICS purport a high degree of factuality, as they are written specifically for the assessment of their content. However, they are also promotional which is also driven by the assessment context, and there may be some opinion included especially in testimonial quotes.

Stance: Research articles are low on grammatical stance markers (as introduced in section 3.2.3); see Biber and Zhang (2018) for a discussion of the discrepancy between their own research showing academic writing as not displaying overt stance, but also the large body of research (e.g. Afros and Schryer 2009; Hunston 1994; Hyland 2005b) arguing that academic writing is inherently evaluative and showing overt stance. The literature on ICS (e.g. Watermeyer and Hedgecoe 2016) assumes a high prominence of stance because these texts are portrayed as "selling" impact; similarly, in my user survey (introduced in section 4.2.3), responses from impact professionals and academic UoA leads indicated that they assume

that words like “novel” or “unique” are frequent. However, the role of stance markers and other evaluative language warrants further investigation, which will be described in chapter 7.

e. Additional variables

Channel: Research articles are usually consumed in the written mode, which may include reading a paper copy, HTML or PDF. For ICS, there is no free choice for the direct addressees, who must read the texts on screens unless they specifically request a PDF print-out (source: personal communication with an assessor). Onlookers of ICS can choose to view the text as HTML on the database or print PDFs and therefore hard copies as they wish.

Setting: Research articles may be read with a considerable time difference between writing and reading, even many years later. They may also be read in very different geographical settings. ICS, by contrast, are read by the primary addressee within a year of the text being finalised and submitted, so phrases like “over the last 18 months” are still meaningful at the time of the addressee’s reading (although the interpretation may still be unclear, i.e. whether this is relative to the time of reading or of writing, which was a difference of at least 7 months in REF2014 and likely more in many cases). They are also read in the same country, so references to “national and international” can fundamentally be understood in the same way by addressors and addressees (although the actual meaning of “international” may vary, and even “national” may refer to either the UK as a whole or one of its constituent countries).

Topic: The topic of research articles is variable by (sub-)discipline. In ICS, the topic is very roughly determined through the UoA in which they are submitted, but only for the underpinning research section. The “impact” context and REF guidance may influence lexical choice.

On the basis of the variables that constitute a situational analysis, rather than the linguistic features, Biber and Gray identify a matrix with two main parameters along which registers differ: *informational purpose* and *specialised audience* because “communicative purpose and audience have a major influence on grammatical discourse style within the written mode” (2016: 109). Compared to, for example, newspaper writing (high informational purpose but low specialised audience), all academic writing can be placed firmly in the high informational purpose / high specialised audience quadrant, with research articles in an extreme position,

followed by academic books and postgraduate and undergraduate textbooks (Biber and Gray 2016: 69).

Based on the situational analysis described above, in this framework ICS are still closest to research articles as illustrated in Figure 7, because both have a primary audience (addressees, assessors for ICS) which expects specialist discourse and does not want the style to be "dumbed down" to a "journalistic" level (McKenna 2021: 54). The informational content is highly dense, perhaps even more so than in research articles which in many cases are not quite as restricted for space as the 4-page impact case studies.

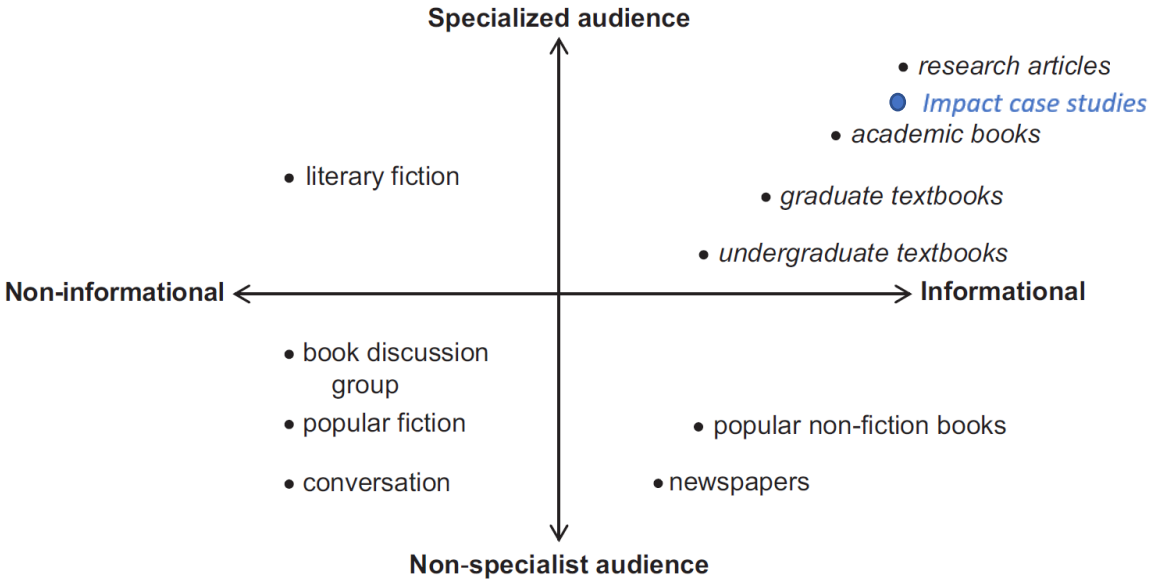


Figure 7: Academic and non-academic registers along two situational parameters (adapted from Biber and Gray, 2016: 69; reproduced with permission of Cambridge University Press through PLSclear)

Following the comparison and especially the different communicative purposes of research articles and ICS, it can be concluded that elements of persuasion are to be expected in these texts. ICS are even more space restricted than research articles, which, combined with the purported factuality for assessment purposes, poses problems for writers seeking to encode an amount of persuasion that presents ICS in a good light without skewing the assessment. Chapter 7 uses the Appraisal framework to examine how this is done.

5.2 Writing and reading impact case studies

After the overview of the general situation of ICS presented in the previous section, this section zooms into the processing circumstances, that is, reading and writing. As indicated above, there is often a discrepancy between the stated addressor and the actual writer of

ICS, and this fact has drawn some criticism which will be discussed and evaluated in section 5.2.1. The section then moves to the reading experience, initially reporting on findings from a qualitative thematic analysis I conducted with five other coders which includes their subjective experience of reading (section 5.2.2), before turning to a quantitative analysis of textual cohesion and its potential effect on the reader (section 5.2.3).

5.2.1 Who really writes impact case studies?

The overt addressor of an ICS is the submitting university or department, or perhaps a researcher or research group. This would suggest that these texts are prepared by those researchers, or at least by staff within the submitting university. However, this is not always the case, and there is great variation between and within universities. Based on my own experience in supporting REF2021 submissions, ICS at research-intensive or specialist universities were sometimes written mostly or completely by internal or external impact professionals, compared to post-1992 universities where ICS preparation often (not always) rested with the researchers, with occasional feedback from the small number of professional service staff available, with varying degrees of knowledge about ICS or impact. In the earlier 2014 submissions, this disparity is likely to have been even greater, because not all institutions realised how high the stakes were. Some institutions therefore did not prioritise resources for preparing ICS, while others may not have had the resources to prioritise ICS despite recognising their significance. Perhaps as a result, in 2014, submissions from research-intensive universities were often highly standardised, while those from newer universities sometimes appeared hastily put together, often by academics with little guidance or support.

This discrepancy has drawn many researchers and commentators such as Penfield *et al.* to assertions that the case study approach "rewards those who write well or *who can afford to pay for external input*" (2014: 29, my emphasis). In fact, in addition to creating (sometimes temporary) impact officer posts in preparation for REF, many universities invested heavily in external writing assistance (Coleman 2019: 13-15) to ensure that ICS were "easy to understand and therefore evaluation-friendly" (Watermeyer and Chubb 2018: 2) for the assessment panels. Grant *et al.* (2015: 17) assume that "the use of professional writers" has influenced the vocabulary. Similarly, McKenna (2021: 18) points to journalists, communications or press offices that drafted or edited ICS in some universities, albeit with mixed success, mainly due to the danger of overselling with "marketing speak" or

oversimplification, which may be perceived as an insult by assessors who are "quite proficient in their respective topic areas". However, it is often unclear whether the term *professional writers* in these criticisms refers to a university's marketing team or press office, or to external impact consultants, or both. This is an important distinction: as consultant, I have seen edits from a press office in an ICS I reviewed. Most of these were not helpful because in an attempt to achieve brevity, they oversimplified the message for a different, more general audience and thereby jeopardised the precision required for specialist assessment. This experience supports McKenna's (2021) criticism, if indeed professional writers are understood as those writing professionally for the public, rather than impact specialists.

A different angle is raised by Brauer *et al.* (2019: 68), who see the role of professional writers as a "huge ethical issue" because they are hired to "craft high-impact case studies", implying that the intent is deceit by presenting an impact as more significant and far-reaching than it is, through writing craft. However, it should be noted that the use of professional writers was not explicitly banned by the REF guidelines. The purported ethical problem is then that writers who work across UoAs are not embedded in the values and discourse of the UoA for which they are writing in the same way as the academic would be. However, in my experience of working with academics across UoAs and with professional writers, some writers specialise in certain disciplines where they may be socialised academically (e.g. through a PhD in that discipline), and all consultants that I know about work closely with the respective academics where at all possible (unless, for example, the researcher has left the university and this is the reason why a different writer was brought in). It is possible that Brauer *et al.* have similar concerns to Moran and Browning (2018: 258), who assert that external writers were hired to "sex up" ICS where there was a lack of evidence, and caution against "advertising gurus who hold that evidence and analysis are insufficient". This may be true for some "advertising gurus", but impact consultants generally emphasised the primary importance of the impact and see it as their role to make the available evidence and analysis visible and noticeable, so that an ICS can get the credit it deserves (e.g. Reichard 2024), rather than holding that "evidence… [is] insufficient" (Moran and Browning 2018: 258).

Brauer *et al.* (2019: 68) also point out that academics may "lack the presentation and writing skills to present their impact", which would provide a reasonable explanation for drafting in help. By contrast, Dunlop (2018: 289) extrapolates from the 2014 ICS in UoA21 (Politics) that

academics are quite able to express themselves and that therefore "any outsourcing of impact writing to commercial companies may be at best unnecessary and at worst contrary to the original intent of impact's introduction to the audit". However, this argument is circular since commercial writers may well have been involved in the creation of the ICS she inspected and on the basis of which she claims that academics do not require writing assistance. Moreover, even though "the audit" has various purposes (Dunlop does not specify what she reads as "the original intent", but see section 2.1.1 above on the purposes of impact assessment in REF), an academic's writing skill is not part of the assessment construct. Therefore, the question of who is involved in achieving a convincing presentation is not related to the intent of assessing impact in REF.

Universities tend not to advertise the approach they took for ICS creation, and indeed this may have varied within institutions and submissions for many reasons. It may have made a difference in some, or even many, cases, but not all: the Main Panel A report states that some HEIs used "professional writers to develop and present their impact submissions, while others relied entirely or in part on academic staff" and remarks that "[n]either strategy was uniformly successful" (HEFCE 2015b: 11). Therefore, it is not possible to attribute particular persuasive power to documents written by one group (such as external consultants) over another (such as impact officers, press officers or researchers), and for this reason, such a sub-division is not reflected in my comparisons across sub-samples.

5.2.2 Qualitative assessment of reading experience

While the process of writing an ICS is a black box for those not involved and therefore cannot be systematically researched retrospectively, the texts are now accessible for everyone to read. As part of the Thematic Analysis (Analysis 2 in Table 3, based on Sample B), the research team (two academic and three impact staff, in addition to me, as described in section 4.3.3) tested the role of "reader" with particular regard to formatting, style and specifically the use of adjectives. Coders were asked to record salient examples for these areas, which were then consolidated into themes. Table 10 summarises formatting that the coders considered helpful and unhelpful for interpreting the ICS.

Table 10: Examples of formatting identified from the qualitative analysis of high and low-scoring ICS

Examples of effective formatting	Examples of less effective formatting
Headings	
• Meaningful and used in a consistent way throughout the ICS • Correspond to structure that may be signposted in Section 1 "Summary of the impact" (or at the start of the relevant Section) • One or two levels of subheadings	• Text broken up too much at the expense of a coherent narrative • Titles of research projects or names of researchers as headings can give the impression that these are the focus of the ICS, rather than the impact
Bullet points, lists	
• Details of impact by beneficiary • Highlighting the central research questions of the underpinning research • In Section 2: sub-dividing research findings	• Bullet lists that are not subsequently elaborated • Points in a list seem unrelated • Formatting used to highlight irrelevant details, drawing attention away from claims for reach and significance • Long testimonials as block quotations can give the impression of taking over from the main narrative
Bold or italics	
• Bold is used for impacts, beneficiaries, researcher names, dates, references to Section 3/5 • Italics for testimonial quotes	• Italics are less effective than bold for highlighting impacts/beneficiaries

It is important to note that "effective" and "less effective" are labels applied subjectively by the six coders, and they are not tied *a priori* to high- or low-scoring ICS. From the raw findings as recorded in the shared coding spreadsheet, I could identify that out of the high-scoring ICS, 58% were highlighted as containing effective formatting, while 13% were highlighted as containing less effective or even confusing formatting. For low-scoring ICS, the figures were reversed: 47% were marked as containing less effective formatting, compared to 18% that were considered as having made effective use of various formatting features. There were also examples of ICS that did not stand out as either effective or ineffective, and most ICS included some effective and some less effective features.

To corroborate the findings around subheadings from the thematic analysis, the text files from Sample A (which includes all known 4* ICS) were tagged for subheadings and paragraph headings as described in section 4.4.2, which allowed quantification of this typographical feature. High-scoring ICS were more likely to clearly identify individual impacts

via subheadings and paragraph headings, with the difference between high- and low-scoring texts found to be statistically significant ($p<0.0001$, 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 UoAs combined in Main Panel D, a more discursive academic writing style is prevalent in research publications (see e.g. Hyland 2002: 232) using fewer visual or typographical distinctions such as headings, bullet points and lists. The difference in the number of headings used in high- and low-scoring ICS from those disciplines suggests that high-scoring ICS showed greater divergence from disciplinary norms than low-scoring ICS. 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 ICS. In addition, it should be noted that the conversations reported on by Watermeyer and Hedgecoe (2016: 7) were divided in their views about subheadings, which, while generally welcomed, were reported as having the potential to interrupt the reading flow.

The finding in Table 10 (under "Bullet points, lists") that long testimonials were encountered in low-scoring ICS and were considered problematic for the integrity of the narrative, especially if they were separated typographically, is at odds with Gow and Redwood (2020: 94). They claim that "[o]verall, the presence of quotation – or, even, close reference – was a strong indicator of the high-calibre impact being described. The more extensive the quotation, the stronger the sense of impact." However, this is an occasion where their research design of only investigating high-scoring ICS without reference to low-scoring texts produced misleading conclusions, as the findings from our thematic analysis show. A testimonial quotation may be a helpful ingredient for a convincing ICS and therefore warrant inclusion in Gow and Redwood's eight characteristics of top-scoring ICS, but the fact of its inclusion holds no value for claims about significance and reach, as is implied by the first of the two sentences quoted above. The fact that some low-scoring ICS quoted copious amounts of text from testimonials shows that there is certainly no correlation between the amount of quoted material and the score of the ICS in 2014, as implied by the second quoted sentence (from Gow and Redwood 2020: 94).

Table 11 presents examples of stylistic features that the research team considered as potentially having affected the clarity of the claims, based on the Thematic Analysis of the reader experience on Sample B.

Table 11: Examples of stylistic features identified through the qualitative analysis

Feature	Stylistic features that were seen as helpful	Stylistic features that were seen as unhelpful or problematic for the reader
Clarity of writing	• Claims are made directly • Avoids long, complex sentences and breaks text into paragraphs, sub-sections and lists where relevant	• Long sentences, unnecessarily complex language, or hard to follow even if technical vocabulary is not used • Text not broken up visually and may be poorly organised or signposted • Long-winded descriptions, poor explanations • Spelling mistakes and grammatical errors
Use of technical terms and acronyms	• Avoids jargon such as "-isms" and "lenses" • Explains necessary technical terms and context • Spells out (sparingly used) acronyms	• 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 from research to impact	• No coherent narrative linking research to pathways and impacts or linking different pathways and impacts together • Swapping between first and third person

One specific question in the close reading of ICS was the way in which adjectives were used, because they are an easily understood linguistic means to show evaluation. The aim was to ascertain to what extent adjectives were used effectively, and whether and how they were used in a way that may have jeopardised the positive evaluation that they invited the reader to make. As this was part of a broader thematic analysis, findings were recorded at text level, not at the level of each instance, and percentages refer to the number of texts in which issues were found, not the number of issues found in each sub-corpus. In 38% of low-scoring ICS there was potentially misleading use of adjectives to describe impacts, compared to 20% of high-scoring ICS. Such use of adjectives (including over-use of superlative, unsubstantiated and/or vague adjectives, such as "transformational") may have given an impression of over-claiming or created a less factual impression than ICS that used adjectives more sparingly and precisely to describe impacts. Further examples from across the corpus are given in Table 12. By contrast, sometimes adjectives were mostly reserved for impacts described in testimonial quotes, giving effective third-party endorsement to the claims rather than using these adjectives directly in the editorial parts of the ICS text. This technique is also rated as successful by Gow and Redwood (2020: 90-91).

Table 12: 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 ICS

Problematic 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 ICS and were often not substantiated with further text or evidence
Vague use of adjectives weakening or casting doubt on claims	• Claims of impact on “many” without an explanation of what “many” signifies in the specific context • "Substantial" is used to describe an 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" without further specification implies that impact was incremental or is only emerging slowly • “Very well received and some very valuable feedback" without providing examples casts doubt on the claim

These findings from the qualitative thematic analysis provide more specific information on the nature of presentational differences and language use than the vague description of “journalistic” or “academic” language by previous writers (e.g. McKenna 2021).

5.2.3 Quantifying readability

So far, I have described the writing process and reading experience for ICS based on literature, my own experience and qualitative research. Given the centrality of claims about reading ease in the discussions of ICS (e.g. Watermeyer and Chubb 2018: 2), it is also appropriate to use quantitative data to explore differences in readability between high- and low-scoring ICS.

In the qualitative thematic analysis, 73% of high-scoring and 53% of low-scoring ICS were considered “easy to read” by the team of researchers, who were not typically specialists in the subject area. However, this finding is based on subjective interpretations of the material, and it could also be questioned because the researchers knew whether they were reading a high-scoring or a low-scoring text, which may have influenced their expectation of readability. Quantitative methods were therefore employed to assess the readability of the ICS in order to balance this. Headline findings were summarised in Reichard *et al.* (2020: 9-13), and details are provided in this section of the thesis.

In order to critically evaluate the claim arising from the qualitative study, the main questions are:

1. Is there a difference in the readability of high- and low-scoring ICS?
2. Which group is easier to read overall, if any?
3. If so, which textual characteristics appear to account for this?

In seeking answers to these questions, I set out to narrow down what language features are more common in successful ICS, and therefore can be interpreted as having potentially contributed to those ICS being more successful in presenting their impact in a convincing way. I was also interested in determining whether ICS that were rated as easier to read were more likely to describe impacts that received higher scores.

It is challenging to measure the difficulty, grade level or readability of a text, and most common measures focus on sentence length or word length (McNamara *et al.* 2014: 78). For texts written by academics for academics, this one-dimensional approach poses several problems:

1. Common measures are more sensitive at lower reading levels that are more relevant for school-age readers (McNamara *et al.* 2014: 79).
2. They do not show where the ease or difficulty stems from and therefore cannot point towards improvement, beyond the assumption that shorter words and sentences are inherently easier to process (Crossley *et al.* 2014: 82).
3. They do not take into account the fact that academic language can be highly specialised and therefore that certain long words (technical terms) are less of a barrier to readability for the audience than length-based readability measures might suggest (e.g. Biber and Gray 2016: 246).

In order to quantify the readability of the texts, they were therefore analysed using the freely available Coh-Metrix online tool developed by McNamara *et al.* (2014). Version 3.0 of this tool, which was the most up-to-date version available at the time of analysis in 2018, provides 106 descriptive indices of language features, which can be divided into three main categories. First, 87 indices describe the text in the framework of a theoretical construct such as the commonly used Type-Token Ratio. Second, eight of the indices that the tool reports are principal component scores derived from combinations of those descriptive indices (Graesser *et al.* 2011), reported both as z-scores and as percentiles (therefore accounting for 16 of the output indices). Third, the tool includes three other existing measures of readability, namely Flesch Reading Ease, Flesch-Kincaid Grade Level and Coh-

Metrix L2 Readability. The Coh-Metrix team developed the principal component (PC) scores from 54 of the textual indices using a 37,520-text corpus to assess the "easability" of a text, and the eight scores that were subsequently incorporated into the tool accounted for 67.3% of the variance in the training texts (McNamara *et al.* 2014). Seven of these eight scores are used in the analysis presented in this section as comprehensive measures of "reading ease" because they assess multiple characteristics of the text, up to whole-text discourse level in order to measure textual cohesion (McNamara *et al.* 2014: 78). The final PC score, "Verb cohesion", is most relevant at lower reading levels (McNamara *et al.* 2012) and was therefore deemed not relevant for the register of ICS. In this study, the seven relevant principal component scores (starting with PC in Table 13) are supplemented by the traditional and more widespread Flesch Reading Ease score of readability measuring the lengths of words and sentences, which are highly correlated with reading speed (Haberlandt and Graesser 1985: 366). Table 13 summarises the measures used and provides both a short description of each and a prediction of how each might show in ICS, based on the literature discussed in section 2.2. Further details of the analysis are provided after the table.

Table 13: Overview of Coh-Metrix measures used in this study

Coh-Metrix Measure label	Name	Short description	Expected finding	Shapiro-Wilk for each sub-corpus (*p*)	
				High-scoring ICS	Low-scoring ICS
RDFRE	Flesch Reading Ease	Based on the length of words and sentences	Expected to be low and showing no difference between high- and low-scoring ICS	0.38	0.15
PCNARz	Narrativity	Connected to word familiarity and world knowledge	Low for all ICS as they are not meant to be written about everyday topics	0.70	0.35
PCSYNz	Syntactic simplicity	Sentence length and use of complex vs simple syntactic structures	Medium for all ICS	0.20	0.005
PCCNCz	Word concreteness	Concrete or abstract words and whether they easily invoke a mental image	No difference expected; Target audience is likely to know abstract or low-frequency words used	0.23	0.17
PCREFz	Referential cohesion	Overlap of ideas between sentences and paragraphs / word overlap	Medium but not too high because new ideas need to be introduced	0.27	0.07
PCDCz	Deep cohesion	Explicit sentence linkers for causal and logical relations	High overall and higher in high-scoring ICS	0.18	0.51
PCCONNz	Connectivity	Explicit links for additive, adversative and comparative relationships (and, but, more than).	In an ICS, this should be high to make relations clear. Expected higher in high-scoring ICS.	0.25	0.11
PCTEMPz	Temporality	Are there different tenses/aspects in the text? Expected to be low in academic texts (esp. for perfect aspect)	Perfect aspect more impact-descriptive than simple past; could be higher for ICS that describe impact/processes than for other academic texts	0.61	0.001

Coh-Metrix uses the following basic definitions to analyse a text (McNamara *et al.* 2014: 61-62):

- Paragraph: marked by a hard return
- Sentence: defined by OpenNLP Sentence Splitter (ApacheOpenNLP Development Community 2011): "The OpenNLP Sentence Detector can detect that a punctuation character marks the end of a sentence or not. In this sense a sentence is defined as the longest white space trimmed character sequence between two punctuation marks. The first and last sentence make an exception to this rule. The first non whitespace character is assumed to be the beginning of a sentence, and the last non whitespace character is assumed to be a sentence end."
- Word: derived from a tree-tagged version of the input texts and defined as leaves on the syntactic tree (as opposed to being defined by a blank space on either side of a continuous string of letters).

The PC scores are reported in Coh-Metrix as percentiles (0-100, a higher score indicates that a text is easier to read) and as z-scores (indicating the distance of each data point from a mean that is set at 0 and where therefore positive and negative scores exist, McNamara *et al.* 2014: 84). The tool provides a score for each of the 106 indices for each text, and a researcher can then calculate the averages for each index across texts, as well as comparing the output scores using inferential statistics. Because a normal distribution of language features across texts in a corpus cannot be assumed (such textual data are often positively skewed, Brezina 2018: 8), a Shapiro-Wilk test (Desagulier 2017: 174; Levshina 2015: 56) was applied to the output scores for each of the chosen scores in each sub-corpus (i.e., 4* and 1*/2* ICS respectively). If this test returns a p-value of <0.05, this indicates that there is a non-normal distribution (shaded grey in Table 13). As shown in Table 13, the majority of measures have a normal distribution in the ICS corpus. Moreover, tests such as the *t*-test are robust even for non-normally distributed data (Brezina 2018: 13). Therefore, it would have been too conservative to apply only non-parametric tests, with a high risk of false negatives (referred to as Type 2 errors in statistics). In order to safeguard against false positives (Type 1 errors) generated through the use of parametric tests on non-normally distributed data, the threshold for statistically significant differences was lowered to $p<0.01$ and the variance of the data was not assumed to be equal. The test that was used to evaluate statistical significance was therefore a 2-tailed Welch 2-sample independent *t*-test, following Levshina

(2015: 96). The effect size was measured using Cohen's D, following Brezina (2018: 190), where D>0.3 indicates a small, D>0.5 a medium, and D>0.8 a high effect size. Comparisons were made between high- and low-scoring ICS in each of Main Panels A, C and D,[12] as well as between all high- and all low-scoring ICS across Main Panels.

The remainder of this section describes the results of these tests, first for Flesch Reading Ease (Table 14) and then the Principal Component Scores. These are summarised in Table 15, and the two scores where a significant difference could be found are discussed further.

Table 14 shows scores for Flesch Reading Ease out of 100, with a higher score indicating that a text is easier to read. While the scores reveal a significant difference between the 4* and 1*/2* ICS, they also indicate that ICS are generally on the verge of "graduate" difficulty, where scores of 30 and lower are interpreted as the texts being suitable for postgraduate readers and scores between 30 and 50 interpreted as being at the optimum level of readability for undergraduate students (Hartley 2016: 1524). As such, the analysis should not be understood as suggesting that these technical documents should be adjusted to the readability of a newspaper article as implied by the findings in McKenna that "some universities […] had their communications or press office edit [ICS] for readability" (2021: 18), but they should be maintained at interested and educated non-specialist level.

Table 14: Average Flesch Reading Ease scores for 4 and 1*/2* ICS by Main Panel, with measures of statistical significance*

	Flesch Reading Ease (out of 100)		**Statistical significance *p* (*t*-test)**	**Effect size (Cohen's D)**
	4* ICS	**1*/2* ICS**		
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

Interestingly, there were differences between the Main Panels. In the Social Sciences and Humanities (Main Panels C and D), the *t*-test results shown in Table 14 indicate that 4* ICS scored significantly higher on reading ease than 1*/2* texts. There was no significant difference between 4* and 1*/2* ICS in Main Panel A (Life Sciences). However, Main Panel A ICS showed, on average, lower reading ease scores than those in Main Panel C and D. This means that their authors used longer words and sentences, which may be explained in part

[12] As explained in section 4.3.2, the text base for Main Panel B was too small to apply statistical tests to a sub-corpus of Main Panel B texts alone.

by more and longer technical terms needed in Main Panel A disciplines even in the ICS register. The difference between 4* and 1*/2* ICS in Main Panels C and D may be explained by the use of more technical jargon in 1*/2* ICS, but indicates that the writers of 4* ICS in these disciplines were able to work around using such language. This interpretation is consistent with findings from the qualitative analysis, as described in section 5.2.2 above (especially Table 11).

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- and word-length are highly correlated with reading speed (Haberlandt and Graesser 1985: 366), Hartley (2016: 1524-1525) is right in his criticism that this index 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 principal component scores introduced above, most did not reveal a significant difference between 4* and 1*/2* ICS or between different Main Panels. Moreover, in most of the indices, ICS across sub-corpora had similar average scores compared to the baseline of general English texts (Graesser *et al.* 2011: 228).[13]

Table 15 shows percentile scores for means because they illustrate more intuitively where the texts sit on a readability measure, with a score of zero meaning 'hard to read' and a score of 100 'easy to read' based on each principal component score. To compare means of these scores with inferential statistics, McNamara *et al.* (2014: 86) recommend using z-scores, rather than the end-user-facing percentiles. They define a z-score as indicating "how many standard deviations an observation […] is above or below the mean, where the mean is set at 0" (McNamara *et al.* 2014: 84).

[13] The reference basis on which Coh-Metrix scores are calibrated within the tool is the TASA corpus, which contains 37,520 texts with an average length of 288.6 words. The texts are "representative of the texts that a typical senior in high school would have encountered from kindergarten through 12th grade" (Graesser *et al.* 2011: 228).

Table 15: Overview of the means and standard deviations of principal component scores (as percentiles)

Index	**Description**	**4* ICS**		**1*/2* ICS**		**Statistical significance *p* (*t*-test for *z*-score)**	**Effect size (Cohen's D)**
		Mean	SD	Mean	SD		
PCNARp	Narrativity	8.48	4.01	8.02	4.25	0.16	
PCSYNp	Syntactic simplicity	41.64	13.12	39.36	14.06	0.09	
PCCNCp	Word concreteness	65.63	16.62	66.90	16.40	0.61	
PCREFp	Referential cohesion	26.29	17.03	30.46	16.77	0.97	
PCDCp	Deep cohesion	53.53	18.53	43.85	17.52	<0.001	>0.5
PCCONNp	Connectivity	9.88	8.76	5.94	8.35	<0.001	>0.5
PCTEMPp	Temporality	39.78	20.36	41.09	21.08	0.5	

There are statistically significant differences between 4* and 1*/2* ICS in two of the measures, highlighted grey in Table 15: *deep cohesion* and *connectivity*.

Deep cohesion indicates whether a text makes causal connections between ideas explicit (e.g. through subordinating conjunctions such as "because", "so") or leaves them for the reader to infer. As shown in Figure 8, 4* ICS had a higher level of deep cohesion compared to general English texts, while 1*/2* ICS tended to sit below the average of the reference texts (see footnote 13 above).

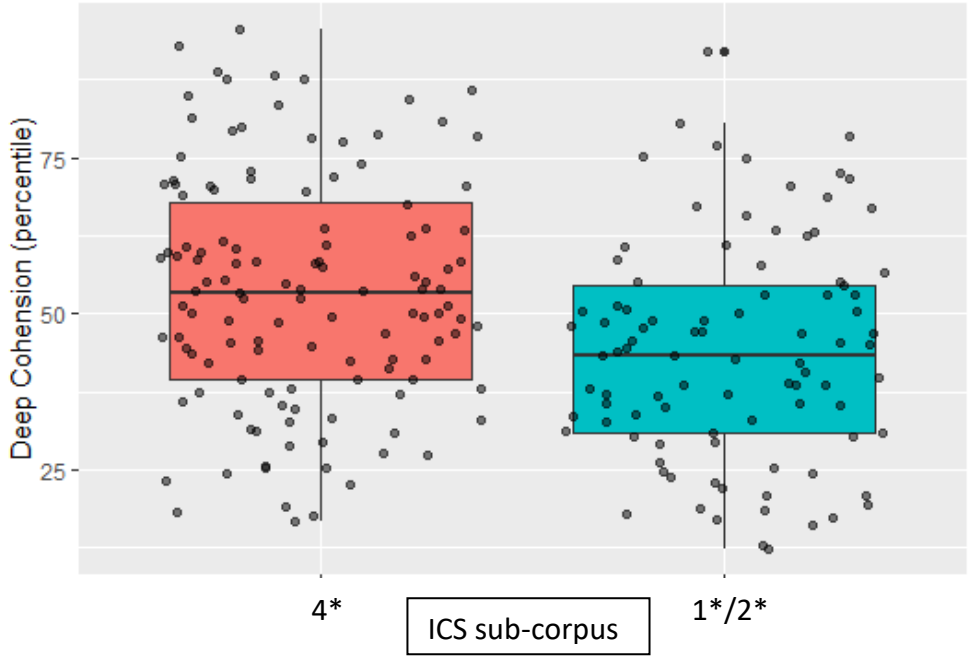


Figure 8: Comparison of Deep Cohesion across 4 and 1*/2* ICS*

Connectivity measures the level of explicit logical connectives (e.g. coordinating conjunctions such as “and”, “or” and “but”) to show relations in the text. ICS were low in connectivity compared to general English texts, as shown in Figure 9: while the average of 1*/2* ICS is near the bottom (6th percentile) with a large number of texts approaching 0, the average of 4* ICS at least approaches the 10th percentile (exact values for averages in Table 15).

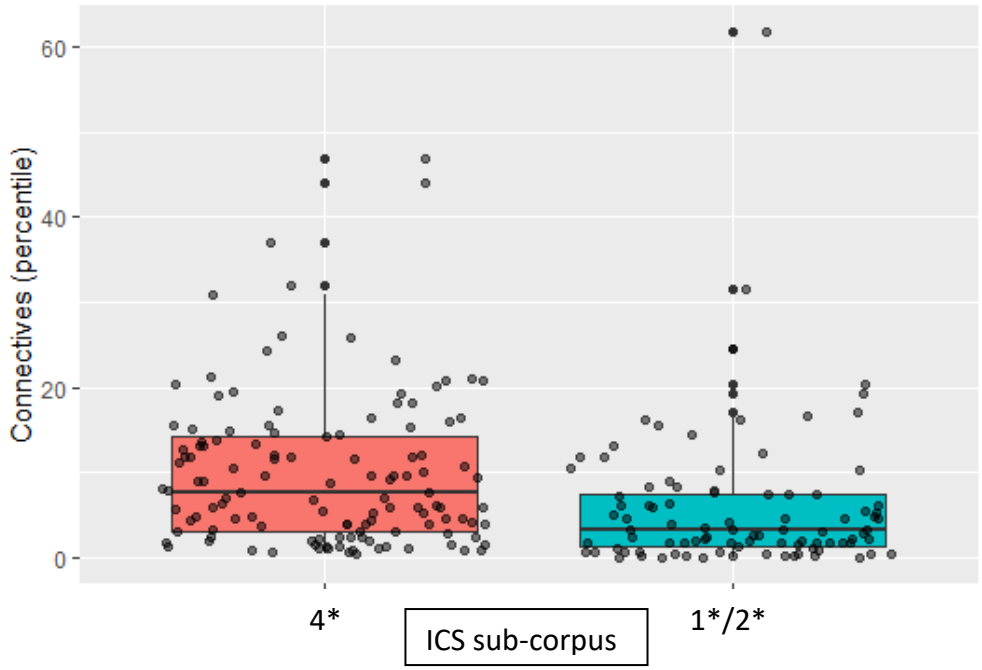


Figure 9: Comparison of Connectivity across 4 and 1*/2* ICS*

Although in terms of the average scores reported in Table 15 the *t*-test indicates significant difference at $p<0.001$, with a medium effect size ($D>0.5$), the box plots in Figure 8 and Figure 9 show a great range and variation in scores within both the 4*- and 1*/2*-corpora. The box plots overlap to the extent that the mean score of each plot is within the box of the comparison plot. These results should therefore be viewed with caution: despite their statistical significance, it appears that the variation within each sub-corpus does not allow firm conclusions about these features being a marker of one or the other group of texts.

For the principal component scores where statistically significant differences could be shown, comparisons between scores within each Main Panel were made, as shown in Table 16.

Table 16: Deep Cohesion and Connectivity – Mean percentiles for 4 and 1*/2* ICS and significance and effect size of difference, by Main Panel (MP*

	Deep cohesion				Connectivity			
	4* ICS	1*/2* ICS	*p* (*t*-test)	Cohen's D	4* ICS	1*/2* ICS	*p* (*t*-test)	Cohen's D
Overall	53.53	43.85	<0.001	>0.5	9.88	5.94	<0.001	>0.5
MP A	53.51	51.72	>0.5	-	12.08	9.93	<0.05	>0.5
MP C	55.98	45.06	<0.01	>0.5	9.16	4.81	<0.001	>0.8
MP D	49.35	38.65	<0.01	>0.5	7.13	5.62	>0.5	-

It is important here to point out that ICS in Main Panel A (Life Sciences) on average used longer words and sentences than those in Main Panels C (Social Sciences) and D (Arts and Humanities), as indicated by the *Flesch Reading Ease* scores (Table 14), and therefore would be seen as harder to read by that measure. However, they also made causal and logical relationships more explicit in the texts, as shown by the relatively higher scores in *Deep Cohesion* and *Connectivity* especially in the 1*/2* sub-corpora (Table 16). This finding would indicate that on these measures, they are easier to read than ICS from other Main Panels. Therefore, no firm conclusions can be drawn regarding the reading ease of texts from any one measure.

Moreover, the finding that ICS in science disciplines are more explicit in expressing causal and logical relationships than those in social sciences and humanities is particularly interesting in light of Biber and Gray's (2016: 121-122) suggestion that science writing is maximally *in*explicit compared to Humanities prose (see section 3.1.3 above). One possible explanation for this discrepancy is that Coh-Metrix includes different measures to Biber and Gray's discussion of explicitness and compressed versus elaborate syntactical structures. This is perhaps unlikely because the reason that Biber and Gray give for the structural compression in science writing is the ever-tighter word limits in journal articles (2016: 42), which is exacerbated through the 4-page limit of ICS. Another possible explanation might be that ICS are indeed a different register to research articles, as discussed in the situational analysis (section 5.1 above), with a different purpose and readership requiring a more explicit style for semi-specialist readers compared to the peer readers of research articles. If this is the case, it is even more interesting that Main Panel A texts scored more highly on these measures because it indicates that the writers of these ICS may have made a larger adaptation to their writing style for this register, both because the texts are on average

grammatically more explicit than those in other disciplines and because the presumed starting point, as implied by Biber and Gray, was even further removed from this level of grammatical explicitness than those of ICS writers in Main Panels C and D.

Support for this explanation could also be found in the related finding that in Main Panels C and D, generally (in 3 out of 4 comparisons in Table 16) there was a clearer difference between 4* and 1*/2* ICS than in Main Panel A, with 4* ICS receiving scores that show greater reading ease according to these measures. This could indicate that writers of 4* ICS in Main Panels C and D, as well as writers of ICS in Main Panel A generally, had a greater level of adapting to this register than writers of 1*/2* ICS in Main Panels C and D. One reason for this could be that a higher degree of guidance, writing and editing by professionals other than the researchers themselves, who are embedded in writing for their disciplinary peers, may have been provided in universities that returned submissions that received top scores in REF, and in those where life sciences were returned. While this is largely speculation, it is compatible with anecdotal evidence and my own experience of working with the respective disciplines.

Overall, the average scores in all measures introduced in Table 13 are fairly similar compared with the reference texts in the Coh-Metrix tool (see footnote 13 above). Given the specificity of the register, this is not surprising. When trying to pinpoint the features that influence readability, the use of connectives (both causal, i.e. Deep Cohesion, and logical, i.e. Connectivity) seems to make the most difference, and both are used more in high-scoring ICS. However, as Figure 8 and Figure 9 show, this finding was not consistent enough within the body of high- and low-scoring ICS respectively to be translated into the conclusion that more connectives in an ICS might have a direct influence on the REF score. Nonetheless, the data do provide some tentative support for previous analyses of ICS that emphasised the narrative cohesion in successful examples (Pidd and Broadbent 2015) and the explicit links formed between research as cause and impact as effect (Reed *et al.* 2019).

A final observation from this quantitative analysis of readability is that this is an analysis where I made comparisons across a defined range of possible differences. The Principal Component Scores were chosen as aggregate scores for the indices, and one was discarded as irrelevant before the statistical analyses. This means that, rather than looking for possible differences and reporting those, I defined several possible differences, namely the seven remaining PC scores. Of these, actual differences can be found in only two, showing that out

of all possible differences related to readability in this comprehensive tool, most do not reach statistical significance. Moreover, I have argued that even those two areas that do demonstrate differences should be viewed with caution. Therefore, the difference between 4* and 1*/2* ICS can be seen as small compared to the possible type and degree of difference that might have been found.

5.3 The content of impact case studies

Following the analysis of the overall situation (section 5.1), the discussion of writers (section 5.2.1) and the reading experience (sections 5.2.2 to 5.2.3), the remainder of this chapter describes findings related to the content of ICS. Many studies have set out to find common themes across ICS (see above section 2.1.4), but this is not the goal here. Rather, this section explores different kinds of content that could in theory appear in ICS regardless of subject matter, with a specific focus on the relative emphasis of research-, impact- and pathway-focused material in ICS Section 1 (thesis section 5.3.1) and the use of evidence to convince the reader of the significance and reach of the claimed impact throughout the whole ICS (section 5.3.2). It is important to consider these questions alongside analyses that focus more strongly on the lexis because it is the content, that is, the significance and reach, of an ICS that is subject to assessments. If differences can be found here, in the material that is being included, this may be more directly related to the ICS scores than word choices are.

5.3.1 Type of material in Section 1 of impact case studies

The first analysis of the content of ICS is concerned with the type of material that can be found in high- and low-scoring ICS respectively, regardless of the Unit of Assessment or more specifically the subject matter of a given ICS. It is based on Sample C, that is, extracts of 76 ICS (namely Section 1 texts), balanced across scores and Main Panels, with Main Panels A and B combined into sub-corpus AB (described in section 4.3.4).[14] Section 1 ("Summary of the impact") had an indicative word limit of 100 and the following guidance in the 2014 REF: "This section should briefly state what specific impact is being described in the case study" (HEFCE 2011: 52). Despite being the shortest, this summary section has special relevance because it sets an expectation for the reader (McKenna 2021: 24). The remainder of the ICS should then follow the focus that is set in that section. Indeed, Kousha and Thelwall (2025)

[14] Note that this is different from Sample A which was described in section 4.3.2 and used in the analyses in section 5.2.3 Readability, where Main Panel B texts were disregarded, rather than integrated in a corpus of science writing.

find that Section 1 is most predictive of the score when sections of ICS are submitted to ChatGPT together with the guidance and a query to assign a score to the ICS based on the guidance.

A first impression is also given of the nature of the impact and its link to the university. The analysis presented now therefore examines the extent to which Sections 1 in the 2014 REF included material related to the following: research, impact, pathway, problem, other.

This analysis was originally conducted in support of the Appraisal analysis described in chapter 7. In her description for her own analysis of EU policy documents on immigration, Tupala highlights that one major component of such an analysis should be "determining who or what is being evaluated" (2019: 13), that is, the target of the evaluation. In her case, this is, for example, the EU (self-evaluation in EU policy documents) versus migrants (other-evaluation in those same documents). In my analysis, it is the evaluation of the various different kinds of content. Given the purpose of an ICS, it could be expected that the Summary will include, as a minimum, some clear reference to the impact that it is putting forward for assessment; the research that underpins the impact; and the pathway connecting the impact to the research. Section 1 may also include a statement of the problem(s) which the research or the impact address. Other material is unlikely to serve the purpose of making a strong argument for research-connected impact solving a problem, and therefore, given the tight word limit, it can be expected that such material is used sparingly, especially in high-scoring ICS.

The texts in Sample C were analysed using the UAM Corpus Tool (details in section 7.1). This allows for several layers of coding, one of which was created to code for the type of material as described in the previous paragraph. Given its purpose of establishing the target of a given instance of evaluative language, this layer was titled "Target" and includes the categories "Research", "Impact", "Pathway", a "Problem" statement or something else ("Other"), as represented in Figure 10:

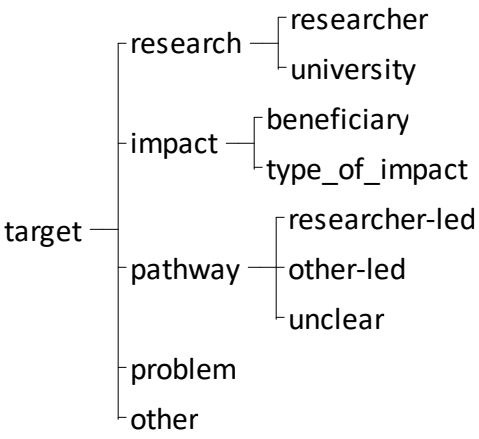


Figure 10: Coding scheme for the "Target" of each text segment

Table 17 shows examples of each type of content from various ICS, along with the number of times this tag was applied ("Tags") and the number of words overall that were covered by this tag ("Words") in the whole of Sample C, that is, the 76 Section 1 texts from selected high- and low-scoring ICS (9,238 words overall).

Table 17: Overview of "Target" categories in Sample C - examples and distribution

Content type	**Example**	**Tags**	**Words**
Research-Researcher	His post-2008 work builds on his pioneering research and the Afghanistan Music Unit, founded in 2002. His scholarship is rooted in research practice, networks, and decades of experience, giving him unique insight into Afghanistan's music and its citizens at home and abroad. (UoA35 Goldsmith *Afghan*)	26	664
Research-University	[…] as a result of UCLan's fire retardant research (UoA13 UCLan *Retardant*)	76	1671
Impact-Beneficiary	Impacts have benefited a range of users (UoA20 Ulster *Amnesty*)	58	479
Impact-type of impact	These impacts are particularly visible in shifts in information society policy at the international level to include greater attention to citizen interests (UoA36 LSE *Citizen*)	161	3278
Pathway-researcher-led	Working closely with police forces, crime prevention practitioners and policy makers, SCS staff have provided evidence, expertise and advice to support particular crime prevention initiatives and approaches to crime prevention more broadly. (UoA22 UCL *Crime*)	98	1449
Pathway-other-led	The research was included in the UK government Advisory Council on the Misuse of Drugs (ACMD, 2009) review of MDMA effects (UoA4 East London *Ecstasy*)	42	782

Content type	Example	Tags	Words
Pathway-unclear	This has had effects on practice in contexts in which national and international EY policy, leadership and pedagogy are developed and produced, enacted and contested. (UoA25 Birmingham City *Early Years Education*)	17	175
Problem	Footrot (FR) causes 90% of lameness in sheep. FR reduces productivity and lowers sheep welfare. (UoA6 Warwick *Footrot*)	21	550
Other	The research has subsequently been used by major national research projects in corruption in local government. (UoA20 Sunderland *Local Integrity*) *This example may have been included to add to the impact claim, but impact on other academic research was ineligible.*	17	190

Details on the process of annotation are provided in the coding manual (see Appendix F), and further details on the piloting process are included in section 7.1.3. The texts were coded in two separate layers, first to identify and classify resources of GRADUATION (findings in section 7.2), and then to segment the texts according to what a stretch of text (e.g. clause or sentence) refers to ("Target"). It is the latter layer that provided the findings in this subsection.

Figure 11 and Figure 12 provide an overview of the distribution of material across the corpus. It can be seen that high-scoring ICS dedicated more words to descriptions or claims of impact than low-scoring ICS (48.33% compared to 30.37%), while low-scoring ICS included more descriptions of research (34.97% compared to just 18.07 in high-scoring ICS).

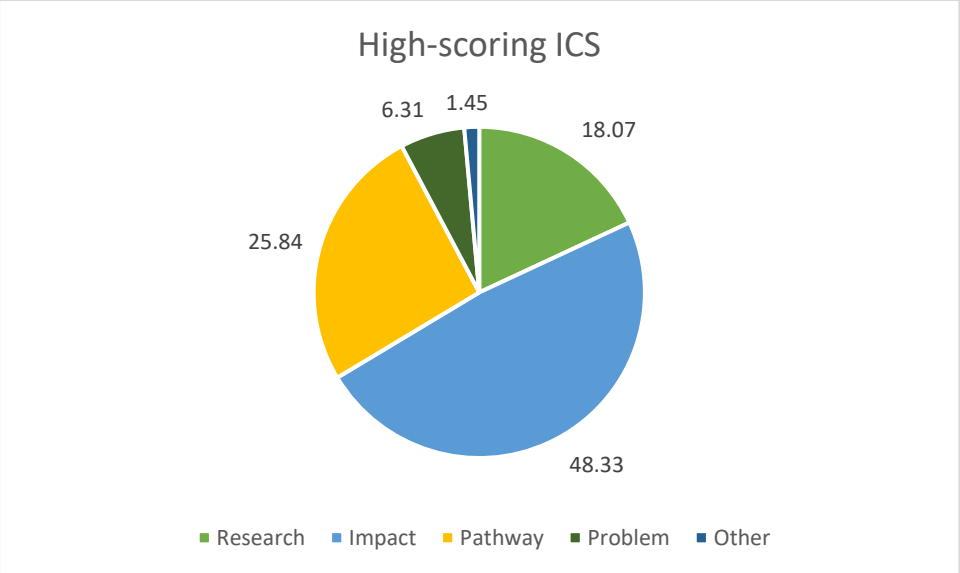


Figure 11: Distribution of material in Section 1 of high-scoring ICS (Sample C)

In fact, as Figure 12 shows, low-scoring ICS include more material on research descriptions (34.97%) than on impact claims (30.37%) in their Sections 1.

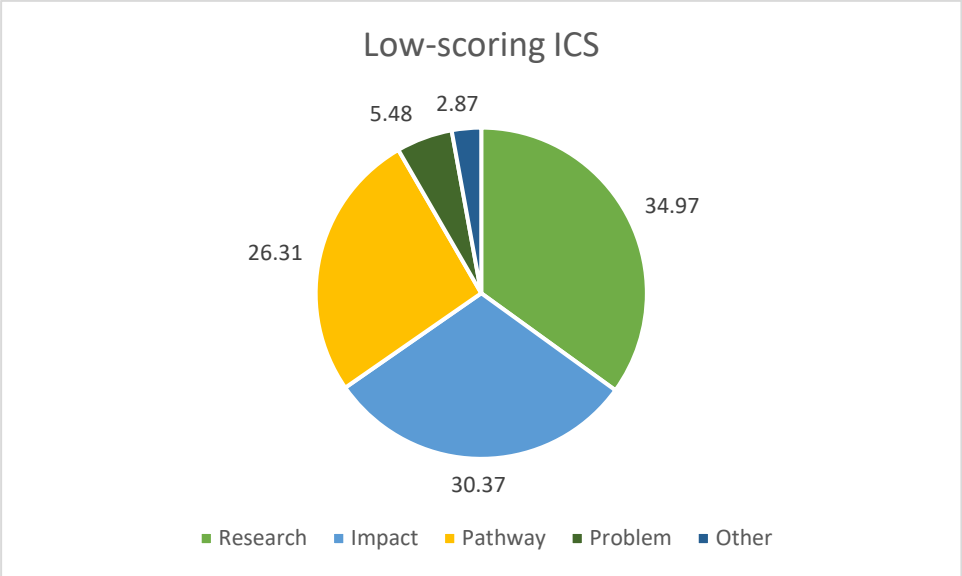


Figure 12: Distribution of material in Section 1 of low-scoring ICS (Sample C)

The remainder of this subsection gives an overview of and commentary on the number of texts that include material related to each of the "Target" categories.

Research is referred to in 72 out of 76 texts (94%). The four texts that do not explicitly include material on the underpinning research in Section 1 are high-scoring Main Panel D (2x) and low-scoring Main Panel AB (2x) texts. Research-related material is split further into information related to a *researcher*, which can be found in 21 texts (27%), and material related to *institutions* (tagged as "university"), which appears in 54 out of the 76 texts (71%). There is an overlap of three texts that have material relating to both a person and an institution. From this, no interpretation can be provided for significant differences. It is interesting, however, that more texts frame their research-related material around the submitting institution, rather than the researcher. This is consistent with the fact that it is institutions, rather than individual researchers, that are assessed in REF.

Impact could be expected to appear in all "Summary of the impact" sections. However, it is referred to in only 69 of 76 texts (90%), which means that seven Section 1 texts do not refer to material classed as *impact* at all, although all of the 4* texts in the sample do. Such material can be further characterised as *type of impact* and *beneficiaries*. *Type of impact* appears in most texts (67 of 76, 88%), but *beneficiary* is specified in only 41 of 76 texts (54%). These are distributed fairly equally among high- and low-scoring ICS (22 high, 19 low).

Conversely, two ICS specify a beneficiary but not the actual nature of the impact. One of these is low-scoring in Main Panel C, but surprisingly the other text that does not specify the nature of the impact is a high-scoring ICS in Main Panel AB (UoA4 Birkbeck *Developmental research*). This text includes clear descriptions of far-reaching pathways and clearly states the beneficiaries, which may have masked the fact that the nature of the impact is not made explicit in the summary text.

Unsurprisingly, none of the seven texts that failed to specify either their impact or their beneficiary in Section 1 achieved a rating higher than 2*. Four texts are in Main Panel C, two in Main Panel D and one in Main Panel AB. Of these, one text only includes material about *Research*. The material in the others is divided mostly into *Research* and *Pathway*, with one including a lengthy *Problem* statement and another including mostly material from the *Other* category, both to the detriment of actual impact claims.

Pathway is referred to in 71 out of 76 texts (93%). The five texts that specify no pathway are distributed across the six sub-corpora, with the exception of Main Panel AB (high-scoring) where all ICS include pathway material. Pathway material can be further specified according to who led the pathway: the *researcher*(s) or *others*. 54 ICS (71%) describe pathways led by *researchers*, spread fairly evenly across sub-corpora but with fewer texts in Main Panel C (low-scoring) that do this. This is consistent with the finding by Gow and Redwood (2020: 83-85) that high-scoring ICS in general included pathways to impact with heavy involvement of researchers, but because their study does not include a control group of low-scoring texts, they do not comment on any difference between scoring brackets. 30 ICS (39%) describe pathways led by *others*, also spread fairly evenly across sub-corpora, and 16 texts include both researcher-led and other-led pathways.

Problem statements are included in 20 Section 1 texts (26%). Of these, 12 are high-scoring texts, which are split fairly evenly across Main Panels, and eight are from low-scoring ICS. It seems therefore that including a problem statement is not a specific feature of either high- or low-scoring ICS.

Other: There are 15 out of 76 texts (19%) that include material not attributable to any of the other categories. Any remaining material was tagged as *Other* because all words in each text had to be annotated with exactly one tag in order to create sub-corpora for use in chapter 7. This includes discourse markers such as “this case study describes”, which itself was treated

as not describing impact. Similar to the *Problem* category, the inclusion of *Other* material does not seem to be a distinctive feature of a particular sub-corpus.

5.3.2 Thematic analysis of full impact case studies

Categorising the main types of material is informative, but it cannot tell the whole story. The complementary thematic analysis (full details in Reichard *et al.* 2020) provides additional insight into the differences in content between high- and low-scoring ICS, specifically in the following areas: the specificity of claims; providing convincing evidence; and articulating links between research and impact. All of these elements are important for persuading the assessor-reader that significant and far-reaching impact arose from the research activity of a university. This section reports on those differences between high- and low-scoring ICS where there may have been scope for a more convincing presentation, while acknowledging that in many cases the lack of evidence or specificity included in an ICS reflects an actually missing link, or limited reach or significance. That is, it may sometimes, but not always, have been possible to make a more persuasive impact claim with more resources to collect and present evidence, but sometimes the presentation of low-scoring ICS is commensurate with the modest claims it could have made in the first place.

One finding from the qualitative thematic analysis was that 84% of high-scoring ICS articulated benefits to specific groups and provided evidence of their significance and reach. This was found for only 32% of low-scoring ICS, which typically focused instead on the pathway to impact without pinpointing and evidencing the benefits. For example, they may describe the dissemination of research findings and engagement with stakeholders and publics without elaborating on the benefits arising from this dissemination or engagement. One way of conceptualizing this difference is using the content/process distinction: whereas low-scoring ICS put relatively more focus on the *process* through which impact was sought (i.e. the pathway used), the high-scoring ICS 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 ICS from all panels, but were less often claimed or evidenced in low-scoring ICS. Where reach was more limited geographically, many high-scoring ICS used context to create robust arguments that their reach was impressive in that context, describing reach for example in social or cultural terms or articulating the importance of reaching a narrow but hard-to-reach or marginalised target group.

Table 18 provides examples of use of evidence that were used in the ICS text to show significance and reach of impacts, and that were considered effective or ineffective respectively by the research team (as explained in section 4.3.3). Details of the ICS from which examples are drawn can be found in Appendix B.

Table 18: Examples of the types of evidence used to show significance and reach of impacts from research

Significance: examples of effective use	**Significance: examples of ineffective use**
• 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 (UoA6 Warwick *Biopesticides*), or peer-reviewed analysis showing that the BBC changed its coverage (UoA36 Cardiff *Newscoverage*) • 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) • 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 or 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 (UoA3 Nottingham *Endoflife*) • 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 data or comparison group to demonstrate that changes could be attributed to actions based on the research
Reach: examples of effective use	**Reach: examples of ineffective use**
• Addressing a challenge that was uniquely felt by a *particular* group, even if on a sub-national scale • Successfully helping 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" (UoA29 Newcastle *Poetry*)	• 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 are based on the global reach of an organisation or initiative using the output of research, without specifying the impact that the research activity or output has had

Table 19 shows how corroborating sources were used in Section 5 of the ICS template to support impact claims. 82% of high-scoring ICS, compared to 7% of low-scoring ICS, were identified as having generally high-quality corroborating evidence. Conversely, 11% of high-scoring ICS, compared to 71% of low-scoring ICS, were identified as having corroborating evidence that was vague and/or poorly linked to claimed impacts.

Table 19: Examples of the use of corroborating evidence identified through qualitative analysis

Examples of high-quality corroborating evidence	**Examples of problematic corroborating evidence**
Credibility of sources	
• Testimonials from high-level decision makers 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 ICS authors, e.g. showing impact on computing speed or randomised control trials	Potential conflicts of interest undermine the credibility of a source which is therefore less persuasive, for example: • 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 a new industry standard, explaining how and where research is cited; evidence of audience or visitor numbers • Link to Government press release showing that 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	• Download figures and other statistics relating to reach of pathway rather than reach of impact • A funding proposal (e.g. original Knowledge Transfer Partnership application) without evidence of how this subsequently contributed to impact • Collaboration agreements for future work • Links to project websites and Facebook pages • Lists of media coverage without explaining what impact they evidence • Links 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

Examples of high-quality corroborating evidence	Examples of problematic corroborating evidence
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 • 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 • No evidence provided to support key claims, e.g. missing economic data or testimonials to corroborate economic impact

Only 50% of low-scoring ICS clearly linked the underpinning research to claimed impacts (compared to 97% of high-scoring ICS). This gave the impression of over-claimed impacts in some low-scoring submissions. For example, one ICS claimed "significant impacts on Greek society" based on enhancing the security of a new IT system in the department responsible for publishing and archiving legislation (UoA11 University of East London *Securesoftware*). 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 (UoA11 University of the West of Scotland, *Enablingtech*). One ICS 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 ICS that appeared to conflate correlation with causation in their attempts to imply attribution between research and impact. Such a lack of explicit linking between research and impact also undermines the persuasive power of an ICS. It is possible that an over-focus on articulating claims to the detriment of evidence may have contributed to an impression of overclaiming, which highlights the crucial role that evidence plays in presenting ICS compared to editorial language choices. That said, it may often be the case that evidence of impact, of the link to the underpinning research, or both is near-impossible to obtain or to separate out from other influences, and therefore those preparing ICS may not have been able to provide the evidence that would have been required in order to establish a convincing link.

5.4 Conclusion

As summarised in chapter 2, several commentators suggest that stylistic features of impact case study writing have a material impact on the score assigned to these ICS. For example, according to Manville *et al*., assessors noted that they "were aware that presentation affected their assessment of the impact" (2015a: 39). Watermeyer and Hedgecoe advise the use of "a compelling but unfussy style" (2016: 6), and McKenna argues that "the trick is to make it easy for the assessors" (2021: 18).

In response to research question 1a (introduced in section 1.2), the studies in this chapter (overview in Table 3 in section 4.2.2 above) suggest a perceived correspondence between reading ease and higher scores, but both the qualitative and quantitative studies suggest that this relationship is weak and there is no evidence that any correlation is causal.

The thematic analysis described in section 5.2.2 shows that a higher proportion of high- than low-scoring ICS used more effective visual and typographic structures (see Table 10). It also suggests that a higher percentage of low-scoring than high-scoring ICS used language that may have given an impression of over-claiming. The thematic analysis also finds a higher percentage of high- than low-scoring ICS to be easy to read. However, the researchers engaged in the thematic analysis knew whether they were reading a high-scoring or a low-scoring text, which may have influenced their expectation of readability. In addition, the results reported above are based on examples of effective and less effective language and formatting use, rather than a systematic categorisation.

In order to be able to more reliably quantify the use of language components that make a text easier or less easy to read, additional quantitative analyses were performed, as summarised in section 5.2.3. CohMetrix was used to determine the readability based on a range of readability tests. Out of eight tests, only three showed a statistically significant difference between high- and low-scoring ICS, namely Flesch Reading Ease, Connectivity and Deep Cohesion. Yet, despite their statistical significance, it appears that the mean values of these measures are so close and the variation within each sub-corpus is so wide (as illustrated in the box plots in Figure 8 and Figure 9) that it is not possible to draw firm conclusions about these features being a marker of one group of texts or the other.

Note that even if high-scoring ICS are written in a more reader-friendly or persuasive way, the difference in writing may not have resulted in the higher scores, if assessors were able to

base their assessment on the merits of the impact rather than the writing. The correlation could instead, for example, be due to differences in available resourcing that affect both the impact itself (quality, significance or reach) and the quality of writing and editing ICS.

However, in response to research question 1b, section 5.3 suggests that at least in one respect writing matters: those ICS that (1) were able to better demonstrate the significance and reach of their impact, (2) showed how the impact was enabled by the university's research, and that (3) also provided strong evidence for both, were consistently provided with higher ratings. This is unlikely a coincidence, given that these aspects are the basis for the assessment criteria.

In particular, the analysis of the type of material in ICS presented in section 5.3.1 shows that, although most ICS provide details about the impact or beneficiary in Section 1, those ICS in the sample that do not are from the low-scoring sub-corpus. In addition, as described in the part of the thematic analysis presented in section 5.3.2, the vast majority of high-scoring ICS clearly articulate the impact and beneficiaries, whereas many low-scoring ICS focus on pathways to impact. Finally, high-scoring ICS are generally able to provide clear evidence for significance and reach of impact as well as the link between research and impact claimed, whereas many low-scoring ICS use generic language for describing impact and reach or provide insufficient evidence of the link between research and impact.

In conclusion, the language of ICS does matter. However, this study suggests that what matters are unlikely to be language choices and a mastery of linguistic persuasion, but the author's ability to clearly demonstrate and evidence the impact of their research.

Chapter 6 Lexical Investigation

As explained in section 2.3, the assumption by many commentators that the language of impact case studies influenced their scores can only be substantiated if there is a difference in language between lower- and higher-scoring ICS. Chapter 5 investigated differences in readability of ICS and in the types of content that tended to be emphasized in Section 1 of ICS. The thematic analysis showed that, overall, differences were content-related, concerning significance, reach and the use of evidence. However, there was not much difference in readability, including grammar and cohesion. Chapter 6 now turns to lexical differences and the question of whether any differences here suggest that high-scoring ICS were written in a more persuasive style.

The purpose of this chapter is to report on a bottom-up comparison that firstly quantifies lexical differences, and secondly examines these differences from various perspectives. The methods of analysis are described in the next section (6.1), but in brief, the chapter compares sequences or strings of words (n-grams, defined in section 6.1.1) that are more frequent in one sub-corpus (e.g. high-scoring ICS) relative to the other sub-corpus (e.g. low-scoring ICS) and vice versa. Those word combinations where a statistically significant difference in frequency can be found were categorised according to their predominant meaning in the ICS context (see section 6.1.4 for methods, section 6.2.1 for results). Independent of this categorisation into themes, they were then marked by the level of authorial or editorial freedom:

(1) those where the possibility to use a word combination is pre-determined by content (e.g. *the house of* [lords/commons]);
(2) those where a writer or editor has a free choice for using a certain word or combination (e.g. *in terms of*, where multiple equivalent wordings are available to replace this phrase);
(3) those where such editorial choice is available in only some cases (for a full discussion of this distinction, see section 6.1.5; results are presented in section 6.2.2).

For those entries where there was editorial choice in at least some cases, which applies to 171 of the 245 entries that are found to have statistically significant differences in frequency between sub-corpora, I have indicated whether the word combination can be construed as

having a persuasive function in the ICS context, and in what way. This resulted in a typology of persuasive functions and language specific to ICS (for a full discussion of the functions, see section 6.1.6; results are presented in section 6.2.3).

As in chapter 5, some interpretation of the findings is given in the various results descriptions (section 6.2). In addition, themes cutting across the chapter are discussed in relation to relevant literature before the conclusion, in section 6.3.

6.1 Method

One option for investigating differences in persuasion is to use existing word lists of persuasion (e.g. derived from Hyland 2005a: 220-224) and search the high- and low-scoring sub-corpora of ICS for occurrences of these words. However, the register of ICS is quite specialist, and it is not overtly persuasive, unlike for example a pamphlet, a political speech or advertising. Indeed, Hyland and Jiang note that items with persuasive (what they call "hyperbolic") functions differ across genres and registers (2021: 191). Existing word lists that are derived from genres with a more obviously persuasive purpose may thus not be helpful in detecting differences in subtle persuasion in ICS, especially as the criteria for assessing ICS are such that persuasion may be hidden and only be noticeable to those familiar with the aims of the genre.

A more data-driven[15] 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). This was briefly explored for the present study: frequency-based wordlists from the corpus were compared to wordlists generated from the British National Corpus (BNC Consortium 2007) to identify keywords for ICS, that is, words appearing statistically more often in ICS than in general written English. This comparison was not assessed as useful, though, because the resulting keywords were mostly words that appear in the REF guidance and are therefore expected (e.g. *research, impact, evidence, underpinning*). This does not

[15] This analysis is not completely data-driven, because I set a number of parameters, such as minimum frequency thresholds and choice of statistical methods, as explained in the remainder of this section (6.1). But the findings are determined by the specific, hidden, lexical profile of these texts, rather than by a pre-existing framework. By contrast, the analysis described in chapter 7 uses and adapts an existing framework. As such, of all analyses in this thesis, those described in this chapter are by comparison the most exploratory and data-driven.

allow conclusions about effective word choice that may have influenced scores, therefore this avenue of investigation was not pursued further.

To mitigate the unhelpful results of the keywords approach, I turned to phraseology, defined broadly as the study of "words occurring together" (Hunston 2011: 5). In addition to units of meaning or formulaic language such as collocations (e.g. Gablasova *et al.* 2017: 156), this includes multi-word units such as n-grams or lexical bundles, which may be co-occurring without creating specific meaning (Durrant 2017: 166, see also 170 for the difference between n-grams and lexical bundles). An n-gram is defined as a sequence of items of length n; that is, in relation to lexical items, a 2-gram is any sequence of two words, a 3-gram is three consecutive words, and so on.

Hunston (2011: 7) lists six features originally identified by Gries (2008, quoted verbatim from Hunston, emphasis in original) along which "phraseological studies" can be defined. The following list explains how the present study applies these features:

1. "The *nature* of the elements involved": lexical n-grams, that is, word forms rather than lemmas, grammatical forms or letters
2. "The *number* of elements involved": at least two consecutive words; the search was performed for up to six consecutive words, but no n-grams longer than four words long were found that appeared often enough to be included in the analysis (see section 6.1.1 for the thresholds applied in this study)
3. "The *number of times* an expression may be observed": in identifying n-grams that may be characteristic of a sub-corpus, minimum thresholds were set for both occurrence across a sub-corpus and text range (see section 6.1.1)
4. "The permissible *distance* between the elements": n-grams are here defined as consecutive words
5. "The degree of *lexical and syntactic flexibility* of the elements involved": for the purposes of this study, n-grams are not units of meaning but co-occurrence of consecutive words, so there is little flexibility; it was decided on a case-by-case basis whether similar phrases were combined (e.g. singular/plural difference) when presenting findings, but none were combined for the analysis itself
6. "The role that *semantic unity* […] play[s] in the definition": no semantic unity is assumed when identifying n-grams, but those combinations where no semantic unity can be

construed even with the help of co-text are mostly disregarded, especially if punctuation is involved (see section 6.1.3 for the principles of eliminating false positives).

For the analyses described in this chapter, I combined the two approaches: the extraction of n-grams (section 6.1.1) and the quantitative analysis of identifying keywords (section 6.1.2). This two-step process allowed me to compare typical word combinations from each sub-corpus to the other. Section 6.1 now describes the steps involved in this: N-grams were extracted from the texts in Sample A (see above section 4.3.2), that is, a sub-corpus of all identifiable 4* ICS (n=124) plus a sub-corpus of all identifiable 1*/2* ICS in those Units of Assessment where guaranteed 4* ICS can be found (n=93) (section 6.1.1). These n-grams were then compared to corresponding lists from other sub-corpora (section 6.1.2), with an additional step of eliminating false positives from the list of statistically significant differences (section 6.1.3). On this final list of statistically significant n-grams, I performed three analyses: coding for emerging themes (section 6.1.4), determining the level of editorial power (section 6.1.5), and suggesting persuasive functions (section 6.1.6).

6.1.1 Extracting n-grams

The first step of the analysis is quantitative and based on lexical frequencies. This then forms a basis for performing qualitative analyses, that is, in this case, checking how lexical items are used so that they can be categorised by their function in the text. Both broad steps are best carried out using specialist software. AntConc (Anthony 2014) is a widely-used and freely available corpus program that, among others, provides the following features that were used in this study: Wordlist generator, n-gram extractor, concordance (Key Word In Context – KWIC lines), concordance plots. Another program that was used is LancsBox (version 3, Brezina *et al.* 2015), which includes a particularly useful tool called “Whelk”. This shows the distribution of occurrences of a word across the different texts in a corpus and thereby helps to spot instances where a word appears to be prominent at corpus level in terms of its frequency, but is used mainly in a single text. The “Whelk” tool is illustrated in Figure 13, showing concordance lines in the upper half and the distribution across texts from one Unit of Assessment, with frequency and relative frequency, in the lower half.

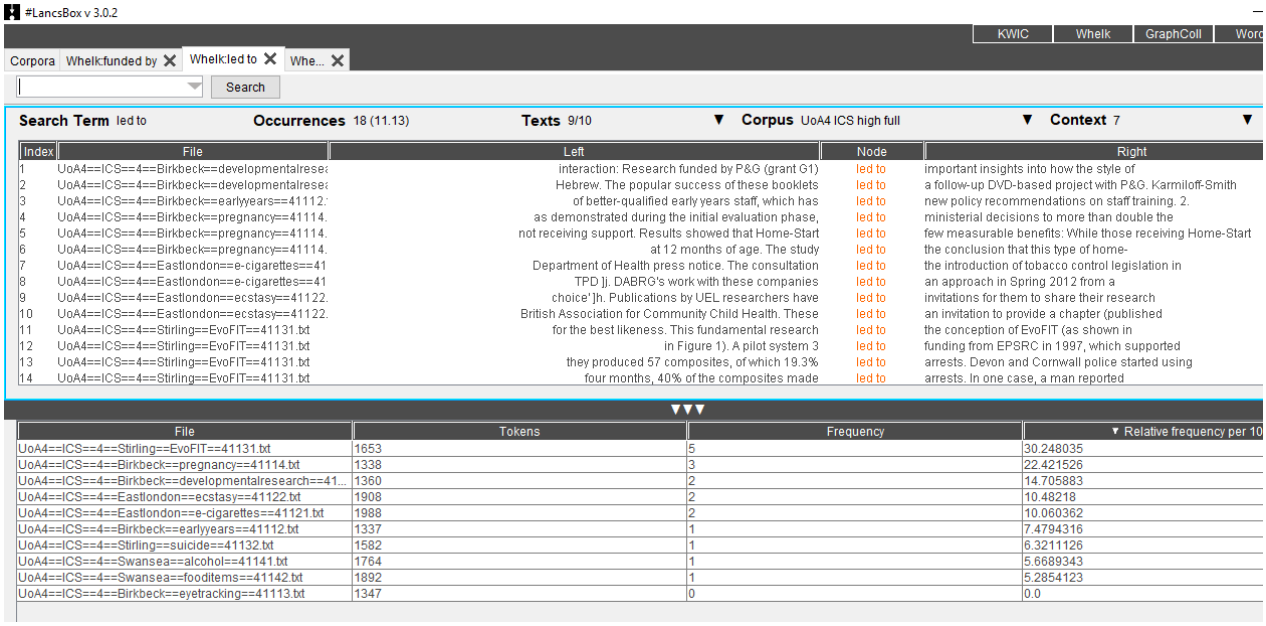


Figure 13: Screenshot of LancsBox 3.0, "Whelk" tool view

In order to identify word combinations for comparison between sub-corpora, n-grams were extracted with AntConc firstly from the corpus of all high-scoring ICS and then separately from the sub-corpora of high-scoring ICS in Main Panel (MP) A, C and D.[16] The corresponding lists were extracted from low-scoring ICS overall and separated by Main Panel. For each sub-corpus, 2-grams, 3-grams and 4-grams were extracted. The word string length (in AntConc: n-gram size) was first set at 2 for minimum and maximum size and then increased until there were no n-grams that met the required range as defined below. There were no 5-grams or above with the required range in either sub-corpus, therefore only 2-, 3- and 4-grams were studied.

To determine the representativeness of a type in a corpus (where a 'type' is a given form of word or, in this case, word combination), dispersion measures were used. While several measures are available that take into account text length and frequency of a type within a text, for this study the basic measure of range was chosen, with minimum thresholds for texts and submissions required for an item to be included in the analysis. It was deemed appropriate here because, rather than focusing on the number of times an item was used, this study investigates the number of texts, or even submissions, that used a particular expression. Moreover, the texts are of broadly comparable length within each sub-corpus

[16] For MP-B, only 6 high-scoring and 2 low-scoring ICS are clearly identifiable and available to the public, from one high-scoring and one low-scoring submission. The MP-B dataset is therefore too small for separate statistical analysis, and no generalisations should be made on the basis of just one submission per sub-corpus.

due to the 4-page limit of the ICS template, and therefore there is less need for the chosen dispersion measure to be sensitive to text length. The drawbacks of using range as a dispersion measure, namely that it is insensitive to both text length and the number of occurrences in a given text (Brezina 2018: 48), were mitigated through designing text range thresholds for extracting word lists in AntConc and through manually checking range thresholds for submissions. The specific problem of range being insensitive to a type appearing predominantly in one text while also appearing in other texts and therefore meeting the range threshold as a false positive was addressed before any qualitative analysis, as described in section 6.1.3.

To extract typical use in each corpus of interest, a *text* range threshold was set in AntConc at 10% of the texts in each sub-corpus: 12 for High (124 texts) and 9 for Low (93 texts). In the high-scoring part, individual submissions from different universities contained between two and 10 texts and these were likely to have gone through an editing process by the same person or team (see section 5.2.1 on the role of professional writers, and section 5.1 more generally on the process of creating ICS). Consequently, a further condition was set, namely that n-grams also had to appear in at least 25% of the *submissions* included in the corpus (n=10 for high-scoring ICS) to be considered "typical" in order to avoid skewing the data through the editorial preferences of a particular person or team, or university style guide. For each 3- or 4-gram which had a range below 20 and was not clearly tied to a particular institution (e.g. 3-gram *University of Bristol*), the Whelk tool (see Figure 13) was consulted to check that the submission range requirement was met and the n-gram was not centred on one submission (i.e. the same UoA and same university). For example, if an n-gram appears in 12 ICS but all of these are part of the same two submissions from the same one university, this would not be considered as meeting the range requirements because the word combination in question may reflect institutional guidelines or editorial preferences, rather than being a more widespread and therefore generalisable feature of ICS. The 124 high-scoring texts in the corpus stem from 40 different submissions, while the 93 low-scoring texts represent 47 submissions. With the thresholds set at 10% of texts and 25% of submissions, this meant that any n-gram identified by AntConc had to occur in at least 12 texts from at least 10 different submissions in the high-scoring sub-corpus or in at least 9 texts in the low-scoring sub-corpus, in order to be considered typical of that sub-corpus and included in the list of study items. The submission range requirement is not as pertinent in

the low-scoring sub-corpus, both because submissions usually consisted of only the minimum two ICS required by REF2014, and because there is less chance of editorial streamlining in these submissions. For comparisons of high- and low-scoring ICS within each Main Panel, the same principles applied. Table 20 shows the thresholds that were used in both the overall and Main Panel comparisons.

Table 20: Thresholds applied for inclusion in quantitative analysis

Main Panel	**No. in High**		**Range thresholds for High**		**No. texts in Low**	**Text range threshold Low**
	Texts	Submissions	Texts	Submissions		
Overall	124	40[17]	12	10	93	9
A	44	11	11	3	12	6
C	37	12	9	3	53	13
D	37	16	9	4	26	6

In addition to the details summarised in Table 20, the following observations should be noted:

- In MP-A, there is a large discrepancy in corpus size, with the sub-corpus of low-scoring ICS being about a quarter the size of the sub-corpus of high-scoring ICS.[18] This means that it was difficult to extract n-grams that are reliably significant (see section 6.1.2) because many entries had to be excluded from significance testing in the quantitative analysis. This was because the expected frequencies needed for reliable testing were often not achieved due to the small size of one of the samples.
- For MP-B, the small sample size, both regarding the number of texts and the fact that these texts stem from one high-scoring and one low-scoring submission only, would have made statistical comparison meaningless. Therefore this was not attempted.
- In contrast to MP-A, the MP-C sub-corpora are of near-equal size (see Table 5 in section 4.3.2 above). This means further that there was no large discrepancy between sub-corpora regarding expected frequencies (see section 6.1.2), and therefore more types could be included in the quantitative and therefore qualitative analysis.
- The size discrepancy of sub-corpora in MP-D sits between those in MP-A and MP-C.

[17] Note that the number of submissions in Main Panels does not add up to this number because the one submission in MP-B is included here but not represented in the breakdown by MP.

[18] Because low-scoring ICS were shorter on average, looking at word counts as displayed in Table 5 (section 4.3.2) the high-scoring corpus is slightly more than four times larger than the low-scoring corpus, despite the number of texts as shown in Table 20 suggesting the opposite.

To illustrate the scale of n-grams that could be extracted and where there was therefore a potential difference in use between sub-corpora, Table 21 shows the numbers of all types from each sub-corpus that met the text range requirements (12 texts for High, 9 for Low). In order to arrive at the overall number of unique types (last column), I combined those from the high- and low-scoring sub-corpora and eliminated duplicates, that is, those types that met the range thresholds in both sub-corpora.

Table 21: Number of extracted n-grams

n-gram	No. of types - High	No. of types - Low	No. of unique types meeting the range threshold in at least one of the sub-corpora
2-gram	1100	950	1360
3-gram	183	183	257
4-gram	22	22	44
Total	**1305**	**1155**	**1661**

To generate lists of n-grams against which the potential study types (meeting the range requirements set out in Table 20) could be compared for statistically significant difference in occurrence, the minimum text range was set at 2. This filtered out n-grams that only appear in one ICS and therefore are not representative of the corpus, and that would have inflated the number of types without adding value to the comparison. At the same time, this excluded word combinations that appeared only once in the whole corpus.

6.1.2 Determining key n-grams

Once the various lists of representative n-grams were extracted, comparisons were made between the sub-corpora of high- and low-scoring texts, both overall and at the level of Main Panels, to detect statistically significant over- and under-use in one set of texts relative to another. The lists of n-grams for each of the high-scoring corpus parts were compared to the corresponding low-scoring parts (High-Overall vs Low-Overall, High-MP A vs Low-MP A etc.). Texts from MP-B were included in the comparison of overall high- versus low-scoring ICS, but not in any of the Main Panel comparisons. Note that this is different to Sample C which was used for one analysis in chapter 5 and for chapter 7, where Sections 1 of MP-B texts were included in the MP-AB sub-corpus (see Table 4 in chapter 4 for the sample overview).

Because the aim was to isolate language use that is typical in one sub-corpus but occurs relatively less often in the other, each sub-corpus (e.g. High-Overall) was treated as a corpus

of interest in turn and compared to the corresponding one (e.g. Low-Overall, functioning as a reference corpus). The lists of 2/3/4-grams that met the range requirements in the corpus of interest were treated as word lists and compared to the full list of 2/3/4-grams in the reference corpus, using the various keyword measures described in the next paragraph. These full lists include all 2/3/4-grams that appear in at least two texts, rather than all 2/3/4-grams including those that appear only once in the corpus, for practical reasons: generating lists of all 2-, 3- and 4-grams that appear in just one of 124 texts would have resulted in extremely long lists. As soon as an n-gram appeared in more than one text, it would be reported as occurring twice and would be included in the analysis. Whether an n-gram appears in a sub-corpus in only one text, or not at all, does not change the assessment of representativeness – in either case, it would clearly not be representative of 124 (high-scoring) or 93 (low-scoring) texts respectively.

There are various statistical tests that are commonly used to classify keywords in a corpus, including Chi-Square, Log Likelihood and Log Ratio. The Chi-Square test is problematic because it is prone to overstating significance the larger the dataset gets (Oakes 1998: 28) and because it assumes normal distribution (in this case: of tokens throughout the corpus), which is often not the case with textual data (McEnery and Hardie 2012: 51). A non-parametric measure of statistical significance that is more appropriate for textual data is Log Likelihood (originally developed by Dunning 1993), which has been extended and applied to keyword measures in corpus linguistics by Rayson (see Rayson and Garside 2000). In the present study, Log Likelihood was therefore used as a measure of the statistical significance of frequency differences where normal distribution cannot be assumed (Rayson and Garside 2000). Brezina's (2018: 85) caution that "relatively small frequency differences […] can reach statistical significance in large enough corpora" can be disregarded here because this specialist corpus is much smaller than the large representative corpora that Brezina refers to in his explanations.

The accepted level of significance in corpus linguistics is $p<0.05$ (cf. McEnery and Hardie 2012: 51). However, Rayson *et al.* (2004) show that Log Likelihood has different levels of reliability at different expected frequencies (i.e. the frequencies that could be expected by chance in the different corpora if they were the same size, on which Log Likelihood is based). Therefore, in accordance with their conclusions, the $p<0.05$ level was applied for expected values greater than 12, but for those n-grams where expected values were lower, a higher

significance level was required for inclusion in the list of significant differences (Rayson *et al.* 2004: 8), as set out in Table 22. This is relevant to the corpus of ICS due to the relatively low number of words in each sub-corpus, resulting in many cases where expected values are low.

Table 22: Significance thresholds for Log Likelihood and expected minimum values for reliable application of the test (based on Rayson et al. 2004: 8)

p<	Log Likelihood	Minimum expected values
0.05	3.84	13
0.01	6.63	11
0.001	10.83	8
0.0001	15.13	1

Once statistical significance was established for an n-gram, Log Ratio (Hardie 2014) was used as a measure of effect size, which quantifies the scale of frequency differences between two datasets, as opposed to the weight of evidence for statistical significance quantified by measures like Log Likelihood. 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: 85) and the Log Ratio was chosen because it is straightforward to interpret: every additional point of the score represents a doubling in size of the difference between two corpora.

Of over 1,000 different n-grams, only a small proportion are used in significantly different frequencies, and of these, not all show an actual difference in use (see section 6.2.1, especially 6.2.1.4). In order to quantify the actual difference found as a proportion of the potential difference detectable through this method, I compared the numbers of all study types, that is, those types from each sub-corpus that met the text range requirements, to the number of study types that showed a significant difference between the study and reference corpus using the significance levels based on expected values as set out in Table 22. For both unique types and significantly different types, I added together those from the high- and low-scoring sub-corpora and eliminated duplicates, to arrive at the overall number of unique types and the number of significantly different types. Table 23 extends Table 21, and the final column shows the percentage of unique types that are significantly different between the high- and low-scoring sub-corpora.

Table 23: Number of extracted n-grams and number of significantly different n-grams

n-gram	High		Low		No. of unique types meeting the range threshold in at least one of the subcorpora		**% of types with significant difference out of types in corpus overall**
	Types	Types with significant difference	Types	Types with significant difference	Types	Types with significant difference	
2-gram	1100	157	950	179	1360	253	**18.60**
3-gram	183	23	183	17	257	26	**10.12**
4-gram	22	7	22	7	44	12	**27.27**
Total	1305	187	1155	203	1661	291	**17.52**

6.1.3 Eliminating false positives

Once statistically significant n-grams were determined, the list of 291 significant n-grams, plus the corresponding lists from the comparison within Main Panels A, C and D, were scrutinised for entries that were problematic for one of the following reasons:

1. Potential sampling bias: Some n-grams stood out as potentially skewing the data because they involved discipline-specific content words. For such instances, the distribution across Main Panels was checked, and if a content word appeared mainly in a UoA with a higher representation of either high- or low-scoring ICS in this corpus, it was deleted. For example, *pupils and* appeared significant in MP-C-Low because many ICS from that sub-corpus come from UoA25 Education. Therefore, education-related n-grams that showed as significant for low-scoring ICS were assumed to be significant due to sampling and were only kept if they also appeared in other sub-corpora (i.e. MP-A, MP-D or Overall comparisons). It is important to note that this process of deleting content words did not apply to n-grams that were more general in meaning but that were determined by the content of an ICS rather than being editorial choice (e.g. *house of lords*). Such n-grams were kept, if they applied to ICS across disciplines.
2. Distribution: Some n-grams appeared infrequently in a sufficient number of ICS to be included in the initial list (clearing the range threshold), but appeared very frequently in one ICS. These n-grams appeared as falsely significant because their frequent use in one text inflated the number of instances in one sub-corpus well beyond what was representative. For these n-grams, the significance was re-calculated counting only as many of the instances in the outlier ICS as the next highest ICS had, in order to gauge

whether their statistical significance relied on a single ICS. For example, the sequence "*e.g.*" in MP A-Low appeared 15 times in one of the ICS and 4 times in the text with the next-highest frequency, so the significance calculation was repeated for MP-A counting only 4 of the 15 instances in that first ICS. The recalculation suggested that the frequency of "*e.g.*" was not statistically significant, and the term was therefore deleted from the list of significant n-grams in the MP-A comparison (note that it was kept in the MP-D comparison, where the occurrences were distributed more evenly and therefore the significance was not skewed by one text).

3. Some sequences turned out to be meaningless, for example those that included punctuation in the original text, or sequences such as *and the*.

This step resulted in a revised list of 190 n-grams that met all of the criteria described above, from 291 unique entries in the overall, all-panel comparison that appeared statistically significant on a purely technical level. The final row of Table 24 compares this to the number of originally extracted unique n-grams. The percentage of meaningful, representative significant entries out of all entries is 11.44%. This shows that the proportion of n-grams that are used significantly more frequently in one sub-corpus compared to the other one is fairly small. As section 6.2.2 will explain, this number is further reduced if those entries are disregarded over which writers have little or no control, that is, those where editorial choice is limited by the subject matter of the research, impact or pathway described in an ICS.

Table 24: Number of extracted n-grams and number of significantly different n-grams, minus false positives

n-gram	High		Low		No. of unique types meeting the range threshold in at least one of the subcorpora		**% of types with significant difference out of types in corpus overall**
	Types	Types with significant difference	Types	Types with significant difference	Types	Types with significant difference	
2-gram	1100	157	950	179	1360	253	**18.60**
3-gram	183	23	183	17	257	26	**10.12**
4-gram	22	7	22	7	44	12	**27.27**
Total	1305	187	1155	203	1661	291	**17.52**
Types left after applying the additional criteria					**1661**	**190**	**11.44**

All figures in Table 24 (which extends Table 21 and Table 23) refer to the overall comparison between the high- and low-scoring sub-corpora only. The additional within-panel

comparisons yielded further n-grams that appeared significantly more often in the high- or low-scoring sub-corpus of just one Main Panel, compared to the other sub-corpus in that Main Panel. Many of these were duplicates of n-grams that are significant in the overall comparison. After eliminating such duplicates, the final number of types with a statistically significant difference in at least one of the comparisons (overall and/or one or several Main Panel comparisons) was 245, rather than the 190 reported in Table 24. It is this figure that underlies all quantitative reporting in the remainder of this chapter.

6.1.4 Coding by theme

From the lists of 245 unique n-grams (Appendix D) where both significance and effect size could be determined statistically and corroborated through manual checking for false positives as described in section 6.1.3 above, common themes emerged that appear to represent different functions. Each n-gram was assigned to a category by making a judgment about its predominant meaning in the texts, as reflected in the contexts captured in concordance lines extracted from the corpus.

In a first step, the entries in the lists of n-grams that were significantly different in the overall comparison were examined to identify common themes, that is, what kind of content seems to be emphasised or what kind of language occurs frequently. For this step, no assumptions were made about the linguistic function that an entry might have (such as grammatical or communicative function), but it did take into account what kind of content may be expected in an ICS and therefore what these entries might be deployed to say. The emerging categories were discussed with an expert on ICS, and a final set of categories was decided (listed and explained in 6.2.1). It has to be noted that this process of designing categories inevitably includes a degree of subjectivity; the aim was to find common themes that would describe typical kinds of content and language. Although replicability was enhanced by discussing categorisation with a second coder, it is possible that someone else going through the same coding process could have read some of the entries differently, even upon consulting concordance lines.

Once established, the categories were applied systematically to n-grams that appeared in any one of the comparisons, that is, those comparing the overall high versus low sub-corpora to each other and those that made comparisons within a Main Panel. Categories were applied on the basis of the contexts shown in concordance lines, and any duplicates that were key in more than one comparison were included, so that any specific uses in Main

Panels could be identified. In addition, the predominant use in the sub-corpus in which the n-gram was key was recorded separately. For some n-grams that were key in the low-scoring corpus, the use in the high-scoring corpus was also recorded, for example where the difference in frequency was small, or where the use in the low-scoring sub-corpus did not converge on a clear trend. In these latter cases, where there were for example two or three different uses, the high-scoring corpus was consulted to see if perhaps one of these uses did not occur there. The reason why this was not done for entries that are key in the high-scoring corpus is that the key n-grams in that corpus tend to be rare in the low-scoring corpus, and therefore the occurrences are unlikely to add substantially to the understanding of the use. This difference is due to the different corpus sizes. Entries where there was still uncertainty after consulting concordance lines in one or both relevant sub-corpora were discussed with two other researchers. The notes on use for each entry were again used in subsequent analyses, as described in sections 6.1.5 and 6.1.6. The categories and examples of use are presented in section 6.2.1.

6.1.5 Coding for editorial power

In addition to developing themes that emerged from the list of n-grams, each entry was coded for whether it was available as an editorial choice to all ICS writers, that is, whether it was within the writer's control to use it (e.g. *in terms of*) or whether it was restricted by the subject matter of the ICS (e.g. *policy makers* which is only available to those writers whose ICS includes engagement with or influence on policy makers).

For some entries, assigning a category was straightforward. For example, in most instances, the string *the government* would be included in an ICS because there was some interaction with a government in the research, impact or pathway described in the ICS, making it content-driven. However, for some entries, this was less obvious. For example, in principle all ICS should be able to refer to a *research team*, but not those where there is only one researcher. References to *researcher(s)* are a matter of editorial choice, that is, they should have been available to all writers and there would have been alternatives such as using names instead, but the specification *team* (or the plural form) is only available in those ICS where more than one researcher's work is described. Similarly, *the project* is a word sequence available to the vast majority of ICS writers, but not those whose research and impact did not include discrete projects (for example, unfunded research, or impact arising

from a body of research where a distinction into discrete projects was not relevant to the overall findings).

After initial coding, the principles for categorising entries were discussed with a second coder who has the dual perspective of having provided input to the writing of many ICS across UK universities, while also crafting their own ICS and making principled editorial choices for describing research and impact processes. Following this, a third category was introduced and I re-coded the entries five weeks after the initial coding round, to add a check of intra-coder reliability. The three final categories were:

1. **Free editorial choice** (e.g. *in terms of*). This label was applied to word combinations that could reasonably have been used by writers of all ICS, regardless of content.
2. **Restricted editorial choice** (e.g. *led by professor*). This label was used where a word combination was open as editorial choice in some, but not all, ICS.
3. **Content-driven** (e.g. *government policy*). This label denotes entries that are driven by the nature of the research, pathway or impact and therefore are only available in those ICS where this applied.

The process of assigning entries to these categories is summarised in Figure 14 and explained in more detail below that.

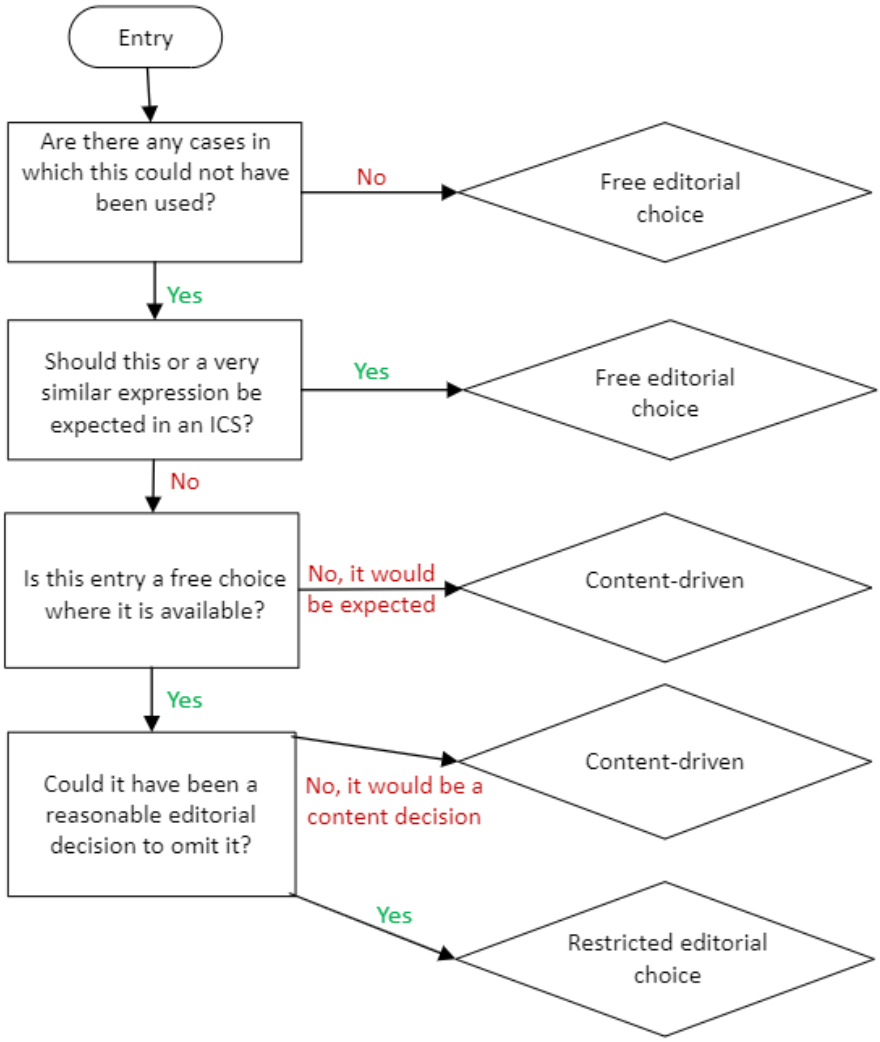


Figure 14: Decision process for coding an entry as editorial choice or content-driven

For each entry, I first considered the extent to which this sequence of words or a very close equivalent is available to all ICS writers, regardless of the nature of the impact or research, to determine whether the word combination could be classed as **free editorial choice**. The following test questions were applied:

a) Are there any cases in which this word combination could not have been used? This is the main inclusion criterion: if the answer was "no", the label "free editorial choice" was applied. For some entries, such as *resulting in*, the question whether the wording could have been used could be considered in different ways, and these were referred to question (b) to see if an entry should be included.

b) Is this or a similar expression something that should be expected across all ICS of at least 1* quality? Some entries appeared to be only applicable where certain impacts had happened or particular pathways to impact were used, for example *resulting in* or *contribution of*. However, all claimed impacts need to have had, or made, a

"contribution" of some sort and include a causal relationship between research and impact, making causal expressions an editorial choice in all cases. Therefore, conceptually the ability to use expressions that conveyed these kinds of meaning is essential for a classifiable ICS (i.e. one that was not deemed ineligible), regardless of how this was worded in specific texts. Entries that were deemed to be available to writers of all ICS because they denote content that is inherently necessary, such as causal (as opposed to temporal, for example) relationships, were treated as free editorial choice.

Those entries that did not necessarily apply to all ICS, that is, those where the answer to question (a) was "yes" and to question (b) was "no", were classed as **restricted editorial choice** or as **content-driven**. The following two test questions were both applied to all such entries to distinguish between these two categories:

c) Is this entry a free language choice in cases where it is available?
d) Could an editor have reasonably chosen not to include this word sequence when writing about this content?

If the answer to both questions was "yes", then the entry was coded as **restricted editorial choice**. For example, entries containing the words *Professor* / *Dr* / *research fellow* were coded as such because there are ICS where there was no involvement of a professor (question a: yes), but for those ICS that did involve a professor it was editorial choice to include or exclude the title (questions c/d: yes). By contrast, entries where the answer to questions c or d was "no" were coded as **content-driven**. For example, references to policy change are unlikely to have been left out due to editorial decisions, therefore the word combination *the policy* is not a language-driven choice.

Entries where there was still ambiguity after the second round of coding using the extended scheme and the test questions, and the small number of entries where the classification had switched from free choice to content-driven between the initial and the refined coding rounds, were discussed with the second coder to inform decisions about their status. The test questions were refined further, drawing on the combined experience of both coders having worked on hundreds of ICS for REF2021. In cases where an entry could be either content-driven (e.g. *"quality of" life*, used in health-related disciplines) or editorial choice

(e.g. *"quality of" the research*, open to all), the use of the expression in the corpus was taken into account, and the label was applied according to the majority of uses.

Further decisions and justifications are recorded in Table 30 to Table 35 in section 6.2.2.

6.1.6 Coding for elements of persuasion

Those entries that were coded as being editorial choice (free or restricted) were then additionally coded for whether they could be seen as containing an element of persuasion. Content-driven entries were not included in this analysis because they would not have been used with primarily editorial intent to persuade the reader, in other words with the aim to unduly "sell" impact (Watermeyer and Hedgecoe 2016: 1), but because they represent the nature of the research, pathway or impact.

As described in section 3.2.1, Dontcheva-Navratilova (2020: 28) argues that repertoires of linguistic means of persuasion are "genre-specific". ICS have a specific purpose with associated assessment criteria, which may confer persuasive meaning to otherwise neutral-looking language. Therefore, generic lists of persuasive language may miss genre-specific meaning. The closest existing list of persuasive language for ICS was developed by Hyland and Jiang (2023). However, their approach is top-down, that is, they used a wordlist of "hyping" terms that they generated initially from different places for a different corpus of research articles. They then ran that list through their ICS corpus, which is much larger than the one used here because it does not include an element of comparison across scoring brackets and the authors were therefore less restricted in their choice of texts. Despite the article stating that the final list could be made available upon request, when contacting the corresponding author, it became clear that the list used in their investigation of ICS could not be made available at the point it was requested because they planned to use it to analyse a further dataset. Moreover, with the experience of having worked on ICS myself, I wanted to take the opportunity of following a bottom-up approach, that is, to identify a more genre-specific repertoire of persuasive language. Therefore, existing lists of persuasive language were not used in this study.

In addition to the need for a bottom-up approach to identifying persuasive language, categorising such language needs to be register-specific. For example, based on their list as originally compiled for research articles, Hyland and Jiang (2021) developed categories of "hype": Certainty (e.g. *significant, crucial*), Contribution (e.g. *necessary, useful*), Novelty (e.g.

first, unique), Potential (e.g. *promising, potential*). However, research articles are a register with a different purpose (cf. section 5.1) and therefore different criteria for persuasion. In their analysis of ICS, the authors acknowledge that hyping was significantly more frequent in ICS than in research articles ($p<0.0001$, Hyland and Jiang 2023: 692), supporting the impression that persuasion has a different role in the two registers. One specific limitation of the research article hype categories is the inclusion of "Potential": while this category is meaningful for emphasising the significance of research described for academic peers in a journal article, it would not have swayed an ICS assessor unduly because "potential" impact, however promising, was ineligible for assessment (HEFCE 2011: 29, point 159). For the purposes of investigating persuasion in ICS, there was therefore a need for developing categories of persuasion specific for this register.

In order to develop categories of persuasion that would be meaningful in ICS assessment, a summary of the use of each entry based on concordance lines was considered from the two perspectives of "writer/editor" and "reader/assessor". The following questions were used to develop the initial codes and then again to assign entries:

1. Could this wording contribute to a positive perception of this ICS?
2. Would an editor choose this wording in order to convince the assessor of the significance or reach of the impact, or the strength of the link to the underpinning research?
3. From the reader perspective, could this entry be construed as having persuasive meaning in an ICS, regardless of its general use, which may or may not be persuasive?
4. If so, what is that persuasive meaning?

In addition to many entries where I did not perceive possible attempts at persuasion, the following categories emerged as ways in which ICS authors may enhance their texts: Credibility, Added Value, Richness and Specificity.

1. Credibility: e.g. emphasising the institution, or a researcher's involvement and credentials, for example by using job roles or using possessive first-person pronouns (*we have*, *our research*)
2. Added value: e.g. emphasising novelty or significance

3. Richness: e.g. giving examples and adding further context, that is, expanding the statement
4. Specificity: e.g. adding focus, or phrases that introduce specific information (such as dates), that is, closing down the statement (in contrast to Richness)

Richness and Specificity may not ordinarily be seen as forms of persuasion, but they are close to INTENSIFICATION and FOCUS respectively in the GRADUATION part of the Appraisal framework, which is a form of categorising evaluative language (see section 3.3). Perhaps more importantly, especially in ICS they add to the ability of an assessor to assign a higher rating, because they give context that can help explain or justify the value of a statement in relation to the assessment criteria.

Once these categories had been assigned to each entry that involves at least some editorial choice where persuasion could be construed, or the entry marked as "no persuasion", a second coder applied the categories following the same process, without access to my decisions. Cross-referencing then showed that there was a high level of inter-coder reliability (83%). I then resolved the discrepancies based on the notes of both coders and recorded the decisions made. Results are detailed in section 6.2.3.

6.2 Results

In this section of chapter 6, the findings from all three analyses are presented in turn. Section 6.2.1 reports on themes as outlined in section 6.1.4. This will be followed by a description of the separation into n-grams where a writer has editorial choice versus those that are determined by content (section 6.2.2), while section 6.2.3 reports on the various types of persuasion that can be found in those n-grams where there was editorial choice for at least some ICS writers.

6.2.1 Themes emerging from key n-grams

The first analysis started with a general assessment of the nature of the significant n-grams. From this, 11 themes emerged that in some cases split into subthemes, resulting in 23 different categories. Some of these categories include mostly n-grams that are key in the corpus of high-scoring ICS (section 6.2.1.1), others include mostly n-grams that are key in the corpus of low-scoring ICS (section 6.2.1.2). There are categories that include a more balanced number of n-grams key for each sub-corpus, but which point to a difference in the way that these categories were realised in the high- and low-scoring REF2014 ICS

respectively (section 6.2.1.3). Finally, in some categories, no clear difference in use can be seen between high- and low-scoring ICS (section 6.2.1.4). Each category represents a certain function in the text, either related to the content (e.g. Significance) or the narrative (e.g. Discourse). In this section, each of these predominant functions will be presented in a table showing the relevant categories with examples of n-grams that reflect them. Full details of all entries and their categorisation, with an overview of use in the relevant corpus of interest as explained in 6.1.4, can be found in Appendix D.

6.2.1.1 Functions typical for high-scoring ICS

As shown in Table 25, the functions that appear predominantly in the high-scoring corpus are related to the significance of the impacts claimed, especially indicating change or scale, or they point to the use of more specific information, as opposed to vague language, for example referring to the timing of the impact.

Table 25: Functions that emerged from n-grams that are key in high-scoring ICS

Category	**Definition**	**Number of n-grams**	**Example n-grams**
Significance – change	indicating the significance of the impact, especially highlighting change	2	- a new - are now
Significance – scale	emphasizing significance through scale of impact	6	- a major (*e.g. development*) - long-term - more than
Significance – official	indicating the significance of the impact, giving official backing	21	- department of (*usually referring to government department*) - policy makers
Date	anchoring an event somewhere in time	4	- after the - since the

Most of the functions typical for high-scoring ICS serve to contextualise the impact, for example by specifying either the scale of the impact or the change, or the scale of the problem, which in turn emphasises the need for change. As Table 25 shows, there are not many functions that seem typical for high-scoring ICS as opposed to low-scoring ICS, and they are not represented by many different entries. The exception is the category classified as Significance-official, which encompasses n-grams mainly related to public policy: government departments, lists of official bodies, mentions of the government directly, or

mentions of office holders. An explanation for this distribution could be that researchers behind high-scoring ICS, especially those in submissions that achieved 100% 4*, may have had more access to such bodies or contacts. This distribution could also indicate that impacts involving government officials or departments, or other official bodies such as the World Health Organisation, were highly valued by assessors. This in turn could be related to an observation that an influence on public policy is often linked to national reach and in-built significance. Indeed, Grant *et al.* (2015: 55) show that the word *policy* appeared at least once in 3,206 of 6,679 ICS in the database (48%), but their work did not distinguish between high- and low-scoring ICS. It is important to note that impact on government policy was not a necessity for a 4* ICS. For example, the sequence *house of* followed by either *commons* or *lords* appears in 27 out of 124 ICS in the high-scoring sub-corpus, that is, just over one fifth (21.8%). Moreover, the only n-gram typical for low-scoring ICS in this category (*policy and practice*) stands out because it is a vague summary phrase, rather than specifying the nature of the change in policy or practice. It therefore illustrates the difference in specificity between high- and low-scoring ICS. All but two n-grams representing this function are driven by content with no or little editorial choice, therefore giving no indication that word choice typical for high-scoring ICS unduly influenced the assessment.

6.2.1.2 Functions typical for low-scoring ICS

Turning to functions represented by n-grams that appear with a statistically significant higher frequency in the low-scoring ICS, it is important to note that some of these also appear often in the high-scoring ICS, but are relatively more frequent in the low-scoring sub-corpus. As set out in Table 26, these are generally related to pathways to impact or to research outputs, or they are indications of unspecific language where potentially opportunities may have been missed to make a more specific point within the tight constraints of the ICS template.

Table 26: Functions that emerged from n-grams that are key in low-scoring ICS

Category	Definition	Number of n-grams	Example n-grams
Beneficiaries	groups of people who benefit from the impact	5	- members of (*e.g. staff, the public*) - practitioners and
Pathway	Indications of pathways to impact	36[19]	- dissemination of - the book - the event
Research – output	Words related to academic publications (not needed in the main text of ICS)	4	- *et al.* - journal of
Research – topic	Words framing the subject of research	5	- explored the - research into
Reach – subnational	Words referring to geographical or other indications of reach at scales smaller than UK countries	7	- city council - of local - the north
Discourse	Signposting for the reader; accidental but frequent word combinations	20	- this case study - the following
Framing	Abstract concepts, often taking the form of *the issue of* – this is more concept-based than text based (compared to the Discourse category)	19	- the area of - the concept of - the issue of - the ways in which

The high prevalence of n-grams related to beneficiaries seems surprising at first, given that specific beneficiaries are a feature of high-scoring ICS (see section 5.3.2). However, this does not mean that low-scoring ICS specify their beneficiaries more explicitly. On the contrary, the fact that most n-grams in this category are key in low-scoring ICS indicates that these use more generic phrases that therefore appear more often, compared to the more specific and therefore more varied wordings used in high-scoring ICS which therefore do not appear in large enough numbers to show up as significant.

Some n-grams coded as pathway-related appear relatively more often in high-scoring ICS (n=9, 25%), usually as part of references to the range of activities within a particular ICS (e.g. *a series of*) or to the nature of the output (e.g. *report on*). However, the majority of entries

[19] Of these 36 n-grams that were pathway-related, 9 were key in the high-scoring, rather than the low-scoring, sub-corpus. Details are provided in Appendix D.

are key in the low-scoring ICS corpus (n=27, 75%). Based on the contexts reflected in the concordance lines, these can be mostly classified as discrete pathway activities (conference/book/event/project) or, within that, *dissemination*-related. The latter example is especially problematic because *dissemination* and *disseminated* are one-directional with a focus on process rather than resulting impact, and in an ICS with tight space constraints it is unlikely to be an effective use of words to describe such activity without specifying what this dissemination led to.

The relative prevalence of pathway-related n-grams, consistent with findings from the thematic analysis (see section 5.3.2), may in some cases stem from a narrative choice to focus on pathways – in the REF2014 dataset, it is likely that this has happened due to a lack of precedence and understanding of what is expected in an ICS. In other cases, however, it may also be representative of the potential content, where there may not have been enough evidencable impact to write a convincing ICS and the narrative therefore had to focus on what was available, namely pathways to impact. It is important to note here that descriptions of pathways to impact are a necessary part of ICS in order to establish the link between research and impact, but this was partly over-emphasised in ICS that scored low in 2014.

Similar to the picture in pathway-related material, the higher-than-expected presence of research-related entries in low-scoring ICS could support the observation that these devote more words to describing the research in relation to its academic outputs in the main text, instead of describing research findings (Section 2) or illustrating the impacts that it led to (Section 4). High-scoring ICS seem to confine mention of their academic outputs to Section 3 "References to the research", which is not included in the corpus.

The finding that sub-national reach is key in the low-scoring sub-corpus could suggest that local reach was less valued by assessors in REF2014. However, this could also be partly due to sampling bias. All entries are key in the overall comparison, rather than in comparisons within a Main Panel, so the focus on *local* could also be due to the high number of MP-C ICS in the low-scoring corpus. MP-C includes UoAs 22 Social work, 25 Education and 26 Sport and Exercise, and many of the ICS from these units in this corpus are based on work with *local authorities* or *city councils*. This is where many of the entries in this category appear. The claim here is not that local reach was typical of low-scoring ICS, but that MPs with higher

numbers of submissions that did not make it over the 3* mark (inclusion criterion for the low-scoring sub-corpus) were likely to work with bodies that have *local* in their name.

As well as this kind of potential sampling bias, there are other factors that could account for some of the differences. For example, it is possible that if local or regional impact was so central that an ICS narrative had to focus on that, it was not deemed impressive in REF2014. The present study did not consider the whole text of an ICS and what else may have been claimed in a given text that contained, for example, the combination *local authority*, and it therefore cannot report on whether claims indicated through discrete n-grams were central to a given ICS. Moreover, even within MP-D, *city council*, *of local* and *the local* are used significantly more at $p<0.05$ in the low-scoring sub-corpus, and within MP-C, the same is true for *and local* and *in local*. It is not true for the other n-grams in this category, though, and none of them are included as significant in the interpretation because of the small sample size and associated caveats with reliability (Rayson *et al.* 2004).

The final groups of n-grams typical for low-scoring ICS are directly related to textual decisions. They are split into Discourse, which includes signposting and word combinations that do not seem to be ICS-specific and therefore do not fit into any of the other categories, and Framing, which includes n-grams that introduce concepts or research questions. One potential inference of the predominance of these n-grams in the low-scoring corpus is that low-scoring ICS have more, or more formulaic, signposts around the text. The observation that most Discourse entries come from low-scoring ICS could also point to disciplinary differences, in that MP-C and MP-D which are over-represented in the low-scoring sub-corpus, use more of those explicit phrases than MP-A/B, which are over-represented in the high-scoring sub-corpus. However, some entries are also key in the comparisons within MP-C and MP-D, so a potential disciplinary difference can at most only partly explain this finding. A similar question about disciplinary bias through the sample composition arises for Framing-related n-grams, with the possibility that writers in MP-C and MP-D are more explicit about their framework or more accustomed to using such phrases. As in the Discourse category, though, some phrases are key in the within-panel comparisons in C/D, meaning that disciplinary difference cannot explain the finding completely and it is likely that a score difference is present.

6.2.1.3 Functions that are expressed differently depending on scoring bracket

Beyond functions that appear to converge on certain wordings in certain sub-corpora, there are functions that have key n-grams represented in both sub-corpora but with different emphases. These include n-grams related to Attribution and to Reach, which is expressed in different ways (see Table 27). The number of n-grams representing each of these functions is often similar in each sub-corpus, showing that the functions are not unique and therefore they are not treated as typical for either sub-corpus.

Table 27: Functions that emerged from n-grams that are key in either sub-corpus, but with different emphasis

Category	**Definition**	**Number of n-grams**	**Example n-grams**
Attribution – link	Cause and effect relationship, terms that point to a result – link between research and impact (or pathway)	High: 13	- cited in - led to the
		Low: 11	- of the research - through the (*refers to an activity, e.g. engagement/work*)
Attribution – agency	Making links and responsibility explicit: attributing research or impact to a researcher/team	High: 11	- by Professor - our research - research fellow
		Low: 8	- by Dr - research team
Reach – national and above	Geographical terms relating to at least a UK country, or to other countries	High: 10	- the world - in England
		Low: 4	- an international - nationally and internationally
Reach – unspecific	Terms that can relate to reach but are not further specified	High: 1	- across the
		Low: 3	- a number of - a range of

When making attributive links, high-scoring ICS were significantly more likely to include phrases that attributed impact to research, such as *cited in [policy documents]*, *used to [inform]* and *resulting in*. Those phrases indicate a focus on the effect, that is, on what the activity led to. In contrast, low-scoring ICS tended to link backwards, foregrounding the research activity (e.g. *[findings] of the research*). One reason for this discrepancy might be that the link back to the research was assumed, such that the text in high-scoring ICS can focus on the impact and link forward to that.

The category Attribution-agency includes, among others, n-grams which seem to indicate career stage. For example, *by Professor* is key in high-scoring ICS, and *by Dr* is key in low-scoring ICS. However, this distinction is only statistically significant in the comparison within MP-A. To investigate whether disciplinary differences play a part, the relevant entries were compared in the high-scoring sub-corpora of MP-A and MP-C directly, as these two sub-corpora in the same scoring bracket have a similar size. Here, two n-grams that include *Professor* are key in MP-A, and *Dr* also appears more often in MP-A but not to a statistically significant degree. Therefore, no generalisations can or should be made about career stages and ICS from this quantitative finding. This question has been investigated systematically in other studies (e.g. Smith and Stewart 2017), and the contribution of the present study is only to investigate the distribution of words where it is partly editorial choice whether the titles or job roles of researchers are included in the text (see also below in section 6.2.2).

Other entries in this category that are key in high-scoring ICS claim responsibility (*our research*) or emphasise the activity (*led by*), whereas those that are key in low-scoring ICS emphasise the submitting institutions (*the university*).

Turning to reach-related n-grams, it is again the level of specificity that differentiates between high- and low-scoring ICS. Even though one n-gram from the high-scoring sub-corpus was coded as Reach-unspecific (*across the*), this indicates more comprehensive (if unspecific) reach, as opposed to the vague quantification found in low-scoring ICS (*a number of*, *a range of*).

Other reach-related n-grams, referring to national and higher scales, show a similar pattern. All entries in this category that are key for low-scoring ICS contain the word *national* (or *international* in one case). By contrast, all entries from high-scoring ICS contain *UK*, *US* or *world*. This is curious, because “UK” and “national” could be read as essentially referring to the same entity. A possible interpretation is that high-scoring ICS specified jurisdictional reach, compared to low-scoring ICS that used more generic terms, leaving the reader in doubt about the actual reach.

None of the four UK countries (England, Scotland, Wales, Northern Ireland) are mentioned significantly more often in either high- or low-scoring ICS (outside of the phrase *in England and*). They do appear in a fairly even distribution, at least if two ICS about Northern Irish history are discounted, where *Northern Ireland* is a content word, rather than a term

denoting reach. *Wales* (n = 50), *Scotland* (n = 71) and *Northern Ireland* (n = 32) appear slightly more often in high-scoring ICS, but the difference is not significant (England: n = 162). An additional factor to take into account is that the dataset includes only submissions that are either high- 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- and low-scoring ICS in the sample from English, Scottish and Welsh universities, but no guaranteed low-scoring submissions from Northern Irish institutions.

6.2.1.4 Functions that appear similarly across scoring brackets

Finally, there are functions where no clear difference can be detected between the ways that high- and low-scoring ICS seem to represent them. The functions and n-grams described in Table 28 illustrate that formulaic use appears in both sub-corpora, even though the exact words may differ. It shows that even though it is possible to measure lexical differences between sub-corpora, these differences are not always meaningful.

Table 28: Functions that emerged from n-grams that are key in either sub-corpus, with no clear difference

Category	**Definition**	**Number of n-grams**	**Example n-grams**
Research – general	n-grams relating to research but not necessarily confined to Methods/Results	High: 6	- research programme - the study
		Low: 9	- body of - research project
Research – results	References to research results	High: 1	- showed that
		Low: 1	- that the
Impact descriptions – awareness	Description of the nature of the impact	High: 1	- public awareness
		Low: 3	- knowledge and - understanding of the
Impact descriptions – capacity		Low: 3	- and skills - professional practice
Impact descriptions – change		High: 3	- to improve - reduction in
		Low: 1	- to establish
Impact descriptions – general		High: 6	- the policy - impacts on (*often as signpost, e.g. in headings)*
		Low: 5	- an impact on - the impact
Partner	Partner institutions or organisations involved in the research or pathway	High: 2	- the national (*e.g. gallery*) - worked with
		Low: 3	- centre for - with the
Significance – general	Emphasising significance but not further specified	High: 3	- the key - up to
		Low: 4	- one of the - the value

In the Research-general category, the meanings of entries from high- and low-scoring categories are fairly similar. Most occur frequently in both corpora despite the statistically significant differences, so this cannot be turned into a recommendation not to use those phrases that happen to appear even more in low-scoring ICS. The examples selected for Table 28 illustrate this for one frequent question, namely whether it is better to frame an ICS around a single research project or a larger body of research. This question can of course not be answered by a purely linguistic search, but it is striking that the two sub-corpora each have key n-grams pointing in these different directions: *research programme* (high) and *body of* (low) indicate a multi-study ICS, whereas *the study* (high) and *research project* (low) are near-synonyms referring to a single study. It may be that these n-grams are used in ICS that

are actually based on a research programme consisting of several research projects. Therefore, this shows that many of the measurable differences do not necessarily point to either a favourable word choice or favourable content.

Combined with the finding that research-related word combinations are key in low-scoring ICS (as described in section 6.2.1.2), the higher number of research-related n-grams in this category raises the question whether research-related material is generally used more in low-scoring ICS. To check this, the word *research* by itself was compared across sub-corpora, and it was found to appear frequently in both sub-corpora. In the sub-corpus of 124 high-scoring ICS, it appears 2019 times, which is normalised to 91 per 10,000 words. In the sub-corpus of 93 low-scoring ICS, it appears 1450 times, normalised to 110 per 10,000 words and therefore used relatively more often. This is a significant but small difference ($p<0.0001$ and effect size measure Log Ratio = 0.27, where at least 0.5 is usually required for Log Ratio to indicate a meaningful effect). Collocations to the right in both sub-corpora are: *on, has, 's, by*, with *'s* appearing further down the list in the low-scoring sub-corpus. The word *impact*, in the collocation *research impact*, appears at no. 19 in a frequency-based collocations list.

The only observation from the Research-results category might be that ICS in general do not seem to refer to research results very often – although given the methodology, this would not be a valid inference because there may be other wordings that are equally used by both high- and low-scoring ICS and that therefore do not show up as over- or under-used. It does stand, however, that the mention of research results seems to be similar across ICS.

The category Impact descriptions does not map onto typologies of impact but includes word combinations that describe the nature of the impact, which can be a specific impact type (in this case, Impact descriptions – awareness/capacity) or a more general n-gram that could be applicable across different impact types (Impact descriptions – change/general). A likely explanation for the relative dearth of n-grams in this category is that impact descriptions, especially those that indicate the type of impact, would be more specific, especially in the absence of an a-priori categorisation demanded by the REF guidance, and therefore they would not cluster in n-grams across ICS. They might cluster if they are less specific, which, based on discussions in this chapter, may be more expected in low-scoring ICS. For example, the n-gram *professional practice* is used as an umbrella term, and a high-scoring ICS that was written with more specific wordings may have used *police practice* or a similar specification of *professional*.

Finally, regarding Significance, outside of the more specific categories above (section 6.2.1.1), there is no obvious difference between the words and phrases that denote significance from high- and low-scoring ICS.

Overall, this analysis shows that most word combinations that appear significantly more often in one scoring bracket, and the functions which are represented through those word combinations, are key for low-scoring ICS. This distribution indicates that, rather than there being many specific phrases or "power words" (Van Noorden 2015: 150) that would have "hyped up" (Hyland and Jiang 2023: 685) ICS to achieve higher scores, any perceived linguistic superiority of high-scoring ICS is more likely to stem from their authors using more specific phrases that do not converge in this scoring bracket across ICS. Moreover, many of the significantly different n-grams are driven by content (e.g. *city council* or *government policy*) rather than being editorial choice, and it is not credible that these could skew the ICS assessment towards higher scores for impacts with less significance or reach. The next section takes a closer look at the distinction between n-grams that were available as choice for writers and those that were pre-determined by the content of an ICS.

6.2.2 Editorial choice or content-driven?

So far, I have described the categories in which the n-grams can be grouped according to content (e.g. related to significance, reach, pathway). A separate approach to categorising statistically significant n-grams encompassed the division into those that are language-driven and those that are content-driven differences, as described in section 6.1.5. The aim was to determine which of the n-grams writers of ICS have editorial control over. Some n-grams are part of the impact, such as the combination *England and Wales*, which mostly appears in the context of NHS trusts and therefore the co-occurrence is due to the legal structure of the NHS, not to authorial decisions. Others (e.g. *resulting in*) are part of the narrative and are therefore a matter of linguistic choice, rather than being pre-determined or restricted by the research or the type of impact.

If language differences are at least partly driven by pre-existing content restrictions, the assumption that language choice may have influenced assessors beyond the substance of the ICS is less valid, because it reduces the number of language differences that can be freely deployed at the editing stage.

6.2.2.1 Quantitative overview

Discounting duplications (of exactly the same word forms) where an n-gram was significant both in a Main Panel comparison and the overall (high versus low) comparison, there are 245 unique entries. Of these, 103 are key in the sub-corpus of high-scoring ICS, and 142 in low-scoring ICS. Table 29 shows the distribution of these entries across the three levels of choice.[20]

Table 29: Number of key n-grams from high-scoring and low-scoring ICS in each of the three editorial categories: free editorial choice, restricted editorial choice and content-driven wordings

	Free editorial choice	Restricted editorial choice	Content-driven
High (n=103)	33	25	45
Low (n=142)	83	30	29
Total	116	55	76

This indicates that most of the terms that stand out as key in high-scoring ICS are content-driven (n=45). By contrast, more than half of the terms that stand out in low-scoring ICS are language-driven, that is, free editorial choice (n=83). This distribution does not suggest that frequently used words alone persuaded assessors to award higher ratings. If the majority of entries that appear more often in high-scoring ICS had been language-driven, then these could more likely resemble the "power words" as suggested by Van Noorden (2015: 150), and editors could (mis)use such a finding to include those words or expressions in order to achieve a higher star rating. However, in this dataset, the relative majority of terms that stand out in high-scoring ICS are content-driven, suggesting that their scores were based on the substance of their content. It is of course possible that high-scoring ICS were embellished or presented their impacts in a particularly favourable narrative, but the evidence suggests that this does not converge on certain n-grams. Conversely, there were more editorial language choices that stood out in low-scoring ICS, for example those relating to academic publications, as explained in section 6.2.1.2. Overall, the distribution shown in Table 29 indicates, if anything, missed opportunities for writing a more specific and convincing ICS,

[20] While it is possible to calculate the percentage of n-grams with a certain level of choice out of the total number of n-grams, this level of specificity might be too easily misunderstood because the figures in Table 29 and the following text relate to the number of n-gram *types*, that is, different n-grams where editors have a certain level of choice. It does not refer to the number of *occurrences*, and some of the n-grams that are, for example, content-driven may occur many more times than others. I feel that adding percentages in the text may risk inviting misunderstandings. The data does not include information on the percentage of n-gram *occurrences* that relate to a specific level of choice. This same caveat applies to the information presented in section 6.2.3.

and certainly not the existence of certain phrases or combinations that could have been freely deployed and were used more by high-scoring ICS.

6.2.2.2 Examples of n-grams in each category

This section discusses examples of entries from the high- and low-scoring sub-corpora, explaining why they were seen as being free or restricted editorial choice, or as content driven. This is a selection, and the full list of entries can be found in Appendix D: List of n-grams that are significantly more frequent in either high- or low-scoring.

Free editorial choice

This label was applied to word combinations that could reasonably have been used by writers of all ICS, regardless of content. Guiding questions, as summarised in Figure 14 (see section 6.1.5), were:

a) Are there any cases in which this word combination could not have been used?
b) Is this or a similar expression something that should be expected across all ICS of at least 1* quality?

Table 30 and Table 31 provide examples of n-grams of free editorial choice that are typical for high- and low-scoring ICS respectively.

Table 30: Examples of n-grams that were free editorial choice and appeared predominantly in high-scoring ICS

Entry	Justification for placement in this category
the key, up to, including the, by the	These n-grams are not specific to ICS so they could have been employed in any text.
research on, the study, showed that	Each classifiable ICS had to be able to write about research, therefore these generic wordings were seen as universally applicable.
since the	This expression was mainly used to refer to specific points in time, including years (the corpus analysis tool was set to ignore numbers, so *since 2009, the…* would be read as *since the*), or to specific points in the narrative (e.g. *since the introduction of…*). One or both of these meanings should have been available to all ICS writers.
led to, result of, resulting in, the basis	For a classifiable ICS, there had to be a causal relationship between doing (or having done) research and the claimed impact. Therefore causal phrases are conceptually open to all ICS writers. A potential objection to this classification could be that these causal expressions could imply linearity, which was not a requirement. However, these expressions could refer to causality and linearity within research or impact, as well as between research, pathway and impact. Moreover, non-linear impacts can still be described in linear ways, and often were in REF2014. Therefore the causality aspect, pointing to free choice, was given more weight in this decision than the linearity aspect which would point to restricted choice.
used the, used to, using the, was used	Entries including *use* also convey causal links. In comparison with *led to* and *resulting in*, they imply a more specific relationship, e.g. that someone uses the research findings or a toolkit or other material forming a pathway to impact. Very strictly speaking, this is not a necessity for writing an ICS of at least 1* quality, but because the conceptual causal connection is such a strong aspect of these expressions, they were classified the same as those in the previous row.
contribution of, impacts on	Similar to the previous n-grams, *contribution of* was included here because a contribution is a necessary component of a claim that some research (activity) has led to an effect. At the same time, there are multiple ways to write about a contribution and about impacts without using either of these words, therefore their inclusion is free editorial choice.

Table 31: Examples of n-grams that were free editorial choice and appeared predominantly in low-scoring ICS

Entry	Justification for placement in this category
focus on, in relation to, the ways in which	These and other n-grams in the Discourse category (see above, 6.2.1.2) are not specific to ICS.
as part of the, one of the	These entries were discussed as potentially a restricted choice because they assume that there was a part-whole relationship, but due to the variability of use observed in the corpus, they were classed as free choice.
university of	References to universities should be open to all ICS, even if technically not all submitting units are placed in a *university of [place]* – a close equivalent was available to all, and references to the institution were free choice. Note also that this entry is key for low-scoring ICS in the overall comparison and therefore discussed in this table, but it is key for high-scoring ICS in the MP-A sub-corpus. The difference in use is that the MP-A high-scoring ICS generally use this to refer to the submitting unit, while half of all uses in the low-scoring ICS, regardless of MP, refer to a partner university.
the value, contribute to	Similar to the justification in Table 30 for *contribution of*, these n-grams were seen as free choice because some added value is necessary for claiming impacts.
through a / the	These sequences normally refer to pathways to impact in this corpus, and not all pathways are technically "through" something, but because the terms convey causal relationships which are necessary for research-to-impact claims, they were classed as free choice.

Restricted editorial choice

This label was used where a word combination was open as editorial choice in some, but not all, ICS. Guiding questions were:

c) Is this entry a free language choice in cases where it is available?
d) Could an editor have reasonably chosen not to include any of the words in this n-gram when writing about this content?

If the answer to both questions was "yes", the entry was included in this category, otherwise it was seen as content-driven.

Table 32 and Table 33 provide examples of n-grams of restricted editorial choice that are typical for high- and low-scoring ICS respectively.

Table 32: Examples of n-grams that were restricted editorial choice and appeared predominantly in high-scoring ICS

Entry	Justification for placement in this category
evidence on	In principle, all ICS needed evidence, but in this corpus, this particular sequence was generally used for giving evidence to parliament – which was in these cases not a free choice. *Evidence* on its own would be free editorial choice, but with *on* it is restricted to this use, which was only available to some ICS. By contrast, *of evidence* is possible to use more widely and was therefore classed as free editorial choice.
after the, are now	These sequences could be seen as causal expressions, but their use leans much more towards temporality or sequence. They were therefore read as restricted editorial choice for cases where temporality or sequence mattered, because in these high-stakes texts causal relationships were normally expressed with a more explicitly causal phrase, as indicated by the high number of causal phrases that appeared across a wide range of ICS.
more than	This entry could be seen similar to *the value* (see Table 31) and similar expressions that are a conceptual necessity for claiming impact, but it does not apply to all ICS, such as those claiming the reduction of harm.
long term, the most	It was not a necessity to describe impact with long-term effects, or those that included superlatives. Similarly, for those ICS that did include such content, other word choices would have been available.
led by professor	References to *professor* and other academic designations or job titles were seen as editorial choice in those cases where a person with that title was involved, which was not universally the case.
school of, institute for	Unlike entries containing *university,* which were classed as free choice, these entries were seen as editorial choice only in ICS submitted by a university divided into schools, institutes or departments respectively. Note though that entries containing *department* were classed as content-driven because the vast majority of uses (not all, hence the mention here) referred to government departments, rather than university departments.
set up	This entry could be seen as content-driven because setting up something is a specific type of impact or pathway. However, there are possible alternative word choices such as *establish*, so this term was classed as editorial choice, albeit not free choice.
cited in	This word combination generally introduced citations in policy documents, guidance, white papers or parliamentary debates. It was therefore seen as restricted to some ICS based on content, but as editorial choice within those.

Table 33: Examples of n-grams that were restricted editorial choice and appeared predominantly in low-scoring ICS

Entry	Justification for placement in this category
research team	References to research teams, including the use of first person plural, were seen as choice but not for those ICS reporting on the research by one academic.
project was, the project, research projects	While the vast majority of ICS were based on research, pathway or impact projects, this does not apply to all.
working in	This n-gram mostly refers to beneficiaries, such as people *working in* a certain environment. This does not apply in all ICS, but there are different words to describe this (e.g. *employed by*).
the development of	Referring to either research or impact descriptions (e.g. training programme, software tool), this n-gram is not available in all possible ICS, but there are synonyms available.
the event	While this entry could be seen as content-driven, it is an umbrella term and a writer can choose to use a more specific word, such as *conference*.
et al, journal of, the paper	It could be seen as free editorial choice to include references to research papers in the main text (note that Section 3 "References to the research" was not included in the corpus), but journal papers or multi-authored publications are not the norm in all disciplines, therefore these entries were classed as restricted editorial choice.

Content-driven

This label denotes entries that are driven by the nature of the research, pathway or impact and therefore are only available in those ICS where this applied, yet are likely to be included where applicable. Table 34 and Table 35 provide examples of content-driven n-grams that are typical for high- and low-scoring ICS respectively.

Table 34: Examples of n-grams that were content-driven and appeared predominantly in high-scoring ICS

Entry	Justification for placement in this category
improve the	Based on the use in this corpus, this refers to specific types of impact: improving e.g. lives, quality or teaching. It is therefore not universally available, but where there was a clear improvement in a certain category, it would have been marked as such.
quality of	Similarly, this combination generally referred to *quality of life* rather than e.g. *quality of the research*, making it a content-driven phrase.
in the UK, in the US, the world	Geographical designations were classed as content-driven because they were unlikely to have been left out where they applied.
Department for	The vast majority of uses in the corpus refer to government departments, in the UK or other countries. Such involvement, whether as pathway or as part of the impact claim, is unlikely to have been edited out.
Government policy	Similar to references to departments, where government policy is part of the pathway or impact claim, it is likely to be a key component and therefore named as such. This combination often appeared in headings or lists of overarching impacts, often with a verb of change to the left and sometimes a country to the right.
Wellcome Trust	While it is conceivable that such funding was not referenced in the main text, it was certainly a content-determined combination, rather than editorial choice, with alternatives such as *prestigious funding* losing meaning in a way that is unlikely to have happened in these high-stakes texts. Note also that Section 3 in REF2014 did not explicitly provide space for grant information, unlike the REF2021 template, and therefore it was more likely in REF2014 that a Wellcome Trust grant was labelled as such in the main text of Section 2 than it would have been in REF2021.

Table 35: Examples of n-grams that were content-driven and appeared predominantly in low-scoring ICS

Entry	Justification for placement in this category
nationally and internationally, of England, the north	See Table 34 for considerations of geographical designations. Additionally, while combinations containing *national* or *UK* could in theory be seen as interchangeable and therefore to some extent as editorial choice, in practice *national(ly)* is generally part of a collocation with that adjective (e.g. a national level, a national event) in this corpus, making it content-driven.
knowledge and, understanding and	Arguably these entries could be seen as conceptually necessary for an ICS, similar to "contribution" and causal links. However, where they refer to researchers, the necessity does not apply: it is not always the researcher's knowledge that matters for creating impact, it can also be their skills (e.g. in statistical analysis) or much more narrowly the results of a research project with a large team. Therefore it is not necessary to highlight the researcher's knowledge, especially in more scientific disciplines in MP-A and MP-B. More importantly, the vast majority of instances of these term in this corpus refers to increased knowledge or understanding of beneficiaries, making these occurrences references to specific types of impact and therefore content-driven.
the book, the conference, the training	Combinations such as these refer to specific pathways to impact and would be difficult to replace at editorial stage.
city council	This refers to a specific type of partnership/pathway or impact.
professional practice	Similarly, references to professional practice describe specific types of impact.

Entries in the three categories allow different kinds of observations. Those that are content-linked can lead to conclusions about the kind of impact or pathway that may have appeared more often in high- or low-scoring ICS, but writers or editors have little control over whether to use them; that is, they can only choose to use them if the content allows it. These n-grams can be indications of the relative value assigned by assessors to certain types of research, impact types, or reach. By contrast, entries that were in principle available to all writers can contribute to a description of language-related differences between sub-corpora. Recommendations derived from each therefore also differ. Those that are content-linked could be addressed to people who are planning and developing impact (e.g. "engagement with policy makers may have been valued by assessors in 2014"). By contrast, language-driven differences could inform recommendations for writers and editors of ICS and can still be useful at the end of a REF cycle, regardless of the pathway to impact. This includes both free and restricted editorial choice.

Most importantly, it is only the editorial differences (whether they are an available choice in all or some cases) that may be potentially problematic for the integrity of ICS assessment if there was a correlation between these choices and the star ratings. Content-driven differences simply illustrate expected differences between high- and low-scoring impact, and if they affect the "oomph" (Gow and Redwood 2020: 69) of an ICS, that would likely still be consistent with the content that was being assessed, and therefore with the score.

6.2.3 Editorial choices that could be linked to persuasion

Following on from the analysis of the degree of choice that ICS writers were likely to have had, those entries that were considered to have been available as editorial choice (free or restricted, as discussed in section 6.2.2) were scrutinized for their potential to have contributed to persuading the reader of the quality of the impact claimed, and therefore to have influenced the assessment. Here, the entries treated as "free editorial choice" were assumed to be more liable to having been used for "dressing up" because they would have been available to all writers. By contrast, those where the entry is restricted as an editorial choice for only some ICS, for example those involving research by a professor, could not have been deployed as persuasive elements in all ICS and therefore their role in selling impact could not have been as broad as that of those entries that were available to all.

6.2.3.1 Quantitative overview

Of the 171 unique entries where there is editorial choice, 65 (38%) could be construed as attempting persuasion. Of these, 36 are key for high-scoring ICS and 29 are key for low-scoring ICS. Note that all figures in this section refer to the number of different n-gram types in each category, rather than the frequency or distribution of the function in a sub-corpus. This would not be meaningful because the entries come from different sub-corpora and figures of use are not normalised. Moreover, a distribution of functions represented by key n-grams would be misleading because it misses all the ones that are not part of a (typically occurring) n-gram.

Table 36 shows how n-grams with persuasive meanings are split into the different categories introduced in section 6.1.6.

Table 36: Distribution of categories of persuasion (number of different n-grams per sub-corpus), overall

	Choice	Of which: Persuasion	Sub-categories of persuasion			
			Credibility	Added Value	Richness	Specificity
High	58	36	22	6	3	5
Low	113	29	15	3	4	7
Total	**171**	**65**	**37**	**9**	**7**	**12**

Taking into account the distinction between the sub-categories of Choice, namely free and restricted editorial choice, Table 37 shows how many of those entries that were considered as free editorial choice were assigned to each category.

Table 37: Distribution of categories of persuasion (number of different n-grams per sub-corpus), free editorial choice

	Free Choice	Of which: Persuasion	Sub-categories of persuasion			
			Credibility	Added Value	Richness	Specificity
High	33	16	10	1	2	3
Low	83	20	9	2	4	5
Total	**116**	**36**	**19**	**3**	**6**	**8**

The split into categories for entries regarded as restricted editorial choice is summarised in Table 38.

Table 38: Distribution of categories of persuasion (number of different n-grams per sub-corpus), restricted editorial choice

	Restricted Choice	Of which: Persuasion	Sub-categories of persuasion			
			Credibility	Added Value	Richness	Specificity
High	25	20	12	5	1	2
Low	30	9	6	1	0	2
Total	**55**	**29**	**18**	**6**	**1**	**4**

From these figures, it can be seen that most differences in word choice that can be detected through key n-grams are **not** related to persuasion. For those wordings that are open to all writers (free editorial choice), only 31% (36 of 116 n-grams) were classified as persuasion-related, and for those available to some (restricted editorial choice) this figure is 52.7% (29 of 55 n-grams). Overall, only 38% of the different entries where there is free or restricted editorial choice were classified as persuasion-related. This means that less than half of all language differences that could be introduced at editorial stage can be seen as expressing persuasion, and therefore claims in the literature that ICS were predominantly exercises in

persuasion to the point of compromising the assessment outcome are not supported by this finding.

6.2.3.2 Examples of n-grams in each category

This section discusses examples of entries from the high- and low-scoring sub-corpora, with examples of where and how they occur in the corpus and an explanation for their placement in the respective category. In the tables (Table 39 to Table 42), the third column ("key in…") refers to the sub-corpus, that is, whether an expression appears significantly more often in high- or low-scoring ICS, and whether this is the case in the overall and/or a Main Panel comparison. The fourth column ("ICS Section") refers to sections in the REF2014 template, where Section 1 is "Summary of the impact", Section 2 is "Underpinning research", and Section 4 is "Details of the impact". This is a selection and the full list of entries can be found in Appendix E.

Credibility

As shown in Table 39, this label was applied to word combinations that emphasised the researcher's or their institution's role in creating impact, or that otherwise enhanced the credibility of a claim, for example by pointing to the involvement of official bodies.

Table 39: Examples of n-grams showing credibility

Entry	Degree of choice	Key in…	ICS section	Justification for category, based on use in the corpus
cited in	restricted	High (overall / MP-C)	4	Use of this word combination either establishes a link between research and (pathway to) impact, or otherwise invokes credibility. **Use:** MP-C: introduces citations in policy documents, guidance, white papers, Lords debate. In other panels: similar but can also refer to mentions in websites, newspapers, reports - very occasionally an academic publication.
led by professor	restricted	High (overall)	2	Even where a professor takes the lead, this does not need to be stated in Section 2 which is by definition a report on activity in the submitting unit, and a person's name could have been used without title. The inclusion of the title can therefore be seen as an attempt to bolster credentials. **Use:** A *team, project, research group, study* or similar led by a professor, very often used together with the name of a university.
contribution of	free	High (overall)	2	This also emphasises the link between academic research and effect. **Use:** Specifying the nature of the contribution of the research, funder or impact.
peer reviewed	free	Low (overall)	2, some 4	The term is added to emphasise quality, that is, acceptability by relevant academics. However, this is not what is being assessed beyond a certain threshold, and high-scoring ICS generally did not emphasise this acceptability in the main text. **Use:** Mostly followed by *publication, book, article, journal*; occasionally *research, conference, award, evaluation.*

Added Value

This label was applied to word combinations, listed in Table 40, that conveyed novelty or significance, which are both ways to emphasise the quality of an impact claim.

Table 40: Examples of n-grams showing added value

Entry	Degree of choice	Key in…	ICS section	Justification for category, based on use in corpus
more than	Restricted	High (overall and MP-A and MP-C)	2 and 4	This n-gram was seen as emphasising added value because it introduces an undisclosed but higher-than-stated number that serves to claim the best-possible impact – the widest reach, solving the biggest problem. **Use:** MP-A: usually followed by a number, then followed by (potential) beneficiaries. Either scale of the impact or scale of the problem. Occasionally scale of pathways (e.g. *more than 30 news articles*) MP-C: usually followed by a number, which is then followed by a pathway or beneficiary word (*net income, teachers*) Overall: 73x with a number; followed by countries, companies, educational institutions, people (*students, visitors*…)
a new	Restricted	High (overall and MP-D)	Any, but mostly 4 (esp. MP-D)	The notion of novelty indicates added value in the context of competitive impact assessment: something new exists that did not exist before and is very likely presented as a desirable addition to the world. **Use:** MP-D: to the left: develop/establish/launch; to the right: approach, direction, research - or description of whatever the impact is (a new website, youth theatre…) Overall: Left collocates: variations of *create, develop, lead to, recommend, establish, implement*. Right: various content words, whatever the new impact is
the key	Free	High (MP-C)	Any	This n-gram helps to articulate added value by elevating the importance of an entry. **Use:** Right collocates: *actors/drivers, messages, recommendations*
one of the	Free	Low (MP-C)	Any	Taking into account the right collocates, this wording seems to emphasise relative importance. However, it is a vague expression that leaves the scale of comparison open. *One of the* could mean “one of two” or “one of ten”. It also precludes any claims of uniqueness, even though it takes the guise of implying rarity. **Use:** one of the *few/first/most important/main*

Richness

The entries in this category (Table 41) expand a statement by giving examples or by otherwise adding context that could help explain or justify how statements met assessment criteria. They enhance the richness of the claim, which gives assessors more opportunity to be convinced of the merit of the claimed impact.

Table 41: Examples of n-grams showing richness

Entry	Degree of choice	Key in…	ICS section	Justification for category, based on use in corpus
e.g.	Free	High (MP-D)	any	High-scoring ICS give examples with minimum word/space investment and maximum signposting. **Use:** usually in brackets to give examples of stakeholders/beneficiaries or pathways.
as well as	Free	Low (MP-C)	any	Listing either the researchers' actions towards impact or beneficiaries
a number of	free	Low (overall and MP-C)	any	This n-gram introduces the notion of range and variety, with an implication of wider-than-stated reach or importance. However, it lacks specificity, which undermines the claim in a similar way as *one of the* (Table 40). **Use:** 12x in relation to dissemination; 4x with *impact*; 15x with people/beneficiaries; some more conceptual co-occurrences, e.g. *barriers, issues, themes, factors*. To the left: *highlighted, identified; resulted in, led to; collaboration with*

Specificity

In contrast to the previous label, which expands a claim, this label was applied to n-grams (including those listed in Table 42) that add focus, thereby enabling assessors to pinpoint the meaning of a claim and its relevance to assessment criteria.

Table 42: Examples of n-grams showing specificity

Entry	Degree of choice	Key in…	ICS section	Justification for category, based on use in corpus
after the	Restricted	High (overall)	2 and 4	Either referring to a point in the narrative of the impact, or to an external event that had an influence or serves as an anchor for the narrative
long term	Restricted	High (overall / MP-C)	any	MP-C: various words describing the impact: *effect, benefits, performance, investment* Overall: *benefits, effects, care* (a content word – significance preserved when discounting this use)
focuses on	Free	Low (overall / MP-C)	2, some 1	Zooming in on the main point of an ICS or research question
co uk	restricted	Low (overall)	4	part of URLs. 9x news outlet sites (e.g. THE, BBC, Guardian), 5x in one ICS the researcher's own (no longer maintained) website; others include Eventbrite and a review on Amazon

Looking across all four of the categories discussed above, many entries do not appear to convey persuasion in a general context. For example, none of them are included in the list of Boosters introduced by Hyland (2005a: 221-222). It is therefore possible that they may not even have had a persuasive function in all instances in this corpus. In most cases, it is not the wording itself but the function it points to that holds the persuasive power, such as adding credibility. This analysis therefore does not in itself provide a list of wordings that could boost the score of an ICS. Rather, it shows what kind of content (e.g. added value) or phrasing (e.g. specificity) could contribute to a convincing narrative, with examples of how this was done in high- and low-scoring ICS respectively in 2014.

6.3 Discussion and conclusion

In this chapter, I have explored the profile of lexical differences between high- and low-scoring ICS, to address research question 1c. Having extracted word combinations that appear significantly more often in the sub-corpora of high- or low-scoring ICS respectively (key n-grams), I analysed these according to function of use, level of editorial choice and potential persuasive connotations. From this, a number of observations can be made.

6.3.1 Themes

Having identified key n-grams and grouped them into themes, it appears that there are more word combinations that are key for low-scoring ICS than for high-scoring ICS. This finding does not support assertions that high-scoring ICS were "hyped up" (Hyland and Jiang 2023: 685) any more than low-scoring ICS. However, there seems to be a certain degree of convergence around themes. One study that correlated words with grade point averages of submissions introduces certain words typical for high-scoring ICS: *million, government, major, global* (Van Noorden 2015: 150). These correspond to the categories of Significance-scale, Reach-national and above, and generally Significance in my analysis, which appear more often in high-scoring ICS. The most typical words in low-scoring ICS were *conference, university, academic, project* in Van Noorden's study, corresponding to my categories of Pathway and Research. While my findings are consistent with those of Van Noorden (2015), they do not support Watermeyer and Hedgecoe's (2016: 7) recommendation to use words from the official guidance: there is no difference between the sub-corpora for the terms *reach** (e.g. reach, reached, reaching) and *significan** (e.g. significant, significance, significantly) in my corpus.

In addition, the tight page limit of the ICS template may have influenced word choice towards the use of more compressed language, as explained in section 3.1.3. This was even more relevant in ICS that were rated highly because they may have had more impact to report. Conversely, the writers of ICS that received low scores may have been less successful at adapting to a different register outside of disciplinary conventions and therefore may have used language that more closely resembles standard academic phrases such as those in the Discourse and Framing categories, which were therefore significantly more frequent, even typical, in that sub-corpus.

Overall, from looking at key n-grams, it cannot be inferred that certain functions did not appear, or did not appear much, in one sub-corpus. It can merely be observed that there were not many word combinations associated with a given function that appeared across ICS. Equally, there may be n-grams that appear frequently with certain functions but that appeared consistently across sub-corpora and therefore did not show up as statistically significant features of one sub-corpus or another in the investigation. A third caveat is illustrated by the category Beneficiaries. Only the low-scoring sub-corpus has key n-grams that refer to beneficiaries (e.g. *members of* certain groups or *the public*; *practitioners and*;

see above section 6.2.1.2), but this does not mean that high-scoring ICS do not specify their beneficiaries. On the contrary, as explained in section 5.3.2 reporting on the Thematic analysis (Reichard *et al.* 2020), they may mark them more specifically and/or have a greater variety of beneficiaries (as illustrated in e.g. Bonaccorsi *et al.* 2021), which therefore do not result in specific word strings appearing more frequently across several ICS.

6.3.2 Choice

Most of the "power words" identified by Van Noorden (2015: 150) are not freely available to all writers of ICS. This supports my finding that linguistic differences that stand out in high-scoring ICS represent differences in content, rather than differences in writing. Content-driven key n-grams were distributed differently in my high- and low-scoring sub-corpora: of 103 n-grams in High, nearly half (n=45, 43.7%) were content-driven, whereas this is true of only 20.4% (n=29) of the 142 entries in Low (see section 6.2.2.1, Table 29).

The kind of difference that content-related n-grams point to can be illustrated with a closer look at Pathways (see section 6.2.1.2, Table 26). A number of n-grams that are key in the low-scoring corpus contain the word *dissemination*, and this focus on dissemination supports the finding from the qualitative thematic analysis (section 5.3.2 above) that low-scoring ICS tended to focus more on pathways or routes than on impact. Moreover, there are several words and phrases specifying types of engagement, such as *the book* and *the event* that could be considered more informational or, at best, consultative, rather than collaborative (c.f. the IAP2 spectrum of public participation available at https://iap2.org.au/resources/spectrum/ and explained in Osborne 2022). Although the data does not allow direct inferences here, it is possible that this may represent a deeper epistemological position underpinning some ICS, where impact generation was seen as one-way knowledge or technology transfer. In this view of a knowledge deficit model (Jones *et al.* 2013), research findings are 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 equally possible, however, that there was no such (mis-)understanding about the role of research and communication in the REF context, and the appearance of that possibility, through the relatively higher use of research- and dissemination-related wordings in low-scoring ICS, stems more from a lack of support and understanding of the role of an ICS – or indeed from the lack of impact. Depending on the kind of mechanism at play in a given case, more knowledge about what is

required in an ICS, and how this can be worded, may or may not contribute to writing more successful ICS.

6.3.3 Persuasion

In a similar way that many of the identified key n-grams are not a matter of editorial choice, of those that are available to all (or many) writers, most cannot be construed as having a persuasive meaning. Out of 245 key n-grams, only 65 (26.5%) were assessed as carrying persuasion. Compared to the number of originally extracted n-grams (Table 24 above), this means that just 4% of all extracted n-grams are significantly different *and* persuasion-related. Instead, most of the differences are related to the emphasis of the narrative (e.g. Pathway), or content-related (e.g. Significance), or signposts (e.g. Discourse).

Even for those 65 n-grams, the persuasive meaning is not outwardly visible, but instead is specific to the ICS context. Many n-grams included as persuasive here would not be seen as such in other contexts, and their persuasive meaning is afforded to them through the assessment criteria. This supports the claim by Dontcheva-Navratilova (2020: 28) that repertoires of persuasion are genre-specific, and runs counter to Hyland and Jiang's (2023) approach of applying existing (if adapted) word lists and categories from one academic context to another. My new bottom-up method of investigating persuasive language across high- and low-scoring ICS also does not support their conclusion that ICS were "hyped" to an extent that the "usefulness and reliability" of assessment was in danger (2023: 2), or Brauer *et al.'*s assertion that researchers are required to "boast about their research" (2019: 66).

6.3.4 Conclusion

Drawing all three analyses together, the overarching finding of this chapter is that Specificity is a feature of high-scoring ICS. This can be seen, for example, in the fact that they anchor events more explicitly in certain times (section 6.2.1.1) and from other examples listed in Table 27 comparing different n-grams representing the same function in high- and low-scoring ICS. Looking at persuasive function, "specificity" is expressed in five n-grams key for high-scoring ICS compared to seven n-grams that are key in low-scoring ICS. However, this number includes discourse markers announcing focus, and two of the seven entries are part of URLs which should have been placed in Section 5 of the ICS template, rather than in the main text included in this corpus. Moreover, applying specificity is in principle an editorial choice – unless the specifics (quantitative or qualitative) are not known, which in turn makes it more difficult to craft a convincing ICS.

Conversely, ambiguity, vagueness and uncertainty are a feature of low-scoring ICS. 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 ICS template, 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 phrase occurred only in the Summary and the Details of the Impact sections (Sections 1 and 4), 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".

Overall, while a relatively small number of differences can be identified and there is a general trend of a difference between specific and vague language, there is a limited amount of difference between high- and low-scoring impact case studies that can be found using this method. This is because functions may be expressed in more specific ways and therefore not show in the lexical analysis by key n-gram. In the next chapter, I therefore report on an analysis of a principled selection of texts with full context. Following the focus on persuasion, I conducted an Appraisal analysis of evaluation on a sample of Sections 1 of impact case studies, bringing the unit of analysis from a (multi-)word level to the text level.

Chapter 7 Evaluative Language: Appraisal

Each impact case study is unique. The previous chapter sought to find commonalities across texts, and differences between groups of texts, on the basis of frequently occurring word combinations. However, given that the topics of ICS vary, it is expected that instances of evaluation may also be expressed differently, and therefore they may not appear sufficiently frequently to be detected by quantitative keyword measures. A manual Appraisal analysis addresses this problem.

The Appraisal system includes the three components listed and described in section 3.3: ATTITUDE, GRADUATION and ENGAGEMENT. In this study, I apply the GRADUATION subsystem in order to describe the use of covert evaluation, which may help to explain the perceived discrepancy between the characterisation of ICS as descriptive and persuasive. The aim of applying it in this respect is to establish the degree to which the persuasive elements of ICS are implicit, masking its nature of a persuasive register.

A second aim of this analysis is to enable more generalised findings about the functions that are used across the corpus. A lexical analysis could be used to generate lists of language items, which could in turn be (mis-)understood or (mis-)used as recommendations for particular words to include in a text in order to increase the chances for a high rating in a future REF exercise. A functional analysis partly safeguards against such "impact bingo" by emphasising the functions that appear more or less frequently in one or another sub-corpus, supplemented by a number of examples for how those functions are realised, rather than providing lists of, for example, "power words" similar to those described by Van Noorden (2015).

The Appraisal framework was introduced in section 3.3, and detailed information about the sample was provided in section 4.3.4. A first set of findings from analysing this sample was described in section 5.3.1, where I focussed on the type of content that can be found in Section 1 of ICS. This chapter goes into more detail about the process of coding for Appraisal (section 7.1), including the coding scheme, and includes findings from the analysis of tagged texts (section 7.2).

7.1 Method

Appraisal research often combines qualitative and quantitative methods. It is qualitative in that it requires manual coding of words or phrases in context and categorises the overt

(explicit) or underlying (implicit) message to the reader, interpreting it against the backdrop of a situational analysis. It can be quantitative because these coded instances can then be counted to highlight differences between groups of texts. In this section, I first explain how certain issues that commonly occur in Appraisal studies are treated in this analysis (section 7.1.1) and then introduce the coding scheme, including notes on how I adapted this from previous studies to fit the register under investigation based on the situational analysis (section 7.1.2). This is followed by an account of the process of tagging language items and increasing the level of confidence in any conclusions (section 7.1.3) and a description of the statistical analyses I applied to tags after coding (section 7.1.4).

Throughout the chapter, I use certain technical terms:

Category: the labels in the coding scheme, e.g. INTENSIFICATION, INVOKED. Mostly used in relation to the coding scheme.

Feature: similar to Category, but used more flexibly. For example, in the analysis, I sometimes combine more than one category into a feature.

Resource: a language item (one or more words) that is interpreted by the coder(s) as representing a certain category, that is, language that is used as a resource to express evaluation or persuasion within the GRADUATION framework.

Tag: Resources are tagged in the analysis tool (UAM Corpus Tool) with the Category label. These category tags can then be quantified within the tool.

7.1.1 Overall considerations

The opportunity to explore implicit language functions in texts through Appraisal analysis comes with a problem that is of course possible in all research: researcher bias. This is, however, exacerbated in studies of Appraisal and particularly the ATTITUDE sub-system. Since ATTITUDE can also be invoked rather than inscribed, there is a possibility that a coder bringing in their own bias might interpret a word or phrase as positive or negative that looks neutral to many other readers. Given the polarity of ATTITUDE, high inter-coder reliability is needed especially in studies that focus on this sub-system of the Appraisal framework. This is less problematic in the study of GRADUATION because here the emphasis is more on degree than polarity. Tupala (2019: 8) suggests that coder bias can at least be made visible, and at best be mitigated, by providing "valid and relevant background information" for decisions of why

a certain term may be seen as positive or negative, or generally evaluative, in a certain context. Her study of institutional policy documents, a register generally thought of as non-evaluative, holds important lessons for a study of REF ICS. In both registers, information is presented as factual, and evaluation is expected to be more covert than in many other registers such as product reviews (as studied e.g. in Biber and Zhang 2018), and this requires more interpretation by the coder. In order to maximise confidence in the conclusions, it is therefore important to try to minimise the amount of interpretation of the coding scheme that is needed to code a given instance of language.

Fuoli (2015: 12) tackles the issue of subjectivity that is present in all attempts to analyse evaluation in language. While it would be impossible to eliminate this subjectivity, he argues that it is possible to compensate for it to a certain extent through recording decisions and rationales. Especially a study like the present one that endeavours to extend qualitative manual tagging to quantitative analysis of tagged features should consider its stance towards subjectivity. Fuoli distinguishes between "individual" and "social" subjectivity (following Martin and White 2005: 62). Whereas individual subjectivity may be more problematic because it rests with the reading of one person, potentially randomly, social subjectivity can occur where a reader is aware of, and aligned with, the target context of the text. If such a target context favours a certain reading where a casual reader may not recognise an item as evaluative, but the target reader would recognise this because they are "socialised" into the context, this could be seen as "social subjectivity". For example, the number "2013" may look neutral to the casual reader, but the REF assessor in 2014 might read it as evidence that a claim is up to date or very new, depending on context, and therefore it may be marked in a certain way. Rather than being seen as problematic, such social subjectivity is to be expected, and it can be helpful to make it explicit. While coding texts, the social subjectivity should be discussed and applied consistently in the texts in that context. This also means that sometimes an expression is classified as a particular Appraisal resource in the literature but carries a different meaning in certain texts. For the study of REF ICS, there is high potential for social subjectivity because the REF context and criteria are rather specific. As Coder 1, my own background information and assumptions for the social subjectivity of ICS is shaped by a comprehensive literature review (chapter 2), extensive experience as reader and writer of such texts, and conversations with many other readers and writers. This experience is briefly described in section 4.2.3, and decisions influenced by

my role are recorded explicitly in the coding manual (Appendix F). Where there may be a discrepancy between the likely intention of the writer and the likely interpretation of the reader, the latter takes precedence because this perspective can be approximated by an outside reader-researcher, while the original writer's intention cannot be ascertained.

Before outlining the seven steps towards greater replicability, Fuoli (2015) discusses a number of potential issues in both identifying and classifying expressions of Appraisal. I will briefly outline these issues and explain how they are treated in this study.

- **Unitising**: It is sometimes ambiguous what should be treated as a *unit* of evaluative language where there are several expressions that could be tagged either together or separately. For example, "very wide recognition" could validly be tagged as one expression or as three (with "very", "wide" and "recognition" all potentially carrying their own evaluative meaning and receiving separate, potentially different, appraisal tags). This has implications especially where a quantitative analysis of tags is planned, because clearly a greater number of tags may skew the results if not applied consistently. For the ICS corpus, the smallest unit that is clearly evaluative in the context will be tagged, for three reasons:
    a. As Tupala (2019: 10) points out, longer units are more prone to ambiguous interpretation, and therefore smaller units are preferred here.
    b. Section 1 of an ICS, which is 100-120 words of advertising the best impacts, can be expected to be the most evaluation-dense part of these texts. Moreover, the focus of this study is on the *kind* of evaluation that is employed, and therefore a manual analysis needs to be granular, allowing for all separate occurrences to be recorded. This includes separating, and tagging as two items, instances of two evaluative items connected with a coordinating conjunction. The exception to this is where a series of experiential (i.e. non-evaluative) expressions is tagged together as REPETITION, which is a technique of INTENSIFYING a proposition and therefore of covert evaluation. Tupala (2019: 8) makes a different decision for similar reasons. The focus of her study is much more on the target of the evaluation, and therefore she counts the number of times a specific entity is evaluated, in conjunction with the polarity of the evaluation for the entity. By contrast, it can be assumed that most evaluation in ICS is positive, and the challenge is to quantify the relative

number of words or expressions used to refer to the various targets by high- and low-scoring ICS respectively. Therefore it is more valuable for the present study to count three items in a list as separate instances (as a deliberate choice by the writer to use three words rather than one), compared to Tupala counting this as one instance.

c. A final reason for coming to a different conclusion to Tupala (2019) is corpus size. The small corpus of impact case studies (7,500 words) can afford fine-grained tagging, which would be less feasible in Tupala's corpus (200,000 words).

- **Discontinuous evaluative expressions**: Sometimes evaluative expressions stretch across words that are not part of the same expression, and where the components of the evaluative expression are not by themselves evaluative and therefore cannot be treated as separate units. Not all programs provide an option to code these parts as belonging to the same evaluative expression without also including the intervening words. Coding an expression as one even though there are unrelated words in between can be seen as problematic because the additional words can confound the length of evaluative expressions if this is later quantified; however, in this study, it is the number rather than the length of evaluative expressions that is quantified. Fuoli (2015: 6) points to Carretero and Taboada (2014) who coded the whole expression, and I will follow their approach because of the restrictions in the UAM Corpus Tool.
- **Inscribed vs invoked evaluation**: This distinction is discussed in section 3.3, where I argue that implicitly invoking evaluation, rather than explicitly inscribing it, makes the stance of a text clear but not obvious. This is one way in which the discrepancy between the stated (factual) and the enacted (persuasive) purpose of the genre can be reconciled, that is, it may be that most of the evaluation is invoked rather than inscribed. In order to ensure that this is captured as a variable, I added the distinction between inscribed and invoked resources into the coding scheme, following Xu (2017).
- **Layering**: Occasionally, it can be ambiguous who or what is being evaluated and whether this is the entity to which the evaluative expression *grammatically* refers, or something that this entity *stands* for (e.g. metonymically: "Berlin" could be used to mean not the city but "the German government"). To record in a principled way what kind of entity is being evaluated, the ICS corpus is coded in a separate layer for

whether a phrase or sentence refers to impact / pathway / research / problem, as described above in section 5.3.1.

- **Multiple function**: A word or expression could be read as having, or being part of, more than one evaluative function. It is therefore necessary to decide whether more than one label can be applied in such cases, to allow for multiple functions to be recorded. In this study, only one label should be applied to each (part-)expression, because multiple tags can lead to double counting and skew the quantitative analysis if there are more labels than coded instances. Especially in a study focusing on GRADUATION, there is less scope for double function than in a study of ATTITUDE, where more fine-grained levels are less easily distinguished especially regarding Judgement (people-focused) and Appreciation (thing-focused) – the options for GRADUATION, especially in the Quantification branch, are clearer.
- **Irrealis**: An issue in coding for ATTITUDE, and to a certain extent ENGAGEMENT, is how to deal with instances irrealis such as "aim, intend, want". Contexts of this type could be seen as *less* positive than the same statement without qualification, and the ATTITUDE could be seen as more invoked than inscribed. In ICS, these instances should be coded less clearly positively because the register requires claims to be made in a retrospective voice and therefore instances of irrealis are not expected. Where they do occur, they are therefore marked. This study does not code for ATTITUDE, but such instances are tagged as GRADUATION:FOCUS:FULFILMENT:SOFTEN because they show that an action is not complete (see next section for how tags are structured).

Overall, the GRADUATION sub-system is not discussed in Fuoli's (2015: 7-12) section on problems with classification, as most issues are more relevant to the other systems.

7.1.2 Coding scheme

Details on the coding scheme, each of the categories, and the process of annotation are provided in the coding manual (see Appendix F). The texts were coded in two separate layers, first to identify and classify resources of GRADUATION, and then to segment the texts according to what a stretch of text (e.g. clause or sentence) refers to ("Target"). This is important because a GRADUATION resource will have a different effect depending on whether it appears in a segment describing the problem or describing the impact. This distinction may also be a reason why the sentiment analysis in Williams *et al.* (2023) did not show sentiment as a prediction for scores: potentially it was the case that negative sentiments in problem

statements may have balanced out positive sentiments in those text segments that describe the solution to those problems.

The “Target” layer was designed on the basis of the thematic analysis and includes the categories set out in Figure 15:

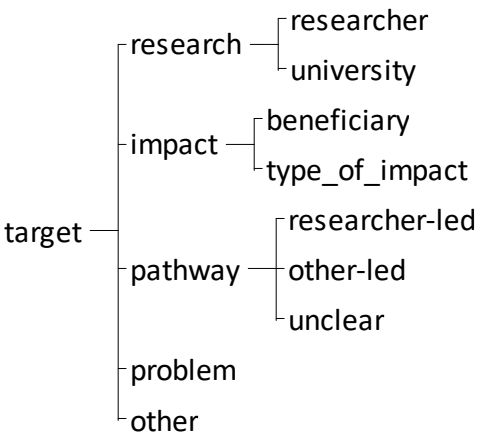


Figure 15: Coding scheme for the "Target" of each text segment

The segments of text (words, phrases, sentences) that have been classified as belonging to one of the “Target” categories can be used as sub-corpora within the UAM tool. In order to indirectly create such sub-corpora, each word in each text needs to be associated with exactly one tag. Analysis of Appraisal resources in each of the “Target” sub-corpora is presented in section 7.2.2.

The GRADUATION layer was created on the basis of existing coding schemes. The first starting point was the scheme by Martin and White (2005) in its iteration that is available on the Appraisal website (White 2003). The website provides an XML file that can be loaded into the UAM Corpus Tool. This initial scheme was adapted based mainly on Hood (2010) and Xu (2017), who had both extended the scheme and applied it to research articles. While Fuoli (2015: 8) clearly states that it is possible to add categories to the system if the existing categories are insufficient for “the specific discursive context under study”, this has to be justified through the situational analysis of the texts. The adaptations in this scheme were based on the situational analysis above (section 5.1) and the reader assumptions about the REF as discussed in the impact literature review. Consequently, for tagging the texts, it was especially important for coders to be sufficiently familiar with the REF context, as well as the Appraisal system. The final coding scheme is presented in Figure 16.

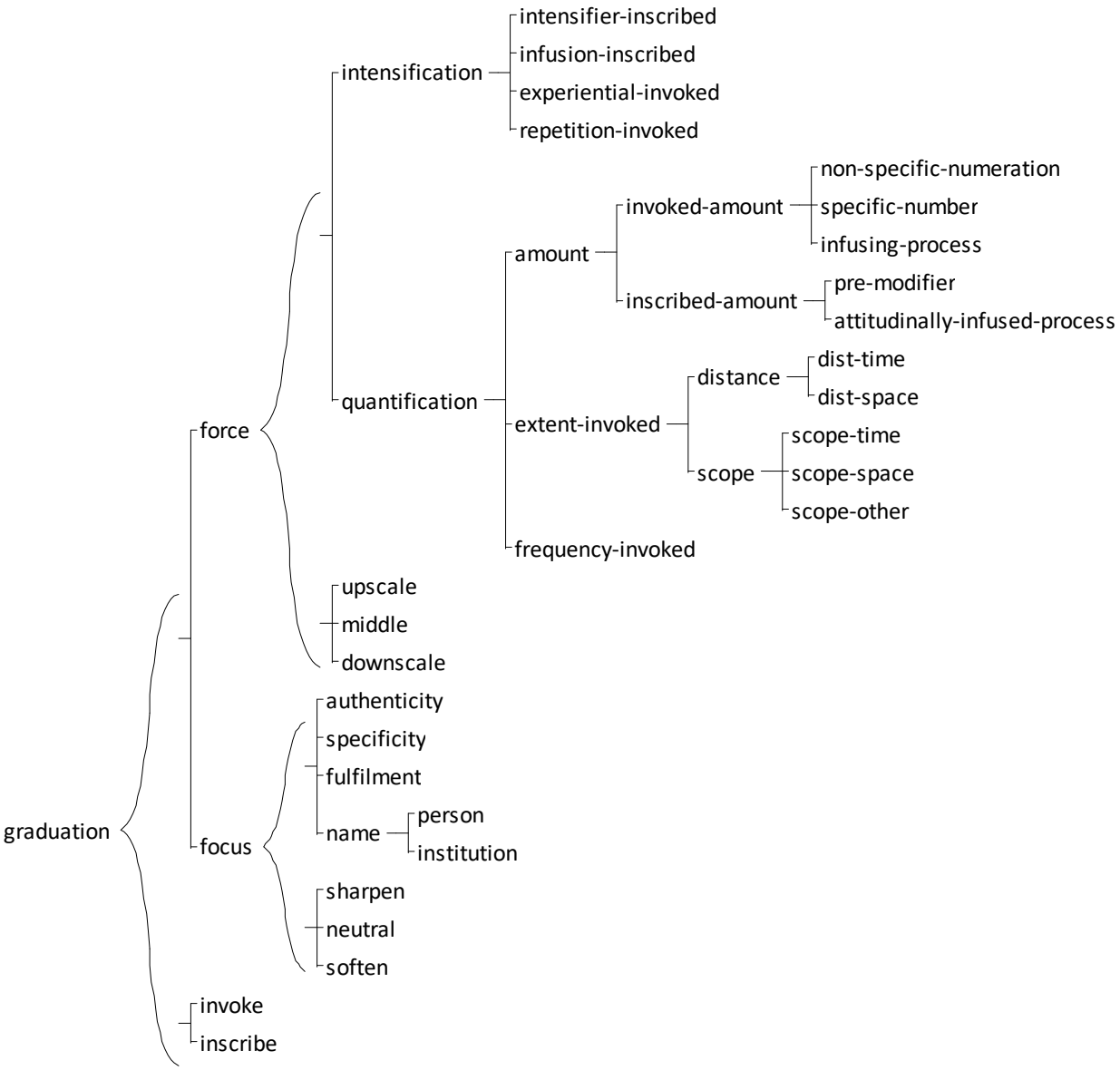


Figure 16: Coding scheme for GRADUATION

Each tag that is applied to a text segment includes three features: the type, e.g. SCOPE-SPACE; the direction, e.g. UPSCALE; and the visibility, that is, either INVOKED or INSCRIBED. The category type, that is, what kind of resource a text segment is seen as, is the most fine-grained type of distinction. In addition, as Martin and White (2005: 152, my emphasis) point out about the ATTITUDE sub-system, the upscaling of a resource can "construe the speaker/writer as maximally committed to the value position being advanced and hence as strongly aligning the *reader* into that value position". This is extremely relevant for the REF assessment context, and therefore I decided to keep the UPSCALE/DOWNSCALE and the SOFTEN/SHARPEN options separate in the coding scheme, rather than having them integrated into the type of GRADUATION, so that one of these labels could be applied as part of every tag. Finally, the

decision to code separately for each instance of GRADUATION whether it is INVOKED or INSCRIBED follows Xu (2017), who made this explicit and added it in her coding scheme.

7.1.3 Process of tagging

In order to facilitate the creation of a detailed coding manual and to assess and ensure reliability, this study used three pilot corpora before the main annotation stage. Each pilot corpus included one high- and one low-scoring Section 1 text each from Main Panels A, C and D. These were drawn from the same UoAs as the analysis corpus where there was choice, but from different UoAs where all texts from a UoA were needed to make up the main sample (see above section 4.3.4). This was necessary especially in Main Panels A/B and was considered an acceptable solution in order to keep the main sample completely separate from texts used for refining the tools for analysis. The pilot corpora were used as follows:

Pilot corpus 1 was coded on 21/09/2021 with the original draft of the coding manual, using a coding scheme based on Martin and White (2005) but heavily influenced by Hood (2010) and Xu (2017). This led to some refinement of the coding scheme, especially by adding more examples to the glosses within the software that are specific to this register. These examples were also added to the coding manual so that they could be easily located by searching the document. Some specific decisions were also recorded with explanations in the coding manual.

Pilot corpus 2 was then coded on 22/09/2021 with the refined coding manual, and several new principled decisions were applied. The main change from Pilot 1 was the decision to code only those instances where the writer had control over the word choice, and not apply tags of evaluation to words that were part of the research topic and technical terms in the problem statements. This is in line with the distinction between editorial choice and content-led n-grams introduced in section 6.1.5. Coding the second pilot corpus also provided further examples for the manual.

After coding Pilot corpus 2, a second coder was brought in to test the coding manual. He coded the first three texts of Pilot Corpus 2 on 03/10/2021 with the first coder in the room as he familiarised himself with the coding manual. This was both a training experience for the second coder to recognise instances of GRADUATION, and helped the first coder with the

development of the scheme, as decisions could be discussed. The scheme was adapted again after this coding round.

After a break of three months (28/12/2021), the final three texts of Pilot corpus 2 were coded together by both coders, with decisions discussed and recorded. This step also served as an intra-coder reliability check for coder 1 when comparing the new coding decisions to those from three months earlier on the same texts, because of the break between the first time of coding these texts and this re-engagement with them in conjunction with the second coder. Where a decision had been made differently, this could clearly be explained with reference to the adapted coding scheme and more refined decisions were arrived at during discussion.

Then both coders independently coded **Pilot Corpus 3** and compared their decisions. Inter-rater reliability was evaluated in relation to four main questions and calculated as follows:

1. *Unitisation: Did both coders identify the same words as instances of* GRADUATION*?*
   Out of 80 instances across both coders, 64 instances were unitised in the same way, that is, 80% agreement. Out of the 16 instances where there were differences, three were words or phrases where both coders saw a GRADUATION resource, but they chose a different length of text to include in the tag, resulting in a different code. For example, "reductions" was clearly seen as GRADUATION. By itself, it was tagged as DOWNSCALE (Coder 2), but with the context "reductions in crime", it was tagged as UPSCALE (Coder 1). After discussion, coders included the following note with the tag in the text: "upscale benefit through downscale of a problem - if it was just 'reductions', it would be 'downscale', but unitisation should include the bad thing so that this can be coded as 'upscale' which is what this should be understood as."
2. *For those instances where coders agreed on Unitisation: Did coders agree on an instance being coded as* INTENSIFICATION / QUANTIFICATION / AUTHENTICITY / SPECIFICITY / FULFILMENT *and their various sub-options, that is, the third level of delicacy (where applicable)?*
   Out of 64 instances, 57 were coded as the same type of GRADUATION (89% agreement). Out of 11 possible categories (with AMOUNT and EXTENT counted as one category each, rather than counting further levels of delicacy), there were discrepancies in five categories: INFUSION, EXPERIENTIAL (2x), INFUSING PROCESS, SPECIFICITY and FULFILMENT (2x). Coders discussed and agreed on a tag, clarifying instructions in

the manual, because some discrepancies had occurred from strict adherence to the manual by one coder who otherwise agreed with the other about the nature of the tag. For example, the manual had shown "new" as QUANTIFICATION:EXTENT:DISTANCE:TIME, but when applying the tag in a text, the coders struggled to decide whether this would be UPSCALE or DOWNSCALE and decided that INTENSIFICATION:EXPERIENTIAL would be more appropriate and moved "new" to that part of the coding manual.

3. *For those instances where coders agreed on the type of* GRADUATION*: Did coders agree on whether an instance was upscale/middle/downscale or sharpen/soften, respectively?*
   This was the case in 90% of tags, and for the other 10% coders agreed after discussion and noted the decision in the manual.
4. *For resources of* FOCUS*: Did coders agree on whether an instance was invoked or inscribed?* There was only one instance of discrepancy, resulting in 95% agreement. The decision was noted in the manual.

In Xu (2017), inter-rater reliability is reported as 85% following verbal coding of a pilot corpus by a second coder, but it is not reported in more detail how this 85% is arrived at – merely that the "classification" made was the same. It is therefore unclear whether this applies to agreement at all levels of delicacy, especially as Xu (2017: 121) adds that instances where "a different Appraisal category" had been assigned were discussed, which could imply ENGAGEMENT vs ATTITUDE vs GRADUATION, rather than further levels of delicacy.

Overall, this detailed process of ensuring and assessing reliability has resulted in a comprehensive and coherent coding manual to support decision making.

The main coding of the Appraisal layer was completed in the span of five days in January 2022. This enhanced the intra-rater reliability because it enabled the coder to frequently refer back to previous instances of a word that they had seen just days earlier. Occasionally this resulted in re-coding of decisions when an expression had been considered again in a different context, but in other instances, it was decided that a term was used with different meanings and should therefore correctly receive different tags in different contexts.

This also highlights that this manual annotation of texts according to the Appraisal system complements the more quantitative, automated components of research described in earlier chapters. Simple keyword searches could not have replaced this annotation process;

the close reading of the whole text segment (Section 1 of an ICS) helped to recognise the context of an expression and therefore place it in different functions according to context.

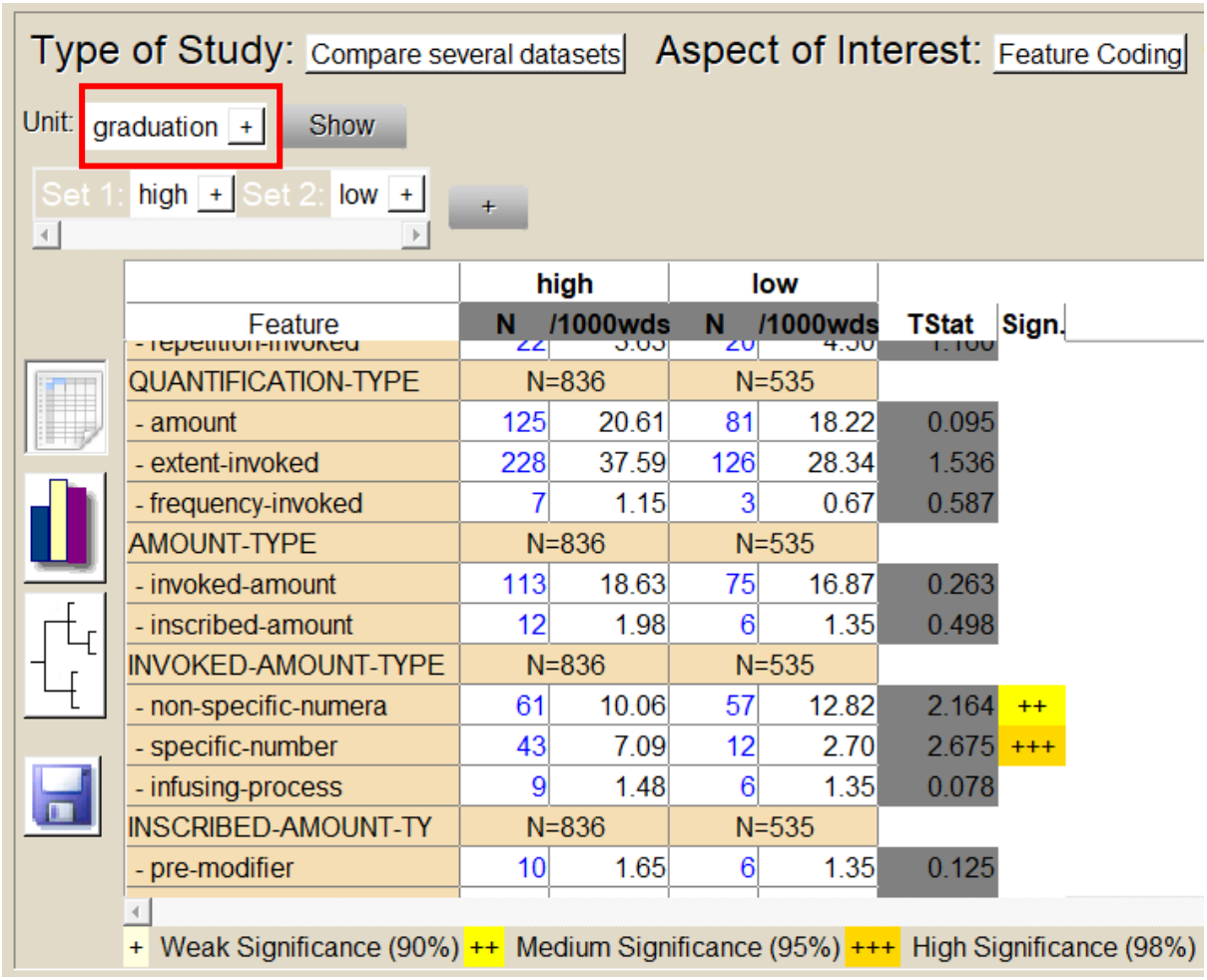
7.1.4 Statistical analysis

Quantitative analysis was approached from two angles. The first is a general exploration, which includes all features in the coding scheme and compares them across scoring brackets and Main Panels (section 7.2.1). Separately, as part of this general exploration, I determined the relative occurrence of features within each type of material ("Target", section 7.2.2). The second angle for quantitative analysis is a deeper exploration of, and reporting on, features that may be of specific interest based on the literature and earlier analyses that I conducted for this study (section 7.2.3). Of these two approaches, the first one is more data-driven and the second one more theoretically informed, although despite this distinction, it is important to acknowledge that even with the more data-driven approach, my own assumptions as a researcher influence the process. This relates, for example, to decisions around the sample, to the adaptations to the coding scheme that I made, and to the way that Appraisal resources were identified and tags applied, as described in the previous section (7.1.3).

For the general exploration, I further investigated those features that appeared significant in order to check whether other factors contributed to that significance. One check was a comparison by score within each Main Panel to eliminate the possibility that a finding that looks like a *score* difference is more likely to be a *panel* difference. A second factor could be instances where significance appeared on the basis of a very small number of occurrences, making the result unreliable when using a chi-square or log likelihood test. Instead, a *t*-test was used. A third potential issue are jagged profiles, where occurrences of features were dispersed across more texts in one sub-corpus but a similar number of occurrences in the other sub-corpus were concentrated in one or two texts. Here, I checked the distribution of hits across texts and whether overuse in one text or submission skewed the result. Concentration of features in one text was expected to happen due to the small number of texts in each sub-corpus (12 texts), but at the same time, the sample is as balanced as possible to avoid a situation where a certain style within a submission (i.e. non-independent texts) skewed the result (see section 4.3.4 for details on the sampling approach and Appendix C for a list of ICS included in this sample).

The general, data-driven, exploration has two components. The first component compares all GRADUATION features between the sub-corpora of high-scoring and low-scoring texts using

t-tests within the UAM Corpus Tool, independent of a feature's position in the coding scheme hierarchy and independent of the other features it could be combined with in the same tag (i.e. category type, direction and inscribed/invoked), as illustrated in Figure 17. This was done first for the corpus as a whole and then split by panel, resulting in four comparisons: Overall High vs Low, MP-AB High vs Low, MP-C High vs Low, MP-D High vs Low.[21] A similar analysis was done for comparisons between panels, regardless of scores, as explained later in this section (following Figure 19).



Feature	high N	high /1000wds	low N	low /1000wds	TStat	Sign.
QUANTIFICATION-TYPE	N=836		N=535			
- amount	125	20.61	81	18.22	0.095	
- extent-invoked	228	37.59	126	28.34	1.536	
- frequency-invoked	7	1.15	3	0.67	0.587	
AMOUNT-TYPE	N=836		N=535			
- invoked-amount	113	18.63	75	16.87	0.263	
- inscribed-amount	12	1.98	6	1.35	0.498	
INVOKED-AMOUNT-TYPE	N=836		N=535			
- non-specific-numera	61	10.06	57	12.82	2.164	++
- specific-number	43	7.09	12	2.70	2.675	+++
- infusing-process	9	1.48	6	1.35	0.078	
INSCRIBED-AMOUNT-TY	N=836		N=535			
- pre-modifier	10	1.65	6	1.35	0.125	

Figure 17: Screenshot of example results page in the UAM Corpus Tool – component 1

The second component serves to illustrate how certain features are split into other features, which enables a comparison of the distribution of all sub-features within a higher-level feature.[22] For example, for the part of the coding scheme shown in red in Figure 18

[21] As explained in section 4.3.4, the sample for this analysis includes a Science sub-corpus of combined Main Panels A and B, rather than either treating B separately or leaving it out (as was done for the sample used in chapter 6). Sub-corpora will be abbreviated as MP-AB, MP-C and MP-D for the remainder of this chapter.
[22] The technical term in SFL is "less delicate". However, the wording "higher-level feature" is used here because it is more intuitively accessible to more readers.

(AMOUNT:INVOKED-AMOUNT), within all AMOUNT tags, this approach enabled highlighting how many tags are SPECIFIC NUMBER compared to NON-SPECIFIC NUMERATION, and how they are distributed across directions (UPSCALE/MIDDLE/DOWNSCALE for FORCE features, SHARPEN/NEUTRAL/SOFTEN for FOCUS features), as shown in Figure 19. These splits were compared in the same way as the comparison of all GRADUATION features (i.e. High-Overall vs Low-Overall and then split by Main Panel, followed by comparisons between the panels).

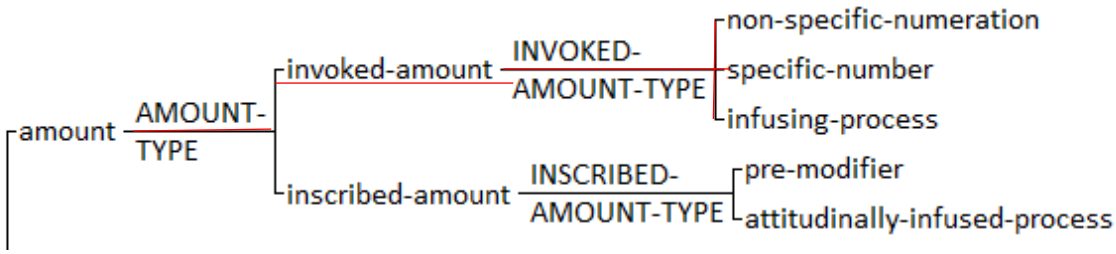
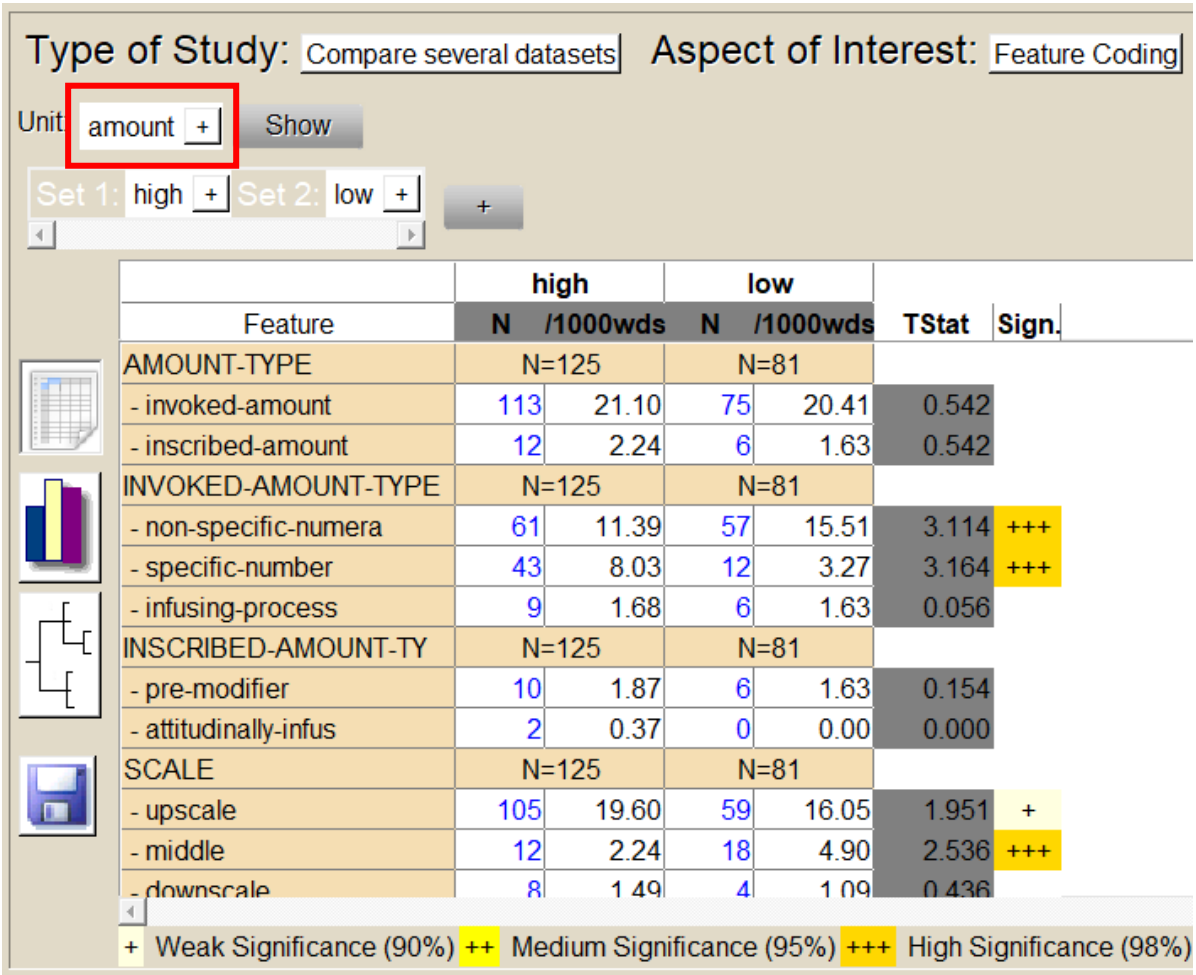


Figure 18: Screenshot of the example part of the GRADUATION coding scheme



Feature	high N	high /1000wds	low N	low /1000wds	TStat	Sign.
AMOUNT-TYPE	N=125		N=81			
- invoked-amount	113	21.10	75	20.41	0.542	
- inscribed-amount	12	2.24	6	1.63	0.542	
INVOKED-AMOUNT-TYPE	N=125		N=81			
- non-specific-numera	61	11.39	57	15.51	3.114	+++
- specific-number	43	8.03	12	3.27	3.164	+++
- infusing-process	9	1.68	6	1.63	0.056	
INSCRIBED-AMOUNT-TY	N=125		N=81			
- pre-modifier	10	1.87	6	1.63	0.154	
- attitudinally-infus	2	0.37	0	0.00	0.000	
SCALE	N=125		N=81			
- upscale	105	19.60	59	16.05	1.951	+
- middle	12	2.24	18	4.90	2.536	+++
- downscale	8	1.49	4	1.09	0.436	

Figure 19: Screenshot of example results page in the UAM Corpus Tool - component 2

The comparison of features across Main Panels (regardless of score) involved the same two components of analysis, namely the overall GRADUATION feature comparison and the more

fine-grained comparison of features within certain parts of the coding scheme. As the UAM Corpus Tool does not include inferential statistics for comparing three groups, the approach taken for this bird's eye view comparison was to isolate one disciplinary sub-corpus and compare it to the combined data of the other two sub-corpora (i.e. MP-AB vs combined MP-C and MP-D; MP-C vs combined MP-AB and MP-D, etc.). This allowed a rapid application of chi-square and *t*-tests to all features within the tool. In order to further probe the significance of any results in those comparisons, a one-way Analysis of Variance (ANOVA) was conducted for certain features. This involved exporting data for each relevant feature, where the number of occurrences was treated as an independent variable and the Main Panel as the dependent variable. The ANOVA was then conducted using Lancaster Stats Tools (Brezina 2018).

Because the analysis with ANOVA required these additional steps, it was only applied to those features that met certain criteria. The main indicator for further investigation was if a feature appeared significant when testing MP-AB versus combined MP-C and MP-D (or in any of the other comparisons), either by chi-square or by *t*-test. In addition, to ensure that no features were excluded in the secondary analysis (ANOVA for three sub-corpora) that may have been borderline significant in the first set of analyses (chi-square and *t*-test comparisons of one sub-corpus against the two others), the following questions were applied to all GRADUATION features to decide whether an ANOVA might be warranted for a more fine-grained comparison:

a) Is this feature at too high a level in the system, combining many different features and therefore a comparison would not be meaningful? For example, the two highest-level features FORCE and FOCUS were not considered.
b) Has this tag been applied often enough to make the comparison meaningful? For example, FREQUENCY-INVOKED only appears in 7 (out of 76) texts overall, spread across 4 out of 6 sub-corpora, so an ANOVA would not be meaningful due to the high number of texts with 0 instances, and because the few texts containing such FREQUENCY resources may be outliers.
c) Are "sister" features included? For example, the "direction" features of MIDDLE and DOWNSCALE were subjected to an ANOVA because there was some significance in other tests, and therefore the third feature in the set, UPSCALE, was also tested. For the same reason, AUTHENTICITY and SPECIFICITY were tested in their function as the other

FOCUS resources because FULFILMENT appeared significant. However, some "sister" features were not tested because there was insufficient data (see criterion b) or because the reliability of tagging is low (e.g. REPETITION, as discussed in section 7.1.1 under Unitisation).

Results from these comparisons are integrated into the results description below (sections 7.2.1 and 7.2.2), where relevant.

In addition to these more data-driven explorations, the tagged texts were also interrogated from a more theoretical perspective (second angle). This was done partly to test assumptions expressed in the impact literature (see section 2.2 above) and partly based on my own analyses described in earlier chapters, including the situational analysis in section 5.1. Several areas for further investigation arose from these two sources. These areas were explored using the same tests and tools described in the current section, but in a more targeted, deep-dive way, as opposed to the overarching principled way described above which was designed to capture emerging insights. Findings are described in section 7.2.3.

7.2 Results

For the interpretation of the results in this section, it is important to keep in mind that the corpus for this analysis consists only of Section 1 ("Summary of the impact") from each ICS, rather than the full texts. The ICS template gives an "indicative word limit" for Section 1 of 100 words, though most ICS exceeded this; on average, a Section 1 in a high-scoring ICS was 139 words long, compared with 103 words in an average Section 1 of a low-scoring ICS. Within this tight word limit, it can be expected that not many additional words are invested to indicate evaluation, and there is limited scope for content descriptions. At the same time, Section 1 is likely where an assessor forms their first impression, and therefore these texts should portray the ICS in the best possible light in the assessment context. Part of this is to give a clear indication for what to expect. Therefore, the nature of the evaluative tone in Section 1, where present, is treated in this study as representative of the ICS as a whole.

As explained above (section 7.1.4), the analysis is approached from two angles: a bird's eye view of the data and a direct interrogation based on pre-existing questions. Section 7.2.1 describes data from the first of these two approaches, which is more data-driven in order to reduce the possibility that potentially interesting findings are missed due to researcher bias. It is followed by section 7.2.2 where the same approach is applied to an analysis of the

"Target" of the material (see also section 5.3.1 for more on this process). The final section of this chapter (section 7.2.3) describes findings from the second angle, addressing specific questions.

7.2.1 Comparison by score and by Main Panel

The results of the first component of analysis, comparing all features across sub-corpora as independent, rather than in relation to each other, show that for the overwhelming majority of features, there is no significant difference between the high- and low-scoring sub-corpora. The comparison includes 49 features, as shown in the coding scheme (see section 7.1.2 Figure 16). Of these, DISTANCE-SPACE was never applied in the corpus, bringing the number of possible comparisons down to 48. Because DISTANCE is split into only two further features, namely DISTANCE-SPACE and DISTANCE-TIME, in this corpus it includes exactly the same tags as DISTANCE-TIME. Therefore, DISTANCE was also excluded from the comparison to avoid double counting of what essentially is only one feature (i.e. DISTANCE-TIME). This leaves a total of 47 feature options, which include the DIRECTION features that are applied to all resources, as well as choices at the intermediate level of delicacy (e.g. INTENSIFICATION vs QUANTIFICATION, which are part of FORCE and are themselves split further into component parts). The majority of these do not occur in significantly different numbers. In a comparison across scores (regardless of MP), only the features shown in Table 43 appear in statistically significantly different numbers.

Table 43: Comparison of GRADUATION features across High-Overall vs Low-Overall (statistical significance)

Feature	**Chi-square**	***t*-test**	**Level of significance**	**Overuse[23] in sub-corpus**
SPECIFIC-NUMBER	8.219	2.675	high (p<0.02)	High
DISTANCE-TIME	6.217	2.497	high (p<0.02)	High
SOFTEN	15.724	3.985	high (p<0.02)	Low
NON-SPECIFIC NUMERATION	5.171	2.164	medium (p<0.05)	Low
UPSCALE	4.736	2.178	medium (p<0.05)	High
DOWNSCALE	3.845	-	medium (p<0.05)/ weak (p<0.1)	Low

The table shows that, out of 47 comparisons across the high- and low-scoring sub-corpora, only six are significantly different, whereas 41 are not. The chi-square tests compare overall

[23] The terms "overuse" and "underuse" are used in a purely descriptive way here as "relative overuse compared to the other sub-corpus".

occurrence in the corpus without taking into account that the dataset is made up of different texts. In order to take dispersion across the different texts into account, the same comparisons were also made using a *t*-test. The level of significance was the same for most features, with the exception of DOWNSCALE. For the difference in the DOWNSCALE feature, the *t*-test within the UAM tool indicated "weak significance", which is the label the tool applies to $p<0.1$. This level of significance was generally discounted in this study, because the number of tags in each comparison is low in this small corpus, and the lower the number, the less reliable a significance test is at higher thresholds (Rayson *et al.* 2004: 8). An explanation for this low level of significance for DOWNSCALE with a *t*-test, despite the apparent difference in absolute numbers across the corpus (4 instances in high-scoring texts compared to 24 instances in the low-scoring sub-corpus), is the distribution of DOWNSCALE tags within the low-scoring sub-corpus: 17 out of the 24 instances are in only two ICS, which intersperse the text with the names of English towns (sub-national geographical references were tagged as DOWNSCALE, and each separate entity such as a town name receives a separate tag, as detailed in the coding manual in Appendix F). Therefore, only 5 out of 6 comparisons in Table 43 should be seen as showing a statistically significant difference, whereas for 42 of the overall 47 features, there is no significant difference, indicating homogeneity across the corpus.

In the comparison across Main Panels, 23 ANOVA tests were run, selected as described above in section 7.1.4. Of these, 17 showed no significant difference (EXPERIENTIAL-INVOKED, QUANTIFICATION, AMOUNT, NON-SPECIFIC NUMERATION, DISTANCE-TIME, SCOPE-TIME, UPSCALE, DOWNSCALE, SOFTEN, AUTHENTICITY, FULFILMENT [OVERALL, SHARPEN, NEUTRAL, SOFTEN], NAME, PERSON, INSTITUTION), while six showed a significant result in one Main Panel compared to at least one other. This is split as shown in Table 44.

Table 44: Overview of significant results from ANOVA comparing Main Panel sub-corpora

Feature	**Overused in…**	**Compared to…**	***p***	**Effect size**
INTENSIFICATION	D	AB	<0.05	small (ω = 0.275)
SPECIFIC-NUMBER	AB	C and D separately	<0.001	medium (ω = 0.396)
SCOPE-SPACE	C	AB	<0.01	medium (ω = 0.343)
SCOPE-OTHER	D	AB C	<0.001 <0.01	medium (ω = 0.409)
MIDDLE	C	D	<0.05	small (ω = 0.279)
SPECIFICITY	C	D	<0.05	small (ω = 0.264)

Of these differences, the feature SPECIFICITY (final row of Table 44) warrants a closer look, because it makes a difference whether something is SHARPENED or SOFTENED. The tag was applied where an expression modified its context to be either more (SHARPEN) or less (SOFTEN) specific. As has been shown throughout the thesis (e.g. sections 5.3.2 and 6.2.1.1), the degree of specificity and explicitness is an area where differences between high- and low-scoring ICS can be found. Comparisons between scoring brackets within each Main Panel show the following: In MP-AB, there are more instances of SHARPENING SPECIFICITY than SOFTENING in both the high- and low-scoring sub-corpora (10 vs 6 instances in High, 8 vs 3 in Low). In MP-C, SPECIFICITY is used notably more to SHARPEN (80% each) in both the high- and low-scoring sub-corpora. In MP-D, in the high-scoring sub-corpus, there is an even split between SOFTENING and SHARPENING, whereas the low-scoring sub-corpus only includes instances of SHARPENING SPECIFICITY, with no SOFTENING tags. This means that, while there are disciplinary differences between the Main Panels in the degree to which SHARPEN takes precedence, there is no indication that high-scoring sub-corpora provide more SHARPENED SPECIFICITY than low-scoring ones. The theme of SPECIFICITY will be revisited below in section 7.2.3.3.

The tables above give an overview of the differences that can be found between sub-corpora when the same comparisons are applied consistently. Conversely, there are features where differences might have been expected based on the literature and previous analyses, but no significant difference is found. Questions based on pre-existing assumptions will be discussed below in Section 7.2.3, but here I want to highlight some features where it is notable that there seems to be no statistically significant difference:

- **QUANTIFICATION**: An underuse in Main Panel D compared to Main Panels AB was expected based on the assumption that science writing includes more reports on quantitative analysis than humanities prose does, but there is virtually no difference between Main Panels AB and D. This highlights that whereas there are differences in the *means* that these panels use to express quantification (see below at FORCE:QUANTIFICATION), the overall *number* of instances seems to be broadly similar.
- **NAME**: This feature was added into the coding scheme during piloting under FOCUS in order to detect differences between sub-corpora in the use of references to the submitting institution or the lead researchers. In Main Panels C and D, there is no significant difference in such mentions between high- and low-scoring ICS. This

indicates that both high- and low-scoring ICS included the names of people or institutions and supports the view that "name-dropping" in itself may not constitute either an advantage or a disadvantage. There is a slight difference in Main Panels AB, in that low-scoring ICS are more likely to name a person and high-scoring ICS are more likely to name an institution. However, the overall numbers are small: in the low-scoring sub-corpus for Main Panels AB, it is two ICS from different submissions that name a person, compared with two ICS from the same submission naming an institution. As this only appears as a difference at UAM's "low significance" level ($p<0.1$), it can be said that in Main Panels AB there is no significant difference either. Comparing the panels against each other, there are significantly more mentions of a researcher's name in Main Panel D than in the other Main Panels; mentions of institutions are nearly equal. This is split into mentions in 6 out of 12 high-scoring and 3 out of 12 low-scoring texts, so the numbers may be too small to be actually meaningful. The number of ICS featuring a name is nearly the same across Main Panels, but whereas names are normally mentioned only once or twice in Main Panels AB and C, they often appear several times in the same text in Main Panel D.

- **SCOPE-SPACE:** This feature was used to identify geographical reach. While there are differences in this feature in Main Panels C and D (see below, FORCE:QUANTIFICATION), in Main Panels AB there is more homogeneity in distribution across scores. Overall, more geographical references were made in high-scoring than in low-scoring Science texts (MP-AB), but the difference is not statistically significant. In both sub-corpora (high and low in MP-AB), international references dominate (SCOPE-SPACE:UPSCALE), with national references not far behind (SCOPE-SPACE:MIDDLE) and no references to sub-national scale (SCOPE-SPACE:DOWNSCALE). Clearly the inclusion of claims for international reach, or at least international relevance, in some low-scoring ICS has not influenced the score in Main Panels AB.

Table 43 above shows the features that emerge as statistically significant in the most top-level comparison: High-Overall vs Low-Overall. In the remainder of this section, these findings are summarised not by individual search in the UAM tool, but by feature or feature group, which enables them to be presented in a more contextualised way. This discussion includes results from all the analyses of GRADUATION described so far, including comparisons across scores and Main Panels.

1) FORCE:INTENSIFICATION

INTENSIFICATION is split into four different kinds of resources: INTENSIFIER (e.g. *very*), INFUSION (e.g. *essential*), EXPERIENTIAL (e.g. *reinforce*) and REPETITION. Of these, EXPERIENTIAL was used most often (just over half of all instances; this tag was applied 120 times, compared to 220 tags of INTENSIFICATION overall). This was the tag applied to terms that added an evaluative element to otherwise non-evaluative expressions, which are to be expected in overtly descriptive texts such as ICS. MP-D has the largest number of INTENSIFICATION resources overall (n=90), a difference that is significant compared to MP-AB (ANOVA, $p<0.05$, see Table 44 above). One possible explanation is that there may be a tendency in the Humanities to provide more narrative ways to express evaluation, or to use a wider repertoire for grading meaning.

REPETITION is underused in MP-AB compared to both MP-C and MP-D. It is used in nearly half of all texts in MP-C (10/28) and MP-D (10/24) but only in three high- and two low-scoring MP-AB texts (5/24). One explanation could be that the style of science writing is less prosaic. An alternative explanation could be that, generally, both impacts and pathways from science research are more linear and direct than those in the social sciences and humanities, where there may be greater variation (and hence opportunity for listing more partner organisations, for example).

A final difference is that in MP-C, the relative number of instances of INSCRIBED evaluation is higher than in the other MPs compared to the number of instances of INVOKED evaluation. Within MP-C, there is a significant difference ($p<0.05$, $t=2.102$) between high- and low-scoring texts: nearly all INTENSIFICATION resources in low-scoring texts are INVOKED, whereas high-scoring texts also include INSCRIBED resources, such as “substantially”, “official”, “radical”.

2) FORCE:QUANTIFICATION

QUANTIFICATION was identified as an important feature both in the literature and in my own previous analyses, so most of the results and discussion, especially on SPECIFIC-NUMBER and NON-SPECIFIC NUMERATION, can be found below in the section 7.2.3.3. However, there are some additional observations arising from the overall feature comparisons and that do not address pre-existing questions, both in the AMOUNT and EXTENT branches of the scheme as represented in Figure 20. These are reported here.

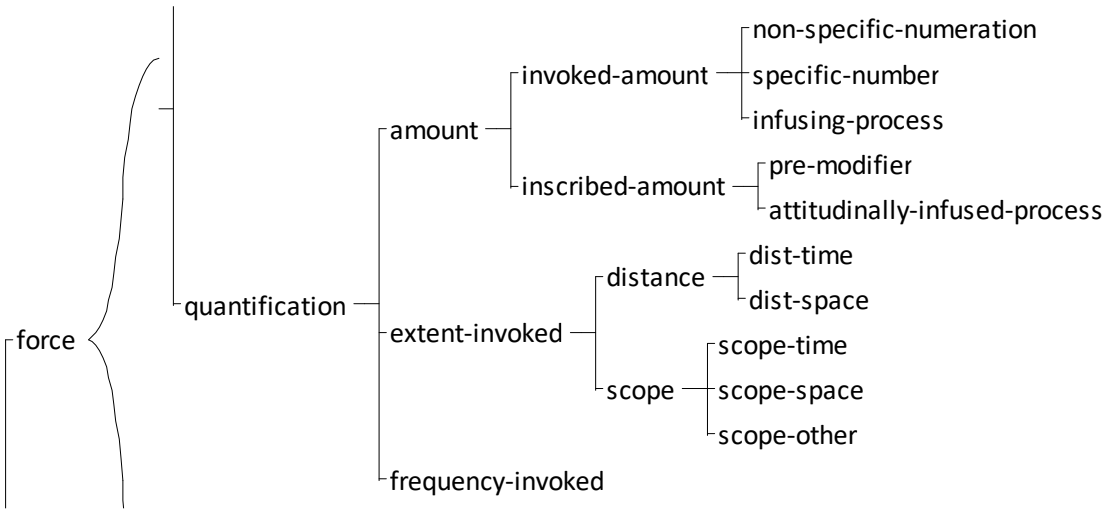


Figure 20: QUANTIFICATION branch of the GRADUATION coding scheme

There is some slight underuse for QUANTIFICATION overall in MP-D when compared to all other MPs combined ($p<0.05$); in an ANOVA comparing the three MP sub-corpora against each other, this underuse is no longer statistically significant, though. Whereas there is some relative overuse of SPECIFIC-NUMBER in MP-AB and of SCOPE-SPACE in MP-C (see ANOVA results in Table 44 above), there does not seem to be a QUANTIFICATION resource that is particularly frequent in MP-D to bring it to the same level of QUANTIFICATION. An ANOVA shows a significant overuse of SCOPE-OTHER in MP-D ($p<0.01$), but with overall smaller numbers – see below (SCOPE-OTHER) for further explanation. One small observation in the AMOUNT section is that INFUSING-PROCESS, that is, using a verb to indicate scaling, is used only twice in the whole MP-AB sample ("accelerate", "increasing"), compared to 7 and 6 times in MP-C and MP-D respectively (mostly "reduce"/"raise"). In these MPs, SPECIFIC-NUMBER (MP-C: 7, MP-D: 9) and INFUSING-PROCESS are used roughly the same number of times, but in MP-AB, SPECIFIC-NUMBER is used much more often (n=39 compared to n=2 for INFUSING-PROCESS).

Turning to EXTENT, there are four key differences within and across sub-corpora:

1. In the MP-AB sub-corpus, there is a difference between scores: EXTENT resources overall, which include DISTANCE and SCOPE and therefore give temporal and spatial information, are significantly more frequent in high-scoring Science texts than low-scoring ones ($p<0.05$, $t=1.996$).
2. In MP-D, there is a score difference for EXTENT:SCOPE-SPACE:MIDDLE, denoting claims to national reach. Such claims appear in half of all high-scoring texts (6/12) but only one low-scoring text (1/12). This shows that high-scoring

Humanities texts were more likely to make explicit claims to national reach than low-scoring Humanities texts.

3. Disciplinary differences can be found in SCOPE-SPACE between the sub-corpora MP-AB (underuse) and MP-C (overuse) at p<0.02. The relative underuse in MP-AB (n=47) indicates that these texts are less specific in showing geographical reach. One explanation could be that their claims have less need of being geographically situated explicitly because the SCOPE may be more obvious from the type of impact. MP-C (n=98) and MP-D (n=58) include many more expressions that can be seen as indicating geographical SCOPE.
4. Looking at the number of texts rather than the number of instances, a higher proportion of MP-AB texts (11/12 High and 9/12 Low) use some geographical expression, compared to a lower proportion (19/26 High and 18/26 Low) in MP-C and MP-D combined, but ICS in the Social Sciences and Humanities that do use geographical references tend to use a higher number of place-based expressions in any one Section 1 text, either through repetition or by listing different entities. Especially MP-C texts use significantly more explicit claims of geographical reach than MP-AB (*t*=4.245) and MP-D (*t*=3.085), which may be connected to the need to specify the site of interventions.

DISTANCE-TIME is a resource that overall appears significantly more often in high-scoring texts (p<0.02, *t*=2.497). In MP-C, the type of time reference that stands out is MIDDLE, appearing nine times in high-scoring and only once in low-scoring texts. These instances are either dates that are close to the submission date (i.e. not emphasising how long ago an impact or pathway activity started) or non-specific sequencing words like *previous* or *subsequently*, making timelines explicit. It is interesting that these instances are more frequent in high-scoring than in low-scoring texts because it seems to show that short-term or recent impacts could also be rated highly. Alternatively, this could be a further indication that explicit information enables assessors to gain a clearer understanding of the impact claims and therefore increases the chance that credit can be awarded.

Within MP-D, all time references tagged as DISTANCE-TIME:UPSCALE appear in the high-scoring texts (e.g. “was founded in 2009”, “since 2008”, “before the research began”). Conversely, all three time references tagged as DOWNSCALE are found in the low-scoring corpus. This means that none of the low-scoring Humanities texts give a start date or other time

reference to indicate change from a baseline, making it harder for assessors to evaluate impact claims. The two texts that do include time references use a vague or future-facing expression, which do not appear at all in high-scoring texts: “future”, “early”/“initial” (in the same ICS) give an impression of preliminarity rather than the required retrospective claim. (Note that the word “future” appears three times overall in high-scoring texts, but in different, more convincing contexts, rather than as time references: e.g. “debates about the future of copyright” in UoA36 LSE *Citizen*.)

DISTANCE-TIME also appears significantly more often in MP-AB compared to combined MP-C and MP-D ($p<0.02$, $t=2.386$). The same is true for SCOPE-TIME ($p<0.02$, $t=3.001$), where MP-AB have 26 instances compared to a combined 24 instances in MP-C and MP-D. This is spread across high- and low-scoring texts, indicating that Science writing overall is more explicit in providing start dates or timelines.

SCOPE-OTHER is a category that was introduced to accommodate expressions indicating scope that was neither temporal nor clearly spatial, often introducing examples. The most frequently tagged word was “both” (n=16), followed by “includ*” (i.e. any form of the verb “include”; n=11), “such as” (n=4) and “across” (n=2), as well as similar expressions that appeared only once each. Half of all instances appear in MP-D (n=25), with a clear overuse compared to MP-AB (n=9) but also MP-C (n=13). The instances in MP-D are mostly "both" but also some other narrative devices to indicate range, which contrasts with the EXTENT categories that stand out in other MPs: geographical range (MP-C) and specific numbers (MP-AB). This may be a sign that in impact arising from Humanities research, there is more narrative or qualitative extent that needs specifying, compared to impact from the (Social) Sciences, where quantitative extent can be expressed with specific numbers. In MP-AB, SCOPE-OTHER appears mainly in High (8 out of 9 instances) and to introduce a list of examples to indicate that what is included is not the full picture (5 out of 9 instances). Of the instances in the other Main Panels (n=38), nearly half are "both" and only 11 tags are of the kind seen in the Sciences ("includ*", "such as"). The remaining instances are more qualitative descriptions of scope, such as "cross-sector" and "in and outside of [specific] communities".

3) DIRECTION

In addition to the type of GRADUATION, each instance also receives a tag for the direction. Resources of FORCE can be tagged as UPSCALE, MIDDLE or DOWNSCALE, whereas resources of

FOCUS can be tagged as SHARPEN, NEUTRAL or SOFTEN. Starting with FORCE, there is an overall overuse of UPSCALE resources in the high-scoring texts ($p<0.05$, $t=2.178$). This overall difference is not borne out in comparisons within each Main Panel, but differences can be seen in the use of this tag. For example, UPSCALE of INVOKED-AMOUNT in the low-scoring sample (n=60) is done relatively more frequently through NON-SPECIFIC NUMERATION (n=44) than through SPECIFIC-NUMBER (n=3), compared to High (n=93) where there are many more instances of SPECIFIC-NUMBER (n=34) and not quite as large a gap between that and NON-SPECIFIC NUMERATION (n=50).

In comparison across Main Panels, a difference can be seen in MP-D compared to all others combined: UPSCALE is split fairly evenly into resources of INTENSIFICATION (44.44%) and QUANTIFICATION (55.56%), compared to a clear skew towards upscaling through QUANTIFICATION (MP-AB: 69.23%, MP-C: 65.64%) over INTENSIFICATION (MP-AB: 30.77%, MP-C: 34.36%) in the other sub-corpora.

Conversely, the difference between high- and low-scoring texts in DOWNSCALE, which is due to overuse in only two texts and therefore should not be treated as of real significance, does become more interesting in more fine-grained comparisons. DOWNSCALE:SCOPE-SPACE does not appear at all in MP-AB, indicating that no sub-national impact is claimed in that sample. Resources of DOWNSCALE appear mostly in high-scoring ICS (7/12 ICS from 4 submissions, n=19) and are mostly connected to SCOPE-TIME (n=5, e.g. “rapidly”) or DISTANCE-TIME (n=5, e.g. “2013”). By contrast, in MP-D there are 28 instances of DOWNSCALE in eight texts (2x High, 6x Low, all from different universities), with the vast majority (23) being connected to SCOPE-SPACE. In both high- and low-scoring MP-D ICS, there is therefore a greater likelihood of claiming sub-national impact than in MP-AB. The only DISTANCE-TIME downscaling instances in MP-D appear in low-scoring ICS and are “future”, “early” and “initial”, which contribute to the impression that the impacts being described here are preliminary at best. No clear trend in either direction can be seen in MP-C, where DOWNSCALE is generally used less than in the other Main Panels.

As indicated in Table 43, SOFTEN is used significantly more often in low-scoring texts across the corpus. However, there are some interesting disciplinary differences. In both Main Panels C ($p<0.02$, $t=2.441$) and D ($p<0.05$, $t=2.334$) independently, SOFTEN is applied more often in low-scoring texts than in high-scoring ones. In Main Panel D, all instances where the FOCUS resource of SPECIFICITY is tagged as SOFTEN occur in high-scoring texts: twice each in the

"pathway" and "problem" segments (any SPECIFICITY resources in the "Research" and "Impact" segments are SHARPEN). High-scoring texts therefore apply more generalisations in these two types of material (see section 7.2.2 for more detail on GRADUATION resources in different types of material). Perhaps more meaningfully, the FOCUS resource of FULFILMENT is tagged as SOFTEN 13x in high-scoring texts and 28x in low-scoring texts, which is a clear overuse in the latter. These tags are distributed fairly evenly over 8 ICS (of 12), compared to 7 (of 12) in the high-scoring sub-corpus.

By contrast, there are significantly fewer instances of SOFTEN in MP-AB than in combined MP-C and MP-D ($p<0.02$, $t=2.37$). Moreover, in contrast to the score difference in those panels, in MP-AB the instances that do occur are nearly balanced between the high- and low-scoring sub-corpora. This supports the impression that Science writing is more factual, with fewer expressions that generalise (SPECIFICITY:SOFTEN) or indicate preliminarity (FULFILMENT:SOFTEN). A potential alternative explanation could be the possibility that there are more science departments at well-resourced universities, or that these departments are more long-standing and therefore have the opportunity to write narratives that span longer time periods of research, and therefore there were few universities in Main Panels AB with really preliminary or vague impacts. However, unlike in Sample A, which has a skew where scores and Main Panels are aligned (i.e. fewer "Low" texts in MP-AB compared to MP-C – see section 4.3.2 above), the Appraisal sample is completely balanced across scores within each disciplinary sub-corpus. Therefore, the difference in writing cannot be explained by sampling bias and is more likely to be tied to disciplinary preferences.

7.2.2 Comparison by type of content

So far, findings have mostly been reported regardless of the type of material that is being evaluated. As described in sections 4.3.4 and 5.3, the texts were also tagged according to the kinds of material that evaluative expressions refer to ("Target"), split into Research, Impact, Pathway, Problem and Other as set out in Figure 15 (see section 7.1.2). Alongside Score (4* or 1*/2*) and Main Panel (AB, C or D), this is another important variable for analysing the distribution of evaluation, because of the possibility that any apparent difference between scores or disciplines may actually be related to the type of material that is emphasised in certain texts. Since all words in all texts were part of exactly one "Target" tag, it was possible to treat the material tagged with each "Target" tag as a sub-corpus, and the two components of the bird's eye analysis described in section 7.1.4 were applied to these sub-

corpora. That is, comparisons were made between all the Research-related material in the high-scoring texts ("High") and all the Research-related material in the low scoring texts ("Low"), and similarly with all Impact- and Pathway-related material.

Comparing material in High and Low that refers to either of these top-level "Target" categories shows that nearly all differences between High and Low disappear. That is, GRADUATION features in material relating to Research in High vs Research in Low, Impact in High vs Impact in Low etc. are virtually equally distributed among these target-related sub-corpora. As Table 45 illustrates, only six GRADUATION features occur with significantly different frequencies in the different types of "Target" material: NON-SPECIFIC NUMERATION, NAME:INSTITUTION, SOFTEN, INTENSIFICATION, INVOKE/INSCRIBE and DOWNSCALE.

Table 45: Significant differences between high- and low-scoring ICS in different types of "Target" material

Significant in (Target)	**GRADUATION Feature**	**Over-use in (Score)**	**t-value / significance (UAM)**
Research	NON-SPECIFIC NUMERATION	low	3.028 / high (p<0.02)
Comment: Many of the FORCE:QUANTIFICATION resources used in Research are found in low-scoring texts, which appear to foreground research quality. Examples: "range of", "extensive"			
Research	NAME:INSTITUTION	high	3.545 / high (p<0.02)
Comment: Institutions are named in research-related material in 23 out of 38 high-scoring ICS, distributed across submissions. By contrast, this was done in only 11 out of 38 low-scoring ICS which were submitted from only 7 different institutions.			
Research	SOFTEN	low	2.689 / high (p<0.02)
Comment: In low-scoring texts, there are 23 instances of SOFTEN in Research-related material, spread across 8 submissions. These tend to be the word "explore" or similarly vague expressions for research, such as "investigated", "potential", "engaged with", "can/may". In high-scoring texts, there are only six instances in five submissions, including "initiated", "provides", "could potentially".			
Impact	INTENSIFICATION	low	2.472 / high (p<0.02)
Comment: The relative over-use of INTENSIFICATION in low-scoring texts is surprising, especially because there is no statistically significant difference within any of the four INTENSIFICATION types. The over-use may be due to the large difference in impact-related material overall. The raw number of GRADUATION tags in impact-related material is 441 for High and 169 for Low, which looks like a stark difference; however, the normalised frequencies are 149 and 123 per 1,000 words respectively, which is a much smaller difference. This clearly reflects a difference in corpus size, that is, in the amount of material dedicated to impact claims in high-scoring versus low-scoring texts (see Figure 11 and Figure 12 in section 5.3.1 above). This difference in corpus size could be the reason why the relatively higher number of INTENSIFICATION resources in low-scoring texts stands out more. Examples in low-scoring texts are "enhanced", "improved", "optimised", whereas high-scoring texts tended to use "significant", "substantially", "remarkable" or "first".			

Significant in (Target)	GRADUATION Feature	Over-use in (Score)	t-value / significance (UAM)
Impact	INVOKE/INSCRIBE	high (invoke) / low (inscribe)	2.036 / medium ($p<0.05$)
Comment: The difference in the use of INVOKED and INSCRIBED expressions is also surprising because there is a slight overuse of INSCRIBE in Low, compared to INVOKE in High. This is contrary to what may have been expected based on assumptions in the literature that explicit attempts at "selling" impact led to higher scores (e.g. Watermeyer and Hedgecoe 2016). This difference seems to suggest that low-scoring texts are more explicit and high-scoring texts more subtle in evaluating their impact descriptions. However, on zooming into the respective instances, the difference can be explained through a combination of attitudinal expressions being modified (connected to research topic) and using certain words of SPECIFICITY which are tagged as INSCRIBED by default. The observation that these instances are centred in a small number of submissions, and that the picture is reversed in Pathway-related material, make the evidence for this apparent difference much less reliable.			
Pathway	DOWNSCALE	low	3.239 / high ($p<0.02$)
Comment: This significant-looking difference in the use of the DOWNSCALE tag is more likely related to different types of material. The only three instances in high-scoring texts are all time references (e.g. "recently", "in 2013"), compared to the 14 instances in low-scoring texts, which are mostly the names of cities. However, when discounting the spatial element, the remaining four instances in Low are "small group", "two", "future", "initial". These terms sound either like real DOWNSCALE or convey more preliminarity than the instances in High. Given the very small number of instances, though, the difference should not be treated as important.			

Nothing appeared significant, or even interesting, when comparing the "Problem" sub-corpora in high- vs low-scoring texts, but this is not surprising given the small amount of material that received the Problem tag (5.95%, 550 words in total; see Figure 11 and Figure 12 in section 5.3.1 for an overview of the distribution of material).

This analysis takes into account the type of material in which certain GRADUATION features are used and therefore what it is that they refer to. This allows for a more direct comparison of language use, that is, comparing GRADUATION of, for example, impact claims with impact claims, rather than combined impact claims and problem statements. The significance of this distinction is that certain features such as DOWNSCALE may be viewed differently in the assessment process whether they refer to a problem or to an impact claim. As indicated earlier, when the high- and low-scoring texts in these sub-corpora (divided by different types of content) are compared, there are hardly any statistically significant differences left. This in turn supports the conclusion that LANGUAGE differences are generally more related to the balance of content in a given text than to editorial choices in writing about specific content.

7.2.3 Specific features of interest

The bird's eye analyses discussed in sections 7.2.1 and 7.2.2 were designed to capture any unanticipated differences from the data, motivated by a desire to "let the data speak". The research design gives a certain amount of direction through the adapted annotation schemes and the detailed instructions in the coding manual, as well as individual coder decisions, but at the stage of analysing the frequency of tags, the comparisons were made without assumptions and the results are reported in full. The analyses described in the present section, by contrast, address specific queries. As described in section 7.1.4, these questions arise from two sources:

1. The literature on REF 2014 impact case studies (see Chapter 2); and
2. Previous analyses on this corpus (see Chapter 5 and Chapter 6).

Each question area is first introduced and contextualised with reference to one or more of these sources and then discussed in light of the GRADUATION analysis.

7.2.3.1 Making claims look good

McKenna (2021: 54) is one of many writers who articulate the accusation that ICS narratives were "drummed up", which he attributes to the use of press officers or other journalistically trained writers being tasked with dressing up the impact stories. In the context of the GRADUATION framework, there are several ways in which such "drumming up" could be explored: using INTENSIFICATION resources; comparing instances of INVOKED and INSCRIBED resources in order to check whether the narrative is "drummed up" visibly or more covertly; and the respective use of UPSCALING and DOWNSCALING resources across expressions. These will be discussed in turn in this section.

For INTENSIFICATION resources, the annotation scheme used in this study distinguishes four types: INTENSIFIER, INFUSION, EXPERIENTIAL and REPETITION (see Figure 16 above, details of the features can be found in the coding manual in Appendix F). For evaluating the claim that narratives are "drummed up", the first three of these resources are most relevant because the final one, REPETITION, is arguably more connected to content than to alternative word choice. A search of these first three INTENSIFICATION resources shows that combined scores for INTENSIFICATION (minus REPETITION) are virtually the same across high- and low-scoring texts (normalised per 1000 words: 19.31/18.96, no statistically significant difference). This supports the view that even if ICS were "drummed up", this is not specific to either high- or low-scoring ones.

Looking at the distribution of INTENSIFICATION resources (minus REPETITION) within Main Panels, the following observations emerge. In MP-AB, there are 55 instances, of which 29 are EXPERIENTIAL, and all but one are UPSCALE, with no notable differences between high- and low-scoring ICS. The picture is similar in MP-D, with 73 instances, of which 52 are EXPERIENTIAL and 71 are UPSCALE, and similar distributions in the high- and low-scoring sub-corpora respectively. In MP-C, there are 50 instances, of which 27 are EXPERIENTIAL and all are UPSCALE. However, there are also 15 instances of INTENSIFIER in the high-scoring MP-C sub-corpus, compared to only 4 such instances in the corresponding low-scoring sub-corpus. This is a statistically significant difference ($p<0.05$, $t=2.070$), and it shows that the high-scoring Sections 1 in MP-C contain more INSCRIBED resources of INTENSIFICATION, contributing to more explicit or even bolder claims.

A second way to test whether high- or low-scoring texts were more overtly "drummed up" than the other is to compare the tags for INVOKED and INSCRIBED resources of GRADUATION. Resources tagged as INSCRIBED are overtly visible as evaluating, even to the non-specialist reader less familiar with the topic. Resources tagged as INVOKED may be recognised as evaluation by a reader who is intimately familiar with both the REF ICS guidance and the conventions of their Unit of Assessment. The "drumming up" is still there, but it is done more covertly. One possibility is that high-scoring ICS used INSCRIBED evaluation resources more liberally and that this may have swayed assessors to award higher scores – this would be supported if more INSCRIBED resources were found in high-scoring ICS. A contrasting possibility is that high-scoring ICS were written with great care to make the same claims more subtly, precisely in order to avoid an impression of overt selling, and that the writers of low-scoring ICS were less careful to conceal explicit claims of excellence and therefore used more INSCRIBED resources.

Table 46 illustrates the distribution of INSCRIBED and INVOKED resources across high- and low-scoring texts (raw numbers).

Table 46: Raw number of resources tagged as INVOKE and INSCRIBE respectively

Score	INVOKE	INSCRIBE
High	772	64
Low	501	34

Statistical analysis within the UAM Corpus Tool shows a non-significant difference (chi-square: p=0.36; *t*-test of normalised scores: *t*=0.911). Therefore, neither possibility seems supported by the data, and there is no clear trend that either the high- or the low-scoring texts indiscriminately "drummed up" their achievements.

A third way in which the supposed "drumming up" might be quantified is through exploring what kind of features are associated with UPSCALE tags. This can in theory be attached to any resource that is tagged as FORCE, that is, to 20 different categories. In the following, I describe selected findings for the distribution of UPSCALE tags within each Main Panel.

In MP-AB, there are 195 UPSCALE tags, of which 135 (69.23%) are attached to QUANTIFICATION resources and 60 (30.77%) to INTENSIFICATION resources. Within the INTENSIFICATION category, most resources are EXPERIENTIAL (n=40, 66.67%), and within the QUANTIFICATION category, there is a nearly even split between AMOUNT (n=69, 51.11%) and EXTENT (n=62, 45.93%), with 2.96% (n=4) FREQUENCY. There is no significant difference between the high- and low-scoring sub-corpora, with the exception of INVOKED-AMOUNT: here, NON-SPECIFIC NUMERATION is significantly over-used in Low ($p<0.05$, $t=2.092$), while SPECIFIC-NUMBER dominates in High ($p<0.05$, $t=2.301$). This is consistent with observations in my other analyses that high-scoring ICS provided more specific detail than low-scoring ones.

In MP-C, there are also 195 UPSCALE tags, of which only 67 (34.36%) are attached to INTENSIFICATION resources and 128 (65.64%) to QUANTIFICATION resources. In the QUANTIFICATION:AMOUNT:INVOKED-AMOUNT category (27 instances), NON-SPECIFIC NUMERATION dominates in both High (n=15, 55.56%) and Low (n=14, 82.35%), with only 7 instances (25.93%) of SPECIFIC-NUMBER in High and none at all in Low.

In MP-D, the 198 UPSCALE tags are split more evenly between INTENSIFICATION (n=88, 44.44%) and QUANTIFICATION (n=110, 55.56%). Especially in the low-scoring MP-D sub-corpus, the split is nearly even (40 and 42 instances respectively), which means that high-scoring ICS have a greater skew towards QUANTIFICATION than low-scoring ones. Within the INTENSIFICATION branch, EXPERIENTIAL resources dominate especially in low-scoring ICS (n=26, 65.00%, compared to n=25, 52.08% in high-scoring ICS; difference is not significant). Similar to MP-C, in QUANTIFICATION:AMOUNT:INVOKED-AMOUNT (21 instances), NON-SPECIFIC NUMERATION clearly dominates in both High (n=14, 66.67%) and Low (n=13, 81.25%), with only 4 instances (19.05%) of SPECIFIC-NUMBER in High and none in Low. This means that both Social Sciences

(MP-C) and Arts and Humanities (MP-D) ICS used different ways to UPSCALE their texts in Section 1 than using explicit numbers.

Due to the covertly promotional nature of this register, it is expected that most GRADUATION resources serve to UPSCALE the claims being made. In order to ascertain whether UPSCALING resources are indeed used more, it is necessary to also check the prevalence and distribution of DOWNSCALING resources across sub-corpora. Across the whole corpus, nearly 75% (74.44%) of all resources of FORCE were classified as UPSCALE. However, the MIDDLE and DOWNSCALE tags were also used, for 17.59% and 7.97% of all instances of FORCE respectively. Therefore, it is interesting to see in what contexts the DOWNSCALE tag was used most often, especially in order to establish whether this is in relation to impact claims or to other information.

Overall, the DOWNSCALE tag was used fairly equally in the high-scoring (n=31; 5.11 per 1,000 words) and low-scoring (n=32; 7.20 per 1,000 words) sub-corpora. A confounding factor in this analysis is that many of the MIDDLE and DOWNSCALE tags were applied to resources of SCOPE-SPACE, where the application of tags was defined in the coding manual as UPSCALE for international, MIDDLE for national and DOWNSCALE for sub-national references. Therefore, many of these tags are pre-determined by the impact itself, rather than by a writer's decision. The tables below are therefore presented with (Table 47) and without (Table 48) the SCOPE-SPACE category.

The following categories are associated with the DOWNSCALE tag at least once in the corpus (number of instances for high/low respectively):

- NON-SPECIFIC NUMERATION (5/3)
- SPECIFIC-NUMBER (3/1)
- DISTANCE-TIME (8/3) – here, DOWNSCALE was attached to references to "recent/current" (3x) or the year(s) immediately before submission, e.g. 2013 (4x). The three instances in Low are "future", "early" and "initial", all being used in a way that conveys preliminary or future claims.
- SCOPE-TIME (5/1)
- SCOPE-SPACE (10/23) – this includes 17 from the same ICS in Low MP-D, as explained above.

As indicated above, especially for DOWNSCALE tags it is interesting to see whether these are applied in impact-related material or other parts of Section 1. Table 47 and Table 48 provide

an overview of the occurrence of downscale tags across sub-corpora (rows) and type of content (columns), both as raw numbers (n) and normalised figures (per 1,000 words).

Table 47: Distribution of resources that DOWNSCALE, including all instances

DOWNSCALE all occurrences	Research		Impact		Pathway		Problem	
	n	per 1000	n	per 1000	n	per 1000	n	per 1000
Overall	5	6.20	33	16.03	17	14.85	8	24.49
High	2	4.74	20	14.56	3	4.19	6	41.38
Low	3	7.79	13	19.01	14	32.63	2	20.00
MP-AB	1	3.12	10	9.27	2	4.28	6	59.41
MP-C	2	7.43	9	13.72	3	8.62	2	27.40
MP-D	2	9.22	14	43.34	12	36.36	0	0

This shows that overall, instances of DOWNSCALE appear mostly in impact- or pathway-related material. Around 60% in each are resources of SCOPE-SPACE (impact: 63.64%; pathway: 58.82%), indicating mention of sub-national reach. One reason why there is very little DOWNSCALE in research-related material may be that impact is often more strongly geographically situated than research.

It is interesting that especially in high-scoring texts, there is a clear dominance of DOWNSCALE in impact-related material, compared to research and pathway. These instances are also spread more evenly across different QUANTIFICATION features (AMOUNT, EXTENT). The picture appears to be different in low-scoring texts, with relatively more instances in Pathway. However, this is at least in part due to two MP-D ICS which include the names of several UK cities, tagged as SCOPE-SPACE DOWNSCALE, and which therefore skew the picture in Low. Therefore, Table 48 presents the figures of DOWNSCALE without geographical references:

Table 48: Distribution of resources that DOWNSCALE, excluding SCOPE-SPACE

DOWNSCALE excluding Scope-Space	Research		Impact		Pathway		Problem	
	n	per 1000	n	per 1000	n	per 1000	n	per 1000
Overall	3	4.98	12	8.14	7	9.098	8	45.98
High	1	2.92	11	10.40	3	5.44	6	46.51
Low	2	7.66	1	2.40	4	18.26	2	44.44
MP-AB	1	3.12	10	9.27	2	4.28	6	59.41
MP-C	0	0	2	6.69	2	16.67	2	27.40
MP-D	2	11.24	0	0	3	16.39	0	0

Excluding place-based DOWNSCALE tags, the following uses are representative in the respective target areas:

- **Research**: The one occurrence in High is "more recently", highlighting temporal sequence. In Low, one of the two occurrences is "had received *little or no* critical attention prior to…", which represents background to or history of the research, similar to the one instance in the high-scoring sub-corpus. The second occurrence in Low is "Early findings suggest that", which indicates that there may not have been enough mature research to generate impact from.
- **Impact**: Occurrences are mainly time references to highlight that something has gone "faster" or "rapidly", or to specify that figures were updated prior to submission in November 2013 ("2013" was coded as downscale).
- **Pathway**: As in the impact-related material, occurrences are mostly time references. Here we have an interesting qualitative difference between high- and low-scoring texts: the instances in High are "more recently", "currently" and "2013", indicating sequence and updates, whereas those in Low include "future" and "initial", indicating preliminarity. The remaining two instances in Low appear to indicate limited reach: "two classrooms" and "small-group meetings", although the latter may not indicate scale because there might have been many small-group meetings, which may have been intended in the pathway design, rather than indicating limited reach.
- **Problem**: This is the only type of material without geographical references, but at the same time it is by far the smallest (see above section 7.2.2; this is also reflected in the high per-1000 figures with very low absolute figures). Of the eight occurrences, six are NON-SPECIFIC NUMERATION, including "low", "lack of", "only", "less than". These highlight the opportunity for the research and pathway to result in impact.

As indicated above, the vast majority of FORCE resources were UPSCALING, especially in the research and impact sections, closely followed by pathway. It is only in the problem statements that a third of all instances are either MIDDLE (22.00%) or DOWNSCALE (16.00%). The observation that ICS emphasise the size of their contribution is therefore supported.

7.2.3.2 Making claims complete

Several authors comment on the need to articulate clear causal links between the research and the impact (McKenna 2021; Reichard *et al.* 2020). Moreover, McKenna (2021: 23)

recommends using “action verbs”. Two GRADUATION resources that are realised through verbs, by definition, are INFUSING PROCESS and ATTITUDINALLY INFUSED PROCESS. These tags are applied to verbs that indicate a change in quantity of some kind, and therefore they could illuminate the numbers of such verbs within and across sub-corpora even across their lexical variation. However, such verbs indicating change were used only 17 times across the whole corpus, and normalised frequencies are virtually the same in the high- and low-scoring sub-corpora. Examples are “raising/raised” (5x), “increase(ing)” (3x), “added”, “restore”, “accelerate”.

The FOCUS resource of FULFILMENT includes expressions of causality such as “led to” (stronger causal link) and “informed” (weaker causal link), but causality is not marked specifically, therefore these tags cannot be used to evaluate McKenna’s assertion directly here. Similarly, action verbs are tagged as FULFILMENT but not marked further as action verbs. FULFILMENT indicates to what extent an action is complete (FULFILMENT:SHARPEN) or incomplete (FULFILMENT:SOFTEN) or not marked in either direction (FULFILMENT:NEUTRAL). Whether an action is more or less complete may vary by the type of content: if an action is expressed through a FULFILMENT:SHARPEN resource, it can convey a more complete impact claim. The distinction into SHARPEN, NEUTRAL and SOFTEN in the FOCUS branch of the GRADUATION coding scheme is parallel to the distinction into UPSCALE, MIDDLE and DOWNSCALE in the FORCE branch, which was discussed in the previous sub-section (section 7.2.3.1). This distinction matters because the REF exercise assesses retrospective impact claims, which should be as complete as possible. Therefore, it may be expected that FULFILMENT:SHARPEN tags dominate either in high-scoring ICS or in the impact-related segments in the overall corpus, or both.

There is no significant difference in the number of FULFILMENT tags between any of the sub-corpora, as confirmed by a *t*-test for high- versus low-scoring ICS (*t*=0.114) and a one-way ANOVA for comparing the three disciplinary sub-corpora against each other (One-way ANOVA: F (2, 73)=1.22; p=0.3). However, the distribution of resources of FULFILMENT into SHARPEN, NEUTRAL and SOFTEN is interesting. When comparing the Main Panels against each other, we find that in MP-AB, SHARPEN is overused compared to combined Main Panels C and D (t=2.887, p<0.02). FULFILMENT:SOFTEN is underused (t=3.349, p<0.02). However, this difference misses the p<0.05 significance threshold in a one-way ANOVA comparing AB, C and D against each other directly.

Within the Main Panel sub-corpora, there are further differences. Of all FULFILMENT resources in MP-AB, half are SHARPEN (n=63, 49.61%), one third are NEUTRAL (n=46, 36.22%) and one sixth are SOFTEN (n=18, 14.17%). There is no significant difference in the use of SHARPENED or SOFTENED FULFILMENT resources between the high- and low-scoring sub-corpora. Looking at MP-D (regardless of score), the FULFILMENT tags are more or less evened out across SHARPEN (n=38, 31.40%), NEUTRAL (n=42, 34.71%) and SOFTEN (n=41, 33.88%). This is similar in MP-C, although SOFTEN (n=34, 25.76%) is trailing behind SHARPEN (n=49, 37.12%) and NEUTRAL (n=49, 37.12%). It seems, therefore, that the preference for FULFILMENT:SHARPEN is most pronounced in science writing.

When comparing high- and low-scoring texts in the disciplinary sub-corpora, there is one statistically significant difference, which can be found in Main Panel C. In both high- and low-scoring texts, there is a comparable distribution of FULFILMENT:SHARPEN (high: n=34, 42.50%; low: n=15, 28.85%) and FULFILMENT:NEUTRAL (high: n=32, 40.00%; low: n=17, 32.69%) tags. However, the high-scoring MP-C sub-corpus contains only 14 instances (17.50%) of FULFILMENT:SOFTEN, whereas the low-scoring MP-C sub-corpus contains 20 such instances (38.46%). Coupled with the difference in text length, that is, the fact that the low-scoring MP-C sub-corpus contains fewer words, this difference is significant (t=2.747). This supports the interpretation that in this Main Panel, low-scoring texts made more tentative or preliminary claims than high-scoring texts.

In Main Panel D, there is an especially interesting distribution. The low-scoring sub-corpus includes overall more FULFILMENT tags than the high-scoring sub-corpus, but the distribution is inverted:

- Low: SOFTEN (28) > NEUTRAL (21) > SHARPEN (12)
- High: SHARPEN (26)> NEUTRAL (21) > SOFTEN (13).

If the previous observation is correct that FULFILMENT:SHARPEN is characteristic of science writing as seen in both high- and low-scoring texts in MP-AB, then it could be concluded that high-scoring Humanities ICS adapted more to this style than low-scoring Humanities ICS, which may have been written in a style more closely aligned with Humanities prose (as described by Biber and Gray, 2016). Writers of high-scoring ICS therefore may have been more successful in transcending disciplinary conventions in this respect.

In addition to the distribution of tags across high- and low-scoring ICS and across and within Main Panels, it is interesting to see what type of content these FULFILMENT resources relate to. Therefore, the remainder of this section reports on an exploration of FULFILMENT resources, and their split into SHARPEN, SOFTEN and NEUTRAL, in the different types of content – research, impact and pathway.

The greatest density of FULFILMENT resources is in pathway-related material (overall 63.33/1,000 words). About half of these are tagged as NEUTRAL (n=79, 45.93%; e.g. “inform*” 20x, “influenc*” 10x, “contributed to” 3x), with a fairly even split of the other half into SHARPEN (n=43, 25.00%; e.g. “led to” 4x, “used” 4x, “achieved” 3x) and SOFTEN (n=50, 29.07%; e.g. “providing” 4x, “engaging with” 3x, “recommendations” 2x).

The impact-related material includes only 35.09 FULFILMENT resources per 1,000 words, and here the clear majority are SHARPEN (n=89, 58.55%), followed by NEUTRAL (n=49, 32.24%) and a small number of SOFTEN (n=14, 9.21%). This picture is the same across scoring brackets and Main Panels. It is interesting that different directions dominate in pathway- and impact-related material because impact claims need to be retrospective, which is best conveyed with an expression conveying completeness of an action, that is, SHARPENED FULFILMENT. Impact descriptions include SHARPENING expressions such as “changed/s” (7x), “resulting in” (3x), “have been adopted” (1x) along with many other unique examples. The SHARPENING expressions that are used in pathway-related material also appear in impact-related material. The NEUTRAL expressions in impact-related material are much more varied than those in pathway-related material and include “sparked”, “impacted on” and “empowered”.

Research-related material contains even fewer FULFILMENT-related expressions, at 19.47 per 1,000 words, with problem-related material containing only four FULFILMENT tags overall.

In order to ascertain whether there are any significant differences in the use of SHARPENING or SOFTENING FULFILMENT resources across scoring brackets, the high- and low-scoring texts within each Target sub-corpus were tested for significance (i.e. FULFILMENT in all research-related material in high-scoring texts vs all research-related material in low-scoring texts, and the same for pathway- and impact-related material). Despite the relatively higher use of SOFTEN resources in low-scoring texts in the pathway-related material, the difference between high- and low-scoring texts in the relative split of FULFILMENT resources (SHARPEN/NEUTRAL/SOFTEN) was not significant in either the pathway- or the impact-related material.

However, there is a difference in research-related material (see Figure 21). In high-scoring texts, more than half of the instances are SHARPEN (n=18, 55.55%; e.g. “based on” 2x, “development” 2x, “generated by”), while in low-scoring texts, two thirds of the instances are SOFTEN (n=20, 66.67%; e.g. “explore*” 8x, “engaged with”, “illustrated”, none of which appear in the high-scoring texts). This translates into significant ($p<0.02$) differences in the use of SHARPENING and SOFTENING resources. This seems to be partly due to the type of research, but especially in the context of REF impact assessment, the research stage could also be playing a part (“explore” indicating an early stage, as opposed to “based on”), with low-scoring ICS being built around research that had not yet matured in the same way.

	high		low			
Feature	N	/1000wds	N	/1000wds	TStat	Sign.
Total Units	18	8.32	30	21.07		
FOCUS-DIRECTION	N=18		N=30			
- sharpen	10	4.62	4	2.81	3.415	+++
- neutral	3	1.39	6	4.21	0.281	
- soften	5	2.31	20	14.04	2.760	+++

Figure 21: Difference of FULFILMENT resources in research-related material in high- and low-scoring texts

In a disciplinary comparison across Main Panel sub-corpora, there is a significant difference in the use of SHARPENING and SOFTENING resources in the Research sub-corpus, but not in any of the other types of material. The overall number of FULFILMENT resources in the Research sub-corpus is similar across Main Panels (MP-AB: n=13; MP-C: 18: MP-D: 17), but in MP-AB, the majority of instances are SHARPEN (n=7, 53.85%), whereas SOFTEN dominates in MP-C (n=11, 61.11%) and MP-D (n=11, 64.71%). Only three instances of FULFILMENT:SOFTEN appear in MP-AB across two ICS, one high-scoring (“could potentially”) and one low-scoring (“explored”, 2x). MP-C and MP-D each include 11 instances of FULFILMENT:SOFTEN in their research-related material, spread respectively across six ICS (MP-C, including 2 high and 4 low) and four ICS (MP-D, including 2 high and 2 low). All four high-scoring ICS in MP-C and MP-D that used FULFILMENT:SOFTEN resources only include one of them, whereas the low-scoring ICS in these Main Panels use several FULFILMENT:SOFTEN resources in their research-related material.

Overall, the exploration of FULFILMENT resources across the corpus shows no statistically significant differences between the complete high- and low-scoring sub-corpora. When exploring the distribution of features in more depth, some observations can be made: MP-

AB ICS contain a relatively higher number of SHARPENED FULFILMENT resources than MP-C or MP-D, indicating a writing style or perhaps an impact stage where more claims of completed impact are made. It also seems that within MP-D, the high-scoring texts included more SHARPENING resources, while SOFTENING resources dominated in the low-scoring texts, aligning the former more closely with MP-AB findings. Within MP-C, the proportion of SOFTENING tags in all FULFILMENT resources is twice as high in low-scoring texts as in high-scoring ones, indicating more tentative or preliminary claims. Finally, markers of FULFILMENT, that is, of more or less completed actions, are most concentrated in pathway-related material, followed by impact-related material, with no significant difference in the percentage of SOFTENING and SHARPENING resources. This contradicts the expectation that SHARPENING resources dominate in impact-related material. The finding that they are most prevalent in pathway-related material may be explained by the fact that this material included descriptions of what was done in order to achieve the impact, and therefore may include more action verbs.

7.2.3.3 Making claims specific

A third area of textual characteristics that are recommended (McKenna 2021: 28) and observed (Reichard *et al.* 2020) is the specificity of information. More specific detail enables assessors to get a more certain picture of the claim and leaves less room for interpretation. In the GRADUATION framework, one resource that enables this can be found in the FOCUS branch, namely SPECIFICITY, which is applied to expressions that SHARPEN or SOFTEN the boundaries of a claim (as opposed to resources of FULFILMENT, which SHARPEN or SOFTEN the completeness of an action). It could be expected that impact-related material is more likely to include more explicitly specific material, especially in high-scoring ICS, and that a difference between high- and low-scoring texts may be less pronounced in research- or pathway-related material.

The low overall numbers of SPECIFICITY tags make statistical comparison difficult, and this may be a reason why no significant differences between high- and low-scoring text within the Target sub-corpora appear (research: n=21; impact: n=33; pathway: n=20). It can be observed, however, that in each Target sub-corpus, resources of SPECIFICITY appear more often in low-scoring texts, per 1,000 words; in the research sub-corpus, this amounts to nearly twice as often (18.26) as in high-scoring texts (9.93). Example expressions in research-related material are “particular(ly)” (6x), “focus* (on)” (6x), “specific” (2x), “identif*” (2x).

The distribution of SHARPEN (n=61) and SOFTEN (n=21) tags is similar across scoring brackets within each Target sub-corpus, but interestingly, the majority of SOFTEN tags appear in impact-related material (n=7 in High, n=5 in Low). Examples are “primarily”, “estimated”, “probably”, which all indicate a degree of uncertainty. This is surprising in light of the advice to be specific. It seems that adding or highlighting specificity and contextualisation in the presentation did not influence the score of an ICS beyond the content that is being specified. However, given that the data basis for this analysis is Section 1 (“Summary of the impact”) with an indicative limit of 100 words, and that the template includes a section titled “Details of the impact” with an indicative limit of 750 words, it is equally possible that more specific detail was included in other parts of the ICS, rather than in the summary section.

In comparisons across Main Panels, no statistically significant differences can be detected either, but there are trends: in MP-AB, resources of SPECIFICITY are spread fairly equally across research-, impact- and pathway-related material, around 10/1,000 words. In MP-C, relatively more resources are used in pathway-related material and also in research- and impact-related material, the normalised figure is higher than in MP-AB. Conversely, in MP-D, very few resources of SPECIFICITY are used in either type of material, with the largest number in research-related material but fewer instances in impact- or pathway-related material than either of the other Main Panels.

Another way to increase the specificity of claims in a text, and one that has received particular attention in the ICS context, is the use of quantitative indicators. These play a role in bridging the gap between the two fundamental approaches to research impact assessment, namely metric and narrative methods, as discussed above in section 2.1.2. Two GRADUATION resources that are especially relevant as such quantitative indicators in the overall QUANTIFICATION branch are SPECIFIC-NUMBER and FREQUENCY (see above Figure 20 for the QUANTIFICATION branch of the coding scheme). Of these, FREQUENCY only appears 10 times, in 7 (out of 76) texts overall, spread across four out of six sub-corpora, which makes any attempt at comparison across sub-corpora meaningless. However, the resource does not appear as often as might be expected in light of the importance placed on quantitative indicators and on providing context for claims, as explained above. Most instances are a form of “yearly” or “annually” (6x in 4 texts), others are “often” and “sometimes” which each appear once, and two are linked to a specific recurring “event” (“per child”, “per start-up”). Here, an explanation may also be that these texts are Sections 1 which are designed as a succinct

summary and therefore may have prioritised other information, although a one-word addition such as “annually” can add much-needed context.

SPECIFIC-NUMBER tags appear much more often (n=55) than FREQUENCY tags. They contrast with NON-SPECIFIC NUMERATION tags, which are even more frequent (n=118). In fact, NON-SPECIFIC NUMERATION is the most frequently used feature within the INVOKED-AMOUNT category across the whole corpus.

High-scoring ICS are expected to have relatively more instances of SPECIFIC-NUMBER than low-scoring ICS. Indeed, a comparison of this feature in high- versus low-scoring texts, regardless of Main Panel, shows that SPECIFIC-NUMBER is used significantly more often (t=2.675, p<0.02) in high-scoring texts than low-scoring ones, and inversely, NON-SPECIFIC NUMERATION is used significantly more in low-scoring texts (t=2.164, p<0.02).

A comparison across scoring brackets within each Main Panel shows interesting observations. Within MP-AB, SPECIFIC-NUMBER appears more often in the high-scoring texts (n=29) than resources of NON-SPECIFIC NUMERATION (n=25). The inverse is found in low-scoring texts, where there are twice as many NON-SPECIFIC NUMERATION (n=20) tags as SPECIFIC-NUMBER tags (n=10). Overall, in MP-AB, NON-SPECIFIC NUMERATION (n=45) is closely followed by SPECIFIC-NUMBER (n=39); by contrast, SPECIFIC-NUMBER is trailing far behind in MP-C (n=7 in High, n=0 in Low) and MP-D (n=7 in High, n=2 in Low). None of the comparisons across scoring brackets within Main Panels show significant differences, though. Moreover, as indicated in section 7.2.1 (Table 44), a one-way ANOVA between disciplinary sub-corpora indicates a statistically significant overuse in MP-AB compared to both MP-C and MP-D separately. With the difference between high- and low-scoring texts disappearing when drilling into the Main Panel sub-corpora, this indicates that there may be a disciplinary difference in the use of SPECIFIC-NUMBER, with uses higher in science writing.

7.3 Discussion and conclusion

In this chapter, I have made several important contributions through tagging the texts in Sample C for features of GRADUATION. First, this enabled a comparison of functional language features based on tags that could be attached to different lexical entities, and therefore functions could be quantified across different expressions. Second, the use of a principled coding scheme enabled me to show that the statistically significant differences that I found cover only a small proportion of the areas where differences could have been found. Third,

through the combination of different layers with different coding schemes in the analysis tool, I could differentiate the use of certain language functions based on the content to which they refer, that is, the different Target types. And finally, more fundamentally, I could start with the situational context of the register of ICS to adapt the coding scheme (refined in several iterations of interacting with the texts themselves, as set out in section 7.1.3) for the functions that should be explored, which were then made concrete by relating them to instances of language, rather than starting with language features and trying to explain their function as was done in chapter 6 (see section 3.3 for more on this distinction of research sequence).

As explained in section 3.3, the appraisal of content in ICS is assumed to be positive, as it is put forward for assessment. The GRADUATION part of the Appraisal system then serves to make visible how this is dialled up or down. My analysis provided an overview of the kind of language that was used to achieve this, without being tied to specific lexical items. In this section, I summarise some examples.

The features that are used most frequently across ICS sit in the QUANTIFICATION branch of the GRADUATION scheme (41.58% of all tags, n=570). This means that devices of QUANTIFICATION play a leading role in pushing the attitude of an ICS, and of these, the vast majority (97.19%, n=554) are INVOKED rather than INSCRIBED. Two thirds (65.44%, n=373) are UPSCALE, followed by MIDDLE (23.51%, n=134) and DOWNSCALE (11.05%, n=63).

Looking at differences between high- and low-scoring ICS across the whole coding scheme, as shown in Table 43, those few features where a statistically significant difference could be found are all part of the QUANTIFICATION branch, namely SPECIFIC-NUMBER, DISTANCE-TIME and NON-SPECIFIC NUMERATION. The first two of these were used relatively more often in high-scoring ICS, pointing towards a higher level of specificity, while the last one is typical of low-scoring ICS.

For the second axis of comparison, namely Main Panel (discipline), there were several differences in how a positive story is pushed. One important example, which is related to QUANTIFICATION, is that ICS in MP-A included relatively more instances of SPECIFIC-NUMBER, whereas MP-C and MP-D include more instances of NON-SPECIFIC NUMERATION than specific-number. This shows that the positive story was pushed in other ways in both Social Sciences (MP-C) and Arts and Humanities (MP-D) ICS than by using explicit numbers (section 7.2.1, Table 44).

The third distinction between corpus parts, which is only applicable in the Appraisal analysis, is that into different types of content (research, impact, pathway, problem and other). In combination with a focus on the direction of GRADUATION, that is, UPSCALE – MIDDLE – DOWNSCALE, this distinction enables a check on whether a positive attitude is pushed in relation to impact or to other parts of the story. As explained above (section 7.2.3.1), the vast majority of FORCE resources in the material that relates to either research or impact were UPSCALING. In those sections that describe pathways to impact, UPSCALING resources also dominate, but not by the same margin. However, in the problem statements, a third of all instances of FORCE are either MIDDLE or DOWNSCALE. This illustrates the importance of distinguishing what exactly a resource of evaluation refers to, so that it does not appear like the DOWNSCALE or negative assessment in some sections balance out the positive assessment in other parts of the text.

One main aim of this analysis was to explain the perceived discrepancy between the characterisation of ICS as description and persuasion respectively. This discrepancy stems partly from the tension in the various purposes of the text for assessment, namely to provide an objective description while at the same time persuading the assessor, and partly from the relative lack of overt stance markers (see section 3.2.3, Figure 2). The aim was therefore to ascertain whether a different mechanism of persuasion may be found in this register, and this led to the inclusion of separate tags for INVOKED and INSCRIBED resources of GRADUATION in the coding scheme used for the Appraisal analysis (following Xu 2017). These categories could then be used as a framework to record, and quantify, covert or invoked meaning in ICS, as well as more explicit ways of showing stance, and thereby bridging the gap between different perceptions of the persuasive nature of the register. In Hood's (2019: 390) words, this "enables the […] writer to maintain a veneer of objectivity while implying stance".

Looking again at all GRADUATION tags across Sample C, INVOKE dominates with 92.85% of all tags (n=1273), with a much smaller number of instances of INSCRIBED GRADUATION (7.15%, n=98). This shows that the vast majority of language items that carry persuasive meaning in the context of ICS do so indirectly, and this is consistent with the perceived discrepancy in the nature of the register. It is also in line with Tupala's (2019) research on policy documents and her argument that recognising language items as carrying implicit evaluation can be contingent on the situational context of a text.

With this overarching finding, the next question, as posed in section 3.3, is whether this differs across different parts of the corpus. As shown in Table 46 and the discussion around it (section 7.2.3.1), there is no clear trend that either the high- or the low-scoring texts had a higher ratio of INSCRIBED and INVOKED resources, and therefore it can be said that neither approach is typical for high- or low-scoring ICS. This means that it is less likely that the use of either explicit (INSCRIBED) or implicit (INVOKED) resources was a more successful means to persuade assessors of the merits of an ICS beyond its content.

A final observation which also supports the conclusion that editorial language choice did not play an undue part in the assessment process is that many of the findings described in section 7.2 are related to content. This shows in two ways. One is that a frequently applied feature in the coding scheme is linked to the geographical scale of the claimed impact, namely SCOPE-SPACE. As discussed in section 7.2.3.1 in the context of the DOWNSCALE tag, SCOPE-SPACE appears a high number of times in a low number of ICS and therefore skews the data. More importantly, it is not in the power of the writer to claim a larger geographical scale than the impact had in the first place, and therefore it was not possible to over-sell this – it may only have been possible to accidentally under-sell the scale of the impact.

The second way in which the findings of the GRADUATION analysis are linked to content is the distinction into types of content through the Target layer. This allowed a finer distinction between the kinds of information that evaluative language items were attached to, and as explained above (section 7.2.2), some of the distribution of features is linked to the distribution of content, rather than scores or disciplines. For example, when looking at the direction of FOCUS resources, it is interesting that FULFILMENT:NEUTRAL resources dominate in pathway-related material and FULFILMENT:SHARPEN in impact-related material, where the retrospective claim to impact is best expressed with language that emphasised that something has indeed happened.

Chapter 8 Conclusion

This study has focused on developing our understanding of impact case studies as a genre with a register appropriate for research impact assessment, in response to calls for metrics-based assessment. Such calls are based on the assumption that language is a confounding variable, as expressed for example by Hyland and Jiang, who assert that ICS were "hyped" to an extent that the "usefulness and reliability" of assessment was in danger (2023: 2). In order to empirically test this assumption, I used a selection of qualitative and quantitative methods and corpus linguistic tools to address this overarching research question:

To what extent could presentation have influenced the scoring of impact case studies in REF2014?

My contribution is therefore an application of empirical corpus analysis to the register of impact case studies, providing a within-register comparison between texts that were awarded top scores and those that received low scores, to the extent that they can be distinguished using the publicly available data. This generates evidence that can be used to evaluate claims and assumptions about ICS which so far have relied on personal experience and intuition, or the analysis of high-scoring ICS without reference to lower-scoring comparators. Since these claims may influence future policy on REF methodology, this study thereby provides empirical evidence that may help inform future iterations of REF. If little difference in presentation can be discerned, then this would constitute evidence that claims about language as a confounding factor may not be justified, strengthening the view that narratives are a fair method of assessment. Conversely, if there are differences, making these visible would help mitigate imbalances between universities that have previous experience of creating high-scoring ICS and those where unclear presentation may have prevented assessors from awarding credit to high-magnitude impacts. Overall, the differences that I found are not significant enough to substantiate claims of unfairness due to language.

8.1 Research contributions

From a linguistics perspective, this thesis is a register analysis. Having established in chapter 3 that registers are defined first by context and then by linguistic features, I showed, on the basis of the situational analysis in chapter 5, why ICS should be written in a more explicit way and therefore form a register separate from other academic writing. In this thesis, REF2014 ICS were examined systematically from various angles. These include a situational analysis

(provided in section 5.1 and summarised in section 8.1.1) and various linguistic analyses (provided in sections 5.2.3 and 6.2). They are complemented by analyses relating to the content of the texts (section 5.3) and a specific focus on evaluative language (chapter 7). The final component of a register analysis is then to compare the situational and linguistic analyses in order to offer explanations from the situation for the prevalence of language features in registers with those circumstances, that is, to provide insights into the factors that may play a part in shaping the distinctive linguistic characteristics of its texts. This comparison is one of the defining features of register studies, as opposed to other kinds of language variation, such as those studied by variationists or found in dialects, where variation is based on convention, rather than interpreted as functional (Biber and Conrad 2019: 69).

8.1.1 Summary of situational analysis

In order to illustrate the specific place that ICS occupy in relation to other academic writing, the separate circumstances of research articles and ICS described in detail in section 5.1 are summarised in this section. Research articles are written for an unknown, unspecified intended reader who can seek additional information about the subject matter if they wish. The target reader is usually specialist enough to understand the material, and in cases where a reader is not a specialist, they may have specifically chosen to read this text and may be able to allocate further time to invest the additional effort required to understand the text to the level that matches their specific reading purpose. It is not essential that the reader understands everything at first reading. There is therefore less need to consider a more general reader when balancing the conciseness, clarity and explicitness of writing in a research article. While articles are often important for either the author(s) or the university, especially in the context of research assessment and related metrics, this is mitigated by the fact that each article is part of a larger body of publications. Research articles are not primarily written for assessment but to contribute to knowledge (although see Tusting 2018, on the change in writing practice for journal articles in the climate of REF assessment).

By contrast, ICS have a different context. Although it is the content of the text that is assessed, as opposed to its presentation, the text on the four pages (in REF2014 – five pages in REF2021) is the only window to that content, and therefore the text needs to be maximally effective at conveying the information on non-academic research impact that is the subject of the assessment. The reader is not allowed to access any additional

information on the content of the text and therefore needs to be convinced by the text itself. There is a specific reader (i.e. the 2-4 panel assessors plus sub-panel chair, Manville *et al.* 2015a: 11) but the writer has only very general information about who those people may be. While the membership of each sub-panel is published, the allocation of ICS to assessors is not, and in 2014 this was handled in different ways by different panels (Manville *et al.* 2015a: 15). Therefore, the degree of specialism and shared understanding is unknown, and it is unclear to what extent a greater shared understanding, where it exists, is permitted to actually be taken into account in the assessment process (Manville *et al.* 2015a). While this situation is similar for grant applications, the REF is a one-time chance every 6-8 years with no opportunity of resubmitting a modified version or obtaining alternative funding. It is linked to a steady, predictable, source of year-on-year funding ("Quality-related" or QR funding) for universities. The stakes of this single text are therefore much higher than for one of a series of research articles (and a single ICS is worth considerably more in the REF than a single research article, e.g. Collett 2024), and in combination with the specific but unknown reader, the need to provide clarity and explicitness for this reader is greater than in much other academic writing.

8.1.2 Main findings

In this section, I summarise the main findings from chapters 5-7 in relation to the research questions set out in section 1.2 and provide possible explanations for them on the basis of the situational analysis.

Research question 1: What features related to the *presentation* of the research, pathway and impact, as opposed to the stated criteria of *significance* and *reach* of the impact and the clarity of *attribution* to the research, may be characteristic of high- or low-scoring ICS and therefore may have influenced the score?

This question is divided into three aspects, which will be discussed in turn:

a. Context and reading experience
b. Balance and emphasis of content
c. Words or phrases

Aspect a: Context and reading experience

The context, as summarised in section 8.1.1, is shown to influence the way that ICS are written, and therefore it was discussed separately in the previous section. It shows that the main addressors were the REF assessors reading the ICS in a high-stakes environment, which placed significant pressure on the texts to be convincing in an immediately accessible way. The ways in which the texts were presented as convincing was tested for this thesis in two different ways. First, in an attempt to test the reading experience, a qualitative thematic analysis found that a higher proportion of high-scoring than low-scoring ICS was found to display effective formatting and was judged as easier to read by the research team taking the perspective of an assessor-reader (section 5.2.2). Second, readability was assessed quantitatively with Coh-Metrix, using the eight most relevant tests included in the tool (section 5.2.3). Of these, three showed a significant difference between high- and low-scoring ICS (namely, Connectivity, Deep Cohesion and Flesch Reading Ease), with high-scoring texts being easier to read than low-scoring ones. Nonetheless, even for the three tests that showed a significant difference, the means are so close and the variation within each sub-corpus is so widespread that it does not allow firm conclusions about these features being markers of one or the other group of texts.

The correlation between scoring and the ease of reading, including effective formatting, should not be taken as evidence of a causal relationship between formatting and style on the one hand and scores on the other hand. There were plenty of well-formatted and easy to read low-scoring ICS as well as some less favourably formatted and difficult to read high-scoring ones. Thus, whilst the ease of reading may make it easier for assessors to understand the impact described in the ICS and therefore seems to convey an advantage, the correlation could also be explained by the fact that many high-scoring ICS originated from better funded research universities who were able and willing to dedicate more resources in the writing process, as well as having had the resources to generate significant impact in the first place.

One way in which the clarity and explicitness required by the reading context is achieved is the use of connectors to increase Connectivity and Deep Cohesion (section 5.2.3). The features combined in these measures are used more in high-scoring ICS than low-scoring ones, both overall and within each Main Panel, meaning that their higher prevalence is associated with higher scores. It does not mean that a greater use of, for example, subordinating conjunctions caused these higher scores, but it may mean that the people

writing ICS that ended up being scored highly had a greater awareness of the need to make relationships between ideas explicit. This is especially pronounced in the texts of Main Panel A (broadly Life Sciences), which have higher Connectivity and Deep Cohesion than the other Main Panels, in both high- and low-scoring ICS. This finding is surprising in light of the work by Biber and Gray (2016: chapter 4) showing that science writing is maximally inexplicit compared to academic writing in other disciplines (see section 3.1.3). My finding therefore supports the assumption that the reading situation and the high stakes of ICS may have influenced language choices at the writing stage.

A second point that can be drawn from that same finding is that, according to Dontcheva-Navratilova (2020), the use of logical connectors, which are subsumed under the Connectivity measure in Coh-Metrix, is linked to persuasion: "Persuasiveness is further enhanced by explicit logical connectors (*however, despite*) […] leading [readers] to the intended discourse interpretation" (2020: 22, italics in original). The purpose of ICS according to the situational analysis is partly persuasion, that is, to lead the readers to the intended interpretation. It is therefore interesting to see that while such connecting language is used statistically significantly more frequently in high-scoring texts, this is overall at a very low level compared to general English texts (see section 5.2.3). This finding supports the view that the persuasive nature of these texts is not borne out in the language choices to the extent that would be necessary for supporting assertions that persuasive language compromised the assessment process.

Aspect b: Balance and emphasis of content

Two of the studies completed as part of this thesis engaged with the content of impact case studies, specifically with those aspects of content that were comparable across topics and disciplines. The themes that emerged from those are (1) the level of specificity, (2) the role that evidence plays and (3) the balance between impact claims and descriptions of pathways.

The results of the thematic analysis summarised in section 5.3.2 show a very strong link between high ratings and ICS that could clearly demonstrate the significance and reach of their impact, including who their beneficiaries were and how they benefited, as well as how the impact was enabled by the university's research. The highly-rated ICS not only had clearer explanations but also provided more specific details in those explanations. A failure to be specific about the benefits and beneficiaries generated by the research could plausibly

be considered a failure in sufficiently evidencing the benefits of the impact.[24] This might partly explain Manville *et al.*'s (2015a: 39) findings that in some cases the overall presentation made it difficult to draw out "the substance" of an impact claim which the assessor felt was in there somewhere. In such cases, assessors noted that they "were aware that presentation affected their assessment of the impact" (Manville *et al.* 2015a: 39).

Moreover, these ICS provided strong evidence for both the impact claims themselves and the link between impact and the university's research. Providing evidence of impact to the extent required for REF is a unique feature of those texts, as opposed to creating impact from research for its own sake or to writing a public-facing informative case study text. Such evidence is needed here because REF ICS are used as a basis for making decisions about high-stakes funding allocations as discussed in section 2.1.1, and therefore they need to earn the trust of the assessors. A related feature that is linked to the reading circumstances described in section 5.1 is that part of this evidence took the form of quotes from testimonials or other sources in the main text of the ICS, because assessors did not routinely have access to those and therefore any information that a writer wanted the reader to have had to be part of the case study text itself.

Finally, as shown in the thematic analysis (section 5.3.2), low-scoring ICS were much more likely to focus on the underlying research or the pathways to impact rather than on the impact itself. A similar conclusion can be drawn from an analysis of Section 1 of the ICS template presented in section 5.3.1 in this thesis. This shows that although most ICS provide details about the impact and/or beneficiary in Section 1, those ICS in the sample that do not are part of the low-scoring sub-corpus.

Aspect c: Words or phrases

Chapter 6 provided a lexical analysis, identifying n-grams that respectively appear more often in the high- or low-scoring sub-corpora. Similar to the findings from the analysis of content, this analysis suggests that high-scoring ICS exhibit more specific references, whereas references and other language in low-scoring ICS are often vaguer. The main contribution of this analysis is the repertoire of n-grams that were found to be key in either sub-corpus, as

[24] It is of course often difficult to evidence, perhaps even quantify, the nature and/or extent of research impact. Therefore, the use of the word "failure" does not imply that in all cases a given claim could have reasonably been evidenced more specifically.

presented in section 6.2.1 and Appendix D. Further, more theoretical, contributions from the investigation of words and phrases relate to research question 2 and to the level of editorial choice summarised in the next section (8.1.3).

Research question 2: What linguistic markers of persuasion and evaluation do ICS feature, and does this differ between high-scoring and low-scoring ICS?

An important contribution from chapter 6 is the categorisation of persuasive language for this particular register, based on the criteria of REF. The situational analysis in section 5.1 and the literature discussed in chapter 2 show that in these texts, certain terms that may otherwise look neutral have specific meanings or significance in the REF context. On this basis, in chapter 6 I established four categories of persuasion in ICS, that is, four ways in which certain word sequences could be seen as adding persuasion to the text: Credibility, Added Value, Richness, and Specificity (definitions in section 6.1.6). Credibility strengthens the link between the research and the impact, which enhances the claim of eligibility. The other three categories are ways of enhancing the claims of significance (especially Added Value and Richness) and reach (especially Specificity). The tables in section 6.2.3.2 provide examples of language that can carry these persuasive meanings, but these language items are illustrative: they occur in the relevant parts of the corpus, but should not be seen or treated as lists of wordings that are recommended for use in ICS.

In section 3.2.1, I raised the possibility of identifying a repertoire of evaluative language in ICS, following Dontcheva-Navratilova (2020: 28). Having completed the research, this takes the shape of establishing ways in which persuasion (chapter 6) and evaluation (chapter 7) are expressed, with illustrative examples of language items through which these ways can be realised.

As shown in section 7.3, language related to QUANTIFICATION as defined in the GRADUATION scheme is the most frequently used way to imply evaluation in Section 1 "Summary of the impact". However, this does not mean that quantitative data are required for high-scoring ICS. The label QUANTIFICATION in this analysis includes numerical data as well as time and space references and non-numerical expressions of volume, such as "a number of". In fact, such resources of NON-SPECIFIC NUMERATION are more frequent across the Section 1 corpus (Sample C) than actual numbers, and they are also more frequent within each sub-corpus except high-scoring ICS in MP-AB, where there are slightly more SPECIFIC-NUMBER resources than NON-

SPECIFIC NUMERATION (section 7.2.3.3). Therefore, while quantitative information plays a leading role in making an ICS convincing from the first glance, there are many other ways in which this is achieved.

Such "language designed to 'convince' the reader" was previously investigated by Derrick *et al.* (2014: 153) in the context of REF2014-related policy documents, where they point out that these texts contain more such language than academic articles. Similarly for ICS, the art of convincing the reader is different to the situation and context of academic articles. This difference in context means that language to convince the reader had to be investigated separately to pre-existing frameworks, unlike in, for example, Hyland and Jiang (2023: 2) who applied their framework of "hype" in research articles to their analysis of ICS. The data-driven framework developed in chapter 6 (sections 6.1.6 and 6.2.3), and the theory-driven framework adapted and applied in chapter 7, contributed to a method of characterising language designed to persuade in this particular register. This had not been attempted before, and it is important both because the context and therefore the repertoire of evaluative language is different to other registers, and because the high stakes of these texts warrant explicit information about this being available to all participants, rather than accepting the advantage that well-resourced writers would otherwise have.

8.1.3 New framework: content-driven versus editorial differences in language

Descriptions of register differences can be based on intuition or on empirical analysis of a corpus (Biber and Conrad 2004: 41). The majority of other commentators, including those who acted as assessors in REF2014, used experience and intuition in their publications to make claims about differences (e.g. McKenna 2021: 18). Those that did base their conclusions on empirical evidence drew only on high-scoring ICS without reference to low-scoring comparators (e.g. Gow and Redwood 2020). My study provides empirical evidence to test those intuitions, based on the use of a range of linguistic methods, as described in section 4.2.2, applied to both high- and low-scoring ICS. The conclusions that can be drawn from the findings of these analyses are rooted in a consistent framework that I introduce for assessing to what extent linguistic differences between high- and low-scoring ICS could be considered an indicator of undue influence of language on the scoring. This framework is based on the insight that certain linguistic features may be dictated by the content described, whereas others are within the discretion of ICS writers. A correlation between certain content-dictated linguistic features and high scores would not be overly concerning in

the context of evaluating the effectiveness of using ICS to assess impact. However, a correlation of high scores and discretionary linguistic features, especially those aimed at persuading the assessor of the merits of that ICS, could strengthen the argument of those who claim that language choices materially impact the assessment outcomes.

In the thesis, I therefore distinguish between three categories of differences that could arise between high- and low-scoring ICS: content-driven, general editorial and persuasive language choices as summarised below (this is described most systematically in section 6.1.5, but the distinction feeds into the categorisation and evaluation in various analyses of this thesis). This distinction is crucial to make before voicing any claims that it was language (or other aspects of presentation) that had an influence on scores. Differences in each category weigh differently on the question of whether they may indicate a causal relationship between language choices and the star rating of a given ICS.

Content: Differences in language can be the result of different content, that is, different impacts being discussed (e.g. the term *government policy*). It is expected that different impacts are being described across the corpus, and a difference between ICS (whether high- or low-scoring) in content words is naturally expected. Therefore, differences in language dictated by the impact described are unlikely to be indicative of undue influence of linguistic factors on ICS scoring.

This does not deny the skill that is needed to shape the narrative of an ICS and the challenge that ICS writers face: where an ICS fails to appropriately describe and evidence the research impact, reach and significance, the assessors could hardly be expected to take those parts of the impact that were not described in the text into account in their assessment. However, any assessment based on written evidence would face this constraint.

Editorial language choices: Differences in language can stem from personal editorial choices (e.g. *in terms of*), which in turn can be influenced by the style guides of individual institutions. Such choices may be incidental or linked to experience with other registers, such as academic writing for peer-reviewed publications. If there are significant differences in language where writers or editors of ICS have free editorial choice, there may be more of an argument that certain language features correlate with high or low scores. This may or may not be problematic for the integrity of the assessment process, depending on the extent to which certain language items are neutral in an ICS context and to which they may

potentially affect the reading experience, making it harder or easier for assessors to evaluate impact claims.

Evaluative or persuasive editorial language choices: Differences in language can be the result of deliberate linguistic choices that are aimed at persuading the reader of the merits of the impact claims and providing evaluation. Such persuasive language could arise in either an overt way, such as the use of boosting adjectives, or in ways that are genre-specific and difficult to detect by anyone not familiar with ICS or with the topic of a given ICS.

A correlation between the use of persuasive language and higher scores could be considered as supporting the view that the skill of "selling" impact linguistically in the ICS is unduly influencing the scoring of ICS. This would be problematic, as the exercise is designed to assess research impact, not the persuasive skills of ICS writers. However, even if a correlation between score and persuasiveness of ICS writing were to be found, on its own it would not provide proof of a causal link. For example, universities or departments achieving high scores in the impact component of REF may also have benefitted from ICS writers who are skilled in persuasive language, for example because there is a correlation between financial resources, high impact and professional writing support, without the persuasion itself actually impacting the score. This is consistent with Williams *et al.*'s (2023: 9) finding that institution is one of the most reliable predictors of score.

8.2 Implications

The format of case studies for assessing the impact of academic research has been criticised in three ways:

(1) they are resource intensive to produce (e.g. Morgan Jones *et al.* 2022: 735);
(2) different types of impact are difficult to compare (e.g. Wilsdon *et al.* 2015: 139); and
(3) language might be a confounding factor in the assessment (e.g. Watermeyer 2019: 80).

This study focuses on the third of these critiques and is largely silent on the other two factors. In respect to the impact of language on scoring, it has been claimed that higher marks can be achieved by those who are best able to "sell" their impact (e.g.Watermeyer and Hedgecoe 2016), making the assessment process a tournament in persuasive linguistics rather than a fair assessment of the positive impact that academic research has on wider society. For example, Hyland and Jiang (2023: 2) assert that ICS were "hyped" to an extent that the "usefulness and reliability" of assessment was in danger; Brauer *et al.* (2019: 66)

lament the requirement for researchers to "boast about their research"; and Smith *et al.* (2020: 38) call it "likely that case studies may perform better than others simply because they are written more persuasively". It is based on such assumptions, rather than on empirical evidence, that some commentators provide advice on the linguistic style that should be used in ICS to enhance the probability of success, such as to "avoid the prosaic yet resist the florid" (Watermeyer and Hedgecoe 2016: 6).

While my findings may be useful in guiding the construction and writing of ICS in future REFs, it is important to recognise that it is not possible to anticipate how future REF panels will construct and interpret guidance and evaluate ICS. Nevertheless, with the level of linguistic granularity provided, it may be possible to draw lessons from my work that can be helpful for training or guidance for academics, professional support staff and impact consultants, addressing gaps in linguistic expertise.

This could include explicitly setting out the ways in which ICS are a separate register from other academic writing, to address the potential pitfall of transferring writing habits and assumptions from research articles to ICS writing. For example, high-scoring ICS 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). By contrast, low-scoring ICS were more likely to contain filler phrases that could be described as "academese" (Biber 2019: 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/or paragraph headings.

On a more granular level, there is evidence that high-scoring ICS had more explicit causal connections between ideas and more logical connective words (*and, or, but*) than low-scoring ICS, and therefore such training or guidance could include information on the effective use of connectors to increase the causal and logical connectivity of research-to-impact claims. This also shows that the core of the ICS register is not about reading ease: while high-scoring ICS in two Main Panels (out of the three that could be analysed in this way) were significantly easier to read, both high- and low-scoring ICS tended to be of "graduate" (Hartley 2016) difficulty.

On a more general level, there are some important points to note for those drawing conclusions about ICS and their role in REF. One is that many language differences are not due to editorial choice but are pre-determined by the content of the ICS, as described in section 6.2.2 and summarised in section 8.1.3. A related conclusion is that it is not possible to "up-polish" an ICS with lesser impact, but it is very possible to miss out on a high rating for an ICS with outstanding impact by not being clear and explicit enough. In other words, to the extent that assessors take into account only clearly demonstrated impact that is evidenced and that is demonstrated to have originated from the university's research when settling on a specific score, it is possible to "undersell" impact using editorial choices, but it is not possible to "oversell" impact by the use of vague claims or assertions that cannot be evidenced.

There is some editorial choice and power about the use of vague language instead of precise language, as well as a focus of attention on the impact of research or pathways of impact. However, this choice only exists for writers of ICS where specific impact has been achieved and has been possible to evidence. For writers of ICS where little impact has emerged or can be evidenced, the equivalent editorial choice to use this level of specificity does not exist, and vague assertions may be all that can be truthfully made and evidenced. As such, the confounding factor of language on the assessment is limited: if ICS that do not demonstrate significance, reach and attribution of impact are not scored highly, this is to be expected.

8.3 Strengths, limitations and further research

This study has several strengths. One unique feature is the whole-framework approach I applied in the language-based analyses, which enabled a switch in perspective from simply describing differences to putting those differences in context quantitatively. I outlined possible differences and then described the subset of those where a difference was detected, showing that the differences are much lesser than they could have been. This is important because it enabled me to argue that the similarities between high- and low-scoring ICS outweigh the differences, and that the differences are not significant or pervasive enough to have influenced assessment outcomes.

Another strength is that I used a selection of different methods to achieve a comprehensive investigation from different angles. This is important because it contributed to the overall confidence in my conclusions that language differences were not a deciding factor in the assessment process, because with this combination of methods, it is less likely that I missed

significant differences than if I had applied only one method of analysis. In particular, the combination of quantitative and qualitative approaches, and the rigour with which I designed and applied the methods, increased the confidence with which conclusions could be drawn.

A further strength is that my comprehensive linguistic exploration was applied to a sample that included both high- and low-scoring ICS. Previous studies either included high-scoring ICS only (e.g. Gow and Redwood 2020) or did not take scores into account in their sample selection at all (e.g. Hyland and Jiang 2023). This prevents these studies from making convincing claims regarding a relationship (whether causal or not) between scores and presentation. The present study addresses this methodological caveat and provides the foundation for linking certain features of language, or content as expressed through certain language use, to assessment outcomes.

Alongside its strengths as a comprehensive linguistic study, this thesis naturally has limitations. One limitation is that it was not possible to access "behind the scenes" observations of how the assessors themselves read and interpreted the submissions. Ethnographic or observational research could address this limitation in future research; however, given the high-stakes nature of this process and the confidentiality of the discussion and the outcome, such access is exceptionally restricted. One example of an observation-based study is Derrick (2018), but it is not possible to match her findings to mine because the ICS in the linguistics sample are a different selection to those whose discussion Derrick investigated.

Another limitation, which arises from the study's nature as a linguistic inquiry, is that the thesis is largely silent on other critiques that are levelled against ICS, such as the resources invested across the sector into preparing case studies. It does not suggest alternative forms of assessment either. It does, however, provide empirical data that can alleviate concerns about the integrity of the exercise based on language, and therefore could be used to endorse the case study narrative approach to research impact assessment.

Moreover, especially for the quantitative components of the study, a larger sample would naturally have been desirable. This was not possible due to the way in which REF results are reported, as described in section 4.1.1. The sample represents less than 3% of the total number of ICS submitted to REF2014. Although the number of ICS was fairly evenly balanced

between Main Panels in samples B and C, the sample only included a selection of Units of Assessment from each Main Panel, where sufficient numbers of high- and low-scoring ICS could be identified (20 and 14 out of 36 Units of Assessment in samples A and B respectively). As such, caution should be taken when generalising from these findings. Finally, it is important to note that the high-scoring component of the sample is not only limited to 4* ICS, but is even restricted to those that were part of a submission where all ICS received the top score. It is therefore not representative of 4* impact across the sector, but at best representative of submissions with an exceptionally high level of success in the exercise. This may be due to high-magnitude impacts in the ICS that these units submitted, or to a generally high level of impact activity in those units and therefore a greater range of options to select from, or indeed to resources that could be allocated to articulating and evidencing impact. However, these caveats were necessary in order to isolate any ICS that could definitively be identified as having received a 4* rating.

A further notable feature of my sample points to directions for future research. In 2013, when ICS were first written, the only sector-wide guidance that was available were the template and the official guidelines. There was no precedent of how these texts could be structured within the template, and what language and style would be most appropriate or most effective. Universities had to complete and submit these texts at the same time, the texts were then used for assessment at the same time, and later released to the public. This constitutes the synchronous but double-blind emergence of a new genre, as described by Wróblewska (2021), and of a new register associated with this genre. Difficult as this was for those in charge of preparing ICS for REF2014, it offers linguists a unique opportunity to study how a new register is negotiated. While the present study only examines the first iteration of producing ICS, it will be possible to extend this to the 2021 and 2029 submissions and trace the development. It would be especially interesting to investigate whether there is a convergence in preferred phrases, that is, fewer differences between high- and low-scoring corpora and higher relative frequencies of certain wordings that are characteristic of high-scoring REF2014 ICS. However, any convergence that may be detected in datasets from later REFs may potentially have been influenced by findings from my analyses, as these were published in Reichard *et al.* (2020) a year before the submission of REF2021 ICS, an article that was downloaded over 10,000 times in its first year. Therefore, the development of this emergent register may not have been as natural as it might have been.

References

Adams, R.J., P. Smart and A.S. Huff. 2017. 'Shades of grey: guidelines for working with the grey literature in systematic reviews for Management and Organizational Studies'. *International Journal of Management Reviews* 19 (4): 432-454.

Afros, E. and C.F. Schryer. 2009. 'Promotional (meta)discourse in research articles in language and literary studies'. *English for Specific Purposes* 28: 58-68.

Anthony, L. 2014. *AntConc*. Version 3.4.4. Tokyo: Waseda University. http://www.laurenceanthony.net/, last accessed 11/07/2017.

---. 2017. *AntFileConverter*. Version 1.2.1. Tokyo: Waseda University. www.laurenceanthony.net, last accessed 11/07/2017.

ApacheOpenNLP Development Community. 'Apache OpenNLP Developer Documentation', https://opennlp.apache.org/docs/2.5.0/manual/opennlp.html#tools.sentdetect.detection, last accessed 19/11/2024.

Auerbach, C.F. and L.B. Silverstein. 2003. *Qualitative Data: An Introduction to Coding and Analyzing Data in Qualitative Research*. New York: New York University Press.

Baguley, T. 2017. 'Psychology Research Excellence Framework 2014 Impact Analysis'. British Psychological Society.

Bandola-Gill, J. and K.E. Smith. 2022. 'Governing by narratives: REF impact case studies and restrictive storytelling in performance measurement'. *Studies in Higher Education* 47 (9): 1857-1871.

Bayley, J. 2023. *Creating Meaningful Impact: The Essential Guide to Developing an Impact-Literate Mindset*. Bingley: Emerald.

Becher, T. and P. Trowler. 2001. *Academic Tribes and Territories: Intellectual Enquiry and the Cultures of Disciplines*. 2nd ed. Buckingham: The Society for Research into Higher Education and Open University Press.

Belcher, B., D. Suryadarma and A. Halimanjaya. 2017. 'Evaluating policy-relevant research: lessons from a series of theory-based outcomes assessments'. *Humanities and Social Sciences Communications* 3 (1): 1-16.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

---. 1990. 'Methodological issues regarding corpus-based analyses of linguistic variation'. *Literary and Linguistic Computing* 5 (4): 257-269.

---. 1993. 'Representativeness in corpus design'. *Literary and Linguistic Computing* 8 (4): 243-257.

---. 2006. *University Language: a Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

---. 2019. 'Text-linguistic approaches to register variation'. *Register Studies* 1 (1): 42-75.

Biber, D. and S. Conrad. 2004. 'Corpus-based comparison of registers'. In C. Coffin, A. Hewings, and K. O'Halloran (eds) *Applying English Grammar: Functional and Corpus Approaches*. London: Arnold. 40-56.

---. 2019. *Register, Genre and Style*. 2nd ed. Cambridge: Cambridge University Press.

Biber, D. and B. Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biber, D. and M. Zhang. 2018. 'Expressing evaluation without grammatical stance: informational persuasion on the web'. *Corpora* 13 (1): 97-123.

Biri, D., K. Oliver and A. Cooper. 2014. 'What is the impact of BEAMS research? An evaluation of REF impact case studies from UCL BEAMS'. Department of Science, Technology, Engineering, and Public Policy, University College London.

BNC Consortium 2007. *The British National Corpus*.

Bonaccorsi, A., F. Chiarello and G. Fantoni. 2021. 'Impact for whom? Mapping the users of public research with lexicon-based text mining'. *Scientometrics* 126: 1745–1774.

Bornmann, L., R. Haunschild and J. Adams. 2019. 'Do altmetrics assess societal impact in a comparable way to case studies? An empirical test of the convergent validity of altmetrics based on data from the UK research excellence framework (REF)'. *Journal of Informetrics* 13 (1): 325-340.

Brauer, R., M. Dymitrow and J. Tribe. 2019. 'The impact of tourism research'. *Annals of Tourism Research* 77: 64-78.

Braun, V. and V. Clarke. 2006. 'Using thematic analysis in psychology'. *Qualitative Research in Psychology* 3 (2): 77–101.

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Brezina, V., T. McEnery and S. Wattam. 2015. 'Collocations in context: A new perspective on collocation networks'. *International Journal of Corpus Linguistics* 20 (2): 139-173.

Brook, L. 2018. 'Evidencing impact from art research: analysis of impact case studies from the REF 2014'. *The Journal of Arts Management, Law, and Society* 48 (1): 57-69.

Burrows, R. 2012. 'Living with the H-Index? Metric assemblages in the contemporary academy'. *The Sociological Review* 60 (2): 355-372.

Cain, T. and D. Allan. 2017. 'The invisible impact of educational research'. *Oxford Review of Education* 43 (6): 718-732.

Caplan, N.A. 2021. 'L2 writing and language learning in academic settings'. In R.M. Manchón and C. Polio (eds) *The Routledge Handbook of Second Language Acquisition and Writing*. New York: Routledge. 268-281.

Carretero, M. and M. Taboada. 2014. 'Graduation within the scope of Attitude in English and Spanish consumer reviews of books and movies'. In G. Thompson and L. Alba-Juez (eds) *Evaluation in Context*. Amsterdam and Philadelphia: John Benjamins. 221-239.

Chowdhury, G., K. Koya and P. Philipson. 2016. 'Measuring the impact of research: Lessons from the UK's Research Excellence Framework 2014'. *PLoS One* 11 (6): e0156978-e0156978.

Chubb, J. 2017. 'Instrumentalism and epistemic responsibility: Researchers and the impact agenda in the UK and Australia'. Unpublished PhD dissertation, University of York.

Chubb, J. and M.S. Reed. 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 Communications* 3 (1).

---. 2018. 'The politics of research impact: academic perceptions of the implications for research funding, motivation and quality'. *British Politics* 13 (3): 295-311.

Coleman, I. 2019. 'The evolution of impact support in UK universities'. Cactus Communications.

Collett, S. 2024. 'What is the value of four star REF outputs and impact case studies?'. *LSE Impact Blog*. https://blogs.lse.ac.uk/impactofsocialsciences/2024/04/08/what-is-the-value-of-four-star-ref-outputs-and-impact-case-studies/, last accessed 02/05/2025.

Congleton, R.D., A. Marsella and A.J. Cardazzi. 2022. 'Readership and citations as alternative measures of impact'. *Constitutional Political Economy* 33 (1): 100-114.

Conrad, S. and D. Biber. 2000. 'Adverbial marking of stance in speech and writing'. In S. Hunston and G. Thompson (eds) *Evaluation in Text: Authorial Stance and the Construction fo Discourse*. Oxford: Oxford University Press. 56-73.

Copley, J. 2018. 'Providing evidence of impact from public engagement with research: A case study from the UK's Research Excellence Framework (REF)'. *Research for All* 2 (2): 230-243.

Crossley, S.A., R. Roscoe and D.S. McNamara. 2014. 'What is successful writing? An investigation into the multiple ways writers can write successful essays'. *Written Communication* 31 (2): 184-214.

Curry, S., E. Gadd and J. Wilsdon. 2022. 'Harnessing the Metric Tide: indicators, infrastructures & priorities for UK responsible research assessment. Report of The Metric Tide Revisited panel'.

Department of Education. 2006. 'Research Quality Framework Assessing the Quality and Impact of Research in Australia: the Recommended RQF'. Canberra: Deptartment of Education, Science and Training. https://webarchive.nla.gov.au/awa/20070926040852/http://pandora.nla.gov.au/pan/76905

/20070926-1259/www.dest.gov.au/NR/rdonlyres/E32ECC65-05C0-4041-A2B8-75ADEC69E159/RecommendedRQF2.pdf, last accessed 14/03/2024.

Derrick, G. 2018. *The Evaluators' Eye: Impact Assessment and Academic Peer Review*. Cham: Palgrave Macmillan.

Derrick, G., I. Meijer and E. van Wijk. 2014. 'Unwrapping "impact" for evaluation: A co-word analysis of the UK REF2014 policy documents using VOSviewer'. *Proceedings of the Science and Technology Indicators Conference*. 145-154.

Derrick, G. and G. Samuel. 2017. 'The future of societal impact assessment using peer review: pre-evaluation training, consensus building and inter-reviewer reliability'. *Palgrave Communications* 3 (1): 17040.

---. 2018. 'Exploring the degree of delegated authority for the peer review of societal impact'. *Science and Public Policy* 45 (5): 673-682.

Desagulier, G. 2017. *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Cham: Springer.

Dontcheva-Navratilova, O. 2020. 'Persuasion: definition, approaches, contexts'. In O. Dontcheva-Navratilova *et al.* (eds) *Persuasion in Specialised Discourses*. Cham: Springer International Publishing. 1-38.

Dotti, N.F. and J. Walczyk. 2022. 'What is the societal impact of university research? A policy-oriented review to map approaches, identify monitoring methods and success factors'. *Evaluation and Program Planning* 95: 102157.

DTZ Consulting and Research. 2006. 'Research Councils UK: Analysis of the external costs of peer review'. London: RCUK.

Duncan, S. and P. Manners. 2017. 'Engaging Publics with Research: Reviewing the REF impact case studies and templates: Executive summary'. Bristol: National Co-ordinating Centre for Public Engagement.

Dunlop, C.A. 2018. 'The political economy of politics and international studies impact: REF2014 case analysis'. *British Politics* 13 (3): 270-294.

Dunning, T. 1993. 'Accurate methods for the statistics of surprise and coincidence'. *Computational Linguistics* 19 (1): 61-74.

Durrant, P. 2017. 'Lexical bundles and disciplinary variation in university students' writing: mapping the territories'. *Applied Linguistics* 38 (2): 165-193.

Finegan, E. 2019. 'Afterword'. *Register Studies* 1 (1): 199-208.

Flowerdew, L. 2004. 'The argument for using English specialized corpora to understand academic and professional settings'. In U. Connor and T.A. Upton (eds) *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins. 11-33.

Friginal, E. and J.A. Hardy. 2014. *Corpus-Based Sociolinguistics: A Guide for Students*. New York: Routledge.

Fulcher, G. 2010. *Practical Language Testing*. London: Hodder Education.

Fuoli, M. 2015. 'A step-wise method for annotating appraisal'. *Functions of Language* 25 (2): 1-26.

Fuoli, M. and C. Hommerberg. 2015. 'Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions.'. *Corpora* 10 (3): 315-349.

Gablasova, D., V. Brezina and T. McEnery. 2017. 'Collocations in corpus-based language learning research: identifying, comparing, and interpreting the evidence'. *Language Learning* 67 (S1): 155-179.

Gardner, S., H. Nesi and D. Biber. 2018. 'Discipline, level, genre: integrating situational perspectives in a new MD analysis of university student writing'. *Applied Linguistics* 40 (4): 646-674.

Gibbons, M., C. Limoges, H. Nowotny, S. Schwartzman, P. Scott and M. Trow. 1994. *The New Production of Knowledge: the Dynamics of Science and Research in Contemporary Societies*. London: Sage.

Gow, J. and H. Redwood. 2020. *Impact in International Affairs: The Quest for World-Leading Research*. London: Routledge.

Graesser, A.C., D.S. McNamara and J. Kulikowich. 2011. 'Coh-Metrix: Providing multi-level analyses of text characteristics'. *Educational Researcher* 40: 223-234.

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

Grant, J., P.-B. Brutscher, S.E. Kirk, L. Butler and S. Wooding. 2010. 'Capturing research impacts: A review of international practice'. Cambridge: HEFCE. http://www.rand.org/content/dam/rand/pubs/documented_briefings/2010/RAND_DB578.pdf, last accessed 09/05/2025.

Grant, W.J. 2023. 'The knowledge deficit model and science communication'. *Oxford Research Encyclopedia of Communication*: Oxford University Press. https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-1396, last accessed 08/04/2025.

Gray, B. 2015. *Linguistic variation in research articles: When discipline tells only part of the story*. Amsterdam: John Benjamins.

Gray, B. and J. Egbert. 2019. 'Editorial: Register and register variation'. *Register Studies* 1 (1): 1-9.

Greenhalgh, T. and N. Fahy. 2015. 'Research impact in the community-based health sciences: an analysis of 162 case studies from the 2014 UK Research Excellence Framework'. *BMC Medicine* 13 (1): 232-232.

Gries, S.T. 2008. 'Phraseology and linguistics theory'. In S. Granger and G. Mernier (eds) *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins. 3-25.

Haberlandt, K.F. and A.C. Graesser. 1985. 'Component processes in text comprehension and some of their interactions'. *Journal of Experimental Psychology: General* 114 (3): 357-374.

Hanna, C.R., L.P. Gatting, K.A. Boyd, K.A. Robb and R.J. Jones. 2020. 'Evidencing the impact of cancer trials: insights from the 2014 UK Research Excellence Framework'. *Trials* 21 (1): 486-486.

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. 2010. 'Units of assessment and recruitment of expert panels'. HEFCE.

---. 2011. 'Assessment framework and guidance on submissions'. HEFCE.

---. 2015. 'Impact Database Licence Arrangements', http://www.hefce.ac.uk/rsrch/REFimpact/licence/, last accessed 15/03/2017.

---. 2015b. 'Research Excellence Framework 2014: Overview report by Main Panel A and Sub-panels 1 to 6'. HEFCE. https://2014.ref.ac.uk/media/ref/content/expanel/member/Main%20Panel%20A%20overview%20report.pdf, last accessed 09/05/2025.

Hill, S. 2016. 'Assessing (for) impact: future assessment of the societal impact of research'. *Palgrave Communications* 2 (1): 16073.

Hinrichs, S. and J. Grant. 2015. 'A new resource for identifying and assessing the impacts of research'. *BMC Medicine* 13 (1): 148.

Hinrichs, S., A. Kamenetzky, L. Borjes and J. Grant. 2015. 'The non-academic impact of international development research in UK Higher Education'. The Policy Institute at Kings'.

Ho, D. 2016. *Notepad++*. 7.3.3. https://notepad-plus-plus.org/. accessed 08/05/2025.

Ho, S.Y.E. and P. Crosthwaite. 2018. 'Exploring stance in the manifestos of 3 candidates for the Hong Kong Chief Executive election 2017: Combining CDA and corpus-like insights'. *Discourse & Society* 29 (6): 629-654.

Hogan, J.M. 2012. 'Persuasion in the rhetorical tradition'. In J.P. Dillard and L. Shen (eds) *The SAGE Handbook of Persuasion: Developments in Theory and Practice* 2nd ed. Thousand Oaks: Sage. 2-19.

Hommerberg, C. and A. Don. 2015. 'Appraisal and the language of wine appreciation: A critical discussion of the potential of the Appraisal framework as a tool to analyse specialised genres'. *Functions of Language* 22 (2): 161-191.

Hood, S. 2010. *Appraising Research: Evaluation in Academic Writing*. London: Palgrave Macmillan.

---. 2019. 'Appraisal'. In D. Schönthal *et al.* (eds) *The Cambridge Handbook of Systemic Functional Linguistics*, Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press. 382-409.

Hughes, T., D. Webber and N. O'Regan. 2019. 'Achieving wider impact in business and management: Analysing the case studies from REF 2014'. *Studies in Higher Education* 44 (4): 628-642.

Humphrey, S.L. and D. Economou. 2015. 'Peeling the onion – A textual model of critical analysis'. *Journal of English for Academic Purposes* 17: 37-50.

Hunston, S. 1994. 'Evaluation and organization in a sample of written academic discourse'. In M. Coulthard (ed.) *Advances in Written Text Analysis*. London: Routledge. 191–218.

---. 2011. *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York: Routledge.

Hyland, K. 1998. 'Persuasion and context: The pragmatics of academic metadiscourse'. *Journal of Pragmatics* 30 (4): 437-455.

---. 2002. 'Directives: Argument and engagement in academic writing'. *Applied Linguistics* 23 (2): 215-239.

---. 2005a. *Metadiscourse*. London: Continuum.

---. 2005b. 'Stance and engagement: A model of interaction in academic discourse'. *Discourse Studies* 7 (2): 173-192.

Hyland, K. and F.K. Jiang. 2021. '"Our striking results demonstrate …": Persuasion and the growth of academic hype'. *Journal of Pragmatics* 182: 189–202.

---. 2023. 'Hyping the REF: promotional elements in impact submissions'. *Higher Education* 87: 685-702.

Hyland, K. and C. Sancho Guinda. 2012. *Stance and Voice in Written Academic Genres*. London: Palgrave Macmillan.

Intellectual Property Office. 'Exceptions to Copyright', https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research, last accessed 13/03/2017.

Jones, M.M., S. Castle-Clarke, C. Manville, S. Gunashekar and J. Grant. 2013. 'Assessing Research Impact: An international review of the Excellence in Innovation for Australia Trial'. Cambridge: RAND Europe. http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR278/RAND_RR278.pdf, last accessed 09/05/2025.

Kellard, N. and M. Śliwa. 2016. 'Business and Management impact assessment in REF2014: analysis and reflection'. *British Journal of Management* 27 (4): 693-711.

Kelleher, L. and A. Zecharia. 2020. 'A Triple Helix systems perspective of UK drug discovery and development: A systematic review of REF impact case studies'. *Industry and Higher Education*: 0950422220969349.

Kelly, D., B. Kent, A. McMahon, J. Taylor and M. Traynor. 2016. 'Impact case studies submitted to REF 2014: The hidden impact of nursing research'. *Journal of Research in Nursing* 21 (4): 256-268.

Kerrigan, S. and J. Callaghan. 2018. 'The impact of filmmaking research'. *Media Practice and Education* 19 (3): 229-242.

Khazragui, H. and J. Hudson. 2014. 'Measuring the benefits of university research: impact and the REF in the UK'. *Research Evaluation* 24 (1): 51-62.

Koester, A. 2010. 'Building small specialised corpora'. In A. O'Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*. London: Routledge. 66-79.

Kousha, K. and M. Thelwall. 2025. 'Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations'. *Journal of the Association for Information Science and Technology*: 1-17.

Koya, K. and G. Chowdhury. 2020. 'Measuring impact of academic research in Computer and Information Science on society'. *Proceedings of the 2020 2nd Asia Pacific Information Technology Conference*. 78-85.

Larsson, T. 2019. 'Grammatical stance marking across registers: Revisiting the formal-informal dichotomy'. *Register Studies* 1 (2): 243-268.

Lei, L. 2021. *Conducting Sentiment Analysis*. Cambridge: Cambridge University Press.

Levshina, N. 2015. *How to do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.

Liardét, C.L. and S. Black. 2019. '"So and so" says, states and argues: A corpus-assisted engagement analysis of reporting verbs'. *Journal of Second Language Writing* 44: 37–50.

Lim, M.A. 2020. 'Impact case studies: what accounts for the need for numbers in impact evaluation?'. *International Studies in Sociology of Education* 29 (1-2): 107-125.

Loach, T., J. Adams and M. Szomszor. 2016. 'Digital Research Report: The Societal and Economic Impacts of Academic Research – International perspectives on good practice and managing evidence'. London: Digital Science.

MacDonald, R. 2017. '"Impact", research and slaying Zombies: the pressures and possibilities of the REF'. *International Journal of Sociology and Social Policy* 37 (11-12): 696-710.

Machen, R. 2020. 'Critical research impact: On making space for alternatives'. *Area* 52 (2): 329-341.

Manville, C., S. Guthrie, M.-L. Henham, B. Garrod, S. Sousa, A. Kirtley, S. Castle-Clarke and T. Ling. 2015a. 'Assessing impact submissions for REF2014: An evaluation'. Cambridge: RAND Europe.

Manville, C., M.M. Jones, M. Frearson, S. Castle-Clarke, M.-L. Henham, S. Gunashekar and J. Grant. 2015b. 'Preparing impact submissions for REF 2014: An evaluation. Findings and observations'. Cambrigde: RAND Europe.

Martin, J.R. and P.R.R. White. 2005. *The Language of Evaluation: Appraisal in English*. Basingstoke: Palgrave Macmillan.

Matthiessen, C.M. 2019. 'Register in Systemic Functional Linguistics'. *Register Studies* 1 (1): 10-41.

Mauranen, A. 2006. 'Speaking the discipline: Discourse and socialisation in ELF and L1 English'. In K. Hyland and M. Bondi (eds) *Academic Discourse across Disciplines*. Bern: Peter Lang. 271-294.

McEnery, T. and A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T., R. Xiao and Y. Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. Abingdon: Routledge.

McKenna, H.P. 2021. *Research Impact: Guidance on Advancement, Achievement and Assessment*. Cham: Springer.

McNamara, D.S., A.C. Graesser and M.M. Louwerse. 2012. 'Sources of text difficulty: Across genres and grades'. In J.P. Sabatini, E. Albro, and T. O'Reilly (eds) *Measuring Up: Advances in How we Assess Reading Ability*. Plymouth: Rowman & Littlefield Education. 89-116.

McNamara, D.S., A.C. Graesser, P.M. McCarthy and Z. Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York: Cambridge University Press.

Meagher, L.R. and U. Martin. 2017. 'Slightly dirty maths: The richly textured mechanisms of impact'. *Research Evaluation* 26 (1): 15-27.

Messick, S. 1989. 'Validity'. In R.L. Linn (ed.) *Educational Measurement* 3rd ed. New York: Macmillan. 13–103.

Midmore, P. 2017. 'The science of impact and the impact of agricultural science'. *Journal of Agricultural Economics* 68 (3): 611-631.

Moon, K. and D. Blackman. 2014. 'A guide to understanding social science research for natural scientists'. *Conservation Biology* 28 (5): 1167-77.

Moran, C.R. and C.S. Browning. 2018. 'REF impact and the discipline of politics and international studies'. *British Politics* 13 (3): 249-269.

Morgan Jones, M., C. Manville and J. Chataway. 2022. 'Learning from the UK's research impact assessment exercise: a case study of a retrospective impact assessment exercise and questions for the future'. *The Journal of Technology Transfer* 47 (3): 722-746.

Morton, S. 2015. 'Progressing research impact assessment: A "contributions" approach'. *Research Evaluation* 24 (4): 405-419.

Nesta. 2018. 'Seven Principles for Public Engagement in Research and Innovation Policymaking'. https://www.nesta.org.uk/report/seven-principles-public-engagement-research-and-innovation-policymaking/, last accessed 08/05/2025.

Nini, A. 2015. *Multidimensional Analysis Tagger*. Version 1.3. http://sites.google.com/site/mulitidimensionaltagger, last accessed 01/04/2017.

Nowotny, H., P. Scott and M. Gibbons. 2003. 'Introduction. "Mode 2" revisited: The new production of knowledge'. *Minerva* 41 (3): 179-194.

O'Donnell, M. 2008. 'The UAM CorpusTool: Software for corpus annotation and exploration'. In C.M. Bretones Callejas *et al.* (eds) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería. 1433-1447.

Oakes, M.P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Oancea, A. 2010. 'The BERA / UCET Review of the Impacts of RAE 2008 on Education Research in UK Higher Education Institutions'. Macclesfield: BERA/UCET.

---. 2013. 'Interpretations of research impact in seven disciplines'. *European Educational Research Journal* 12 (2): 242-250.

---. 2014. 'Research assessment as governance technology in the United Kingdom: findings from a survey of RAE 2008 impacts'. *Zeitschrift für Erziehungswissenschaft* 17 (S6): 83-110.

Olssen, M. and M.A. Peters. 2005. 'Neoliberalism, higher education and the knowledge economy: from the free market to knowledge capitalism'. *Journal of Education Policy* 20 (3): 313-345.

Osborne, C. 2022. 'Towards impact: best practice in community and stakeholder engagement'. In W. Kelly (ed.) *The Impactful Academic*. Bingley: Emerald. 69-82.

Ovseiko, P.V., A. Oancea and A.M. Buchan. 2012. 'Assessing research impact in academic clinical medicine: a study using Research Excellence Framework pilot impact indicators'. *BMC Health Services Research* 12 (1): 478-478.

Parks, S., B. Ioppolo, M. Stepanek and S. Gunashekar. 2018. 'Guidance for standardising quantitative indicators of impact within REF case studies'. Cambridge: RAND.

Penfield, T., M.J. Baker, R. Scoble and M.C. Wykes. 2014. 'Assessment, evaluations, and definitions of research impact: A review'. *Research Evaluation* 23 (1): 21-32.

Pepe, A. and M.J. Kurtz. 2012. 'A measure of total research impact independent of time and discipline'. *PLoS One* 7 (11): e46428.

Perloff, R.M. 2014. *The Dynamics of Persuasion: Communication and Attitudes in the 21st Century*. 5th ed. New York: Routledge.

Phillips, P.A., S.J. Page and J. Sebu. 2020. 'Achieving research impact in tourism: Modelling and evaluating outcomes from the UKs Research Excellence Framework'. *Tourism Management* 78: 104072.

Pidd, M. and J. Broadbent. 2015. 'Business and Management Studies in the 2014 Research Excellence Framework'. *British Journal of Management* 26: 569-581.

Pinar, M. and T.J. Horne. 2022. 'Assessing research excellence: Evaluating the Research Excellence Framework'. *Research Evaluation* 31 (2): 173-187.

Pullinger, R. and O. Varley-Winter. 2017. 'The impact of academic statistics as shown through “impact case studies” submitted to the 2014 REF'. Royal Statistical Society.

Ravenscroft, J., M. Liakata, A. Clare and D. Duma. 2017. 'Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements'. *PLoS One* 12 (3): e0173152.

Rayson, P., D. Berridge and B. Francis. 2004. 'Extending the Cochran rule for the comparison of word frequencies between corpora'. In G. Purnelle, C. Fairon, and A. Dister (eds), *Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004)*, II; Louvain-la-Neuve, Belgium: Presses universitaires de Louvain. 926-936.

Rayson, P. and R. Garside. 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. 1-6.

Reed, M.S. 2018. *The Research Impact Handbook*. 2nd ed. Huntly, Aberdeenshire: Fast Track Impact.

Reed, M.S., M. Ferre, J. Martin-Ortega, R. Blanche, R. Lawford-Rolfe, M. Dallimer and J. Holden. 2021. 'Evaluating impact from research: A methodological framework'. *Research Policy* 50 (4): 104-147.

Reed, M.S. and S. Kerridge. 2017. 'How much was an impact case study worth in the UK Research Excellence Framework?'. *Fast Track Impact Blog*. https://www.fasttrackimpact.com/post/2017/02/01/how-much-was-an-impact-case-study-worth-in-the-uk-research-excellence-framework, last accessed 02/05/2025.

Reed, M.S., B. Reichard, J. Chubb and G. Hall. 2019. 'What makes a 4* impact case study for REF2021?'. *Fast Track Impact Blog*. https://www.fasttrackimpact.com/post/2017/12/19/what-makes-a-4-research-impact-case-study-for-ref2021, last accessed 08/05/2025.

Reichard, B. 2021. 'Impact in international affairs: the quest for world-leading research'. *International Affairs* 97 (1): 239-240.

---. 2024. 'How to turn a 3* impact case study into 4*'. *Bella Reichard Research Impact Consultant Blog*. https://www.bellareichard.co.uk/post/how-to-turn-a-3-impact-case-study-into-4, last accessed 15/11/2024.

Reichard, B., M.S. Reed, J. Chubb, G. Hall, L. Jowett, A. Peart and A. Whittle. 2020. 'Writing impact case studies: a comparative study of high-scoring and low-scoring case studies from REF2014'. *Palgrave Communications* 6 (1): 1-17.

Research England. 2023. 'Research Excellence Framework 2028: Initial decisions and issues for further consultation'. Research England. https://2029.ref.ac.uk/publication/initial-decisions-on-ref-2028/, last accessed 08/05/2025.

Rickards, L., W. Steele, O. Kokshagina and O. Moraes. 2020. 'Research Impact as Ethos'. RMIT University, Melbourne.

Robbins, P.T., D. Wield and G. Wilson. 2017. 'Mapping engineering and development research excellence in the UK: an analysis of REF2014 impact case studies'. *Journal of International Development* 29 (1): 89-105.

Russell, A. and S. Lewis. 2015. 'Documenting impact: an impact case study of anthropological collaboration in tobacco control'. *Anthropology in Action* 22 (2): 14-23.

Saldana, J. 2009. *The Coding Manual for Qualitative Researchers*. Thousand Oaks: Sage.

Scott, M. 1997. 'PC analysis of key words — And key key words'. *System* 25 (2): 233-245.

Shore, C. and S. Wright. 2015. 'Governing by numbers: Audit culture, rankings and the new world order'. *Social Anthropology/Anthropologie sociale* 23 (1): 22-28.

Simpson, B. 2015. 'REF 2014 and impact: reading the runes for anthropology in action'. *Anthropology in Action* 22 (2): 1-4.

Sinclair, J. 2004. 'Corpus and text - basic principles'. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. Online resource.

Śliwa, M. and N. Kellard. 2022. *The Research Impact Agenda: Navigating the Impact of Impact*. Oxford: Routledge.

Smith, K., J. Bandola-Gill, N. Meer, E. Stewart and R. Watermeyer. 2020. *The Impact Agenda: Controversies, Consequences and Challenges*. Bristol: Policy Press.

Smith, K. and E. Stewart. 2017. 'We need to talk about impact: why social policy academics need to engage with the UK's research impact agenda'. *Journal of Social Policy* 46 (1): 109-127.

Smith, S. 1997. 'Power and truth; a reply to William Wallace'. *Review of International Studies* 23 (4): 507-516.

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

Stern, L.N. 2016. 'Building on Success and Learning from Experience: An Independent Review of the Research Excellence Framework '.

Swales, J. and C. Feak. 2000. *English in Today's Research world: A Writing Guide*. Ann Arbor: The University of Michigan Press.

Swales, J.M. 1990. *Genre Analysis*. Cambridge: Cambridge University Press.

Szmrecsanyi, B. 2019. 'Register in variationist linguistics'. *Register Studies* 1 (1): 76-99.

Tavassoli, F., A. Jalilifar and P.R. White. 2019. 'British newspapers' stance towards the Syrian refugee crisis: An appraisal model study'. *Discourse & Society* 30 (1): 64-84.

Terämä, E., M. Smallman, S.J. Lock, C. Johnson and M.Z. Austwick. 2016. 'Beyond Academia – interrogating research impact in the Research Excellence Framework'. *PLoS One* 11 (12): e0168533.

Thompson, G. 2001. 'Interaction in academic writing: learning to argue with the reader'. *Applied Linguistics* 22: 58-78.

Thompson, P., S. Hunston, A. Murakami and D. Vajn. 2017. 'Multi-dimensional analysis, text constellations, and interdisciplinary discourse'. *International Journal of Corpus Linguistics* 22 (2): 153-186.

Thorpe, A., R. Craig, G. Hadikin and S. Batistic. 2018. 'Semantic tone of research "environment" submissions in the UK's Research Evaluation Framework 2014'. *Research Evaluation* 27 (2): 53-62.

Tognini-Bonelli, E. 2004. 'Working with corpora: issues and insights'. In C. Coffin, A. Hewings, and K. O'Halloran (eds) *Applying English Grammar: Functional and Corpus Approaches*. London: Arnold. 11-24.

Toutanova, K., D. Klein, C. Manning and Y. Singer. 2003. 'Feature-rich part-of-speech tagging with a cyclic dependency network'. *Proceedings of HLT-NAACL 2003*. 252-259.

Tupala, M. 2019. 'Applying quantitative appraisal analysis to the study of institutional discourse: the case of EU migration documents'. *Functional Linguistics* 6 (1): 2.

Tusting, K. 2018. 'The genre regime of research evaluation: contradictory systems of value around academics' writing'. *Language and Education* 32 (6): 477-493.

Van Noorden, R. 2015. 'Seven thousand stories capture impact of science'. *Nature* 518 (7538): 150.

Vold, E.T. 2006. 'Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study'. *International Journal of Applied Linguistics* 16 (1): 61-87.

Wagner, S., C. Rahal, A. Spiers and D. Leasure. 2024. 'The SHAPE of research impact'. London: The British Academy.

Warry, P. 2006. 'Increasing the Economic Impact of the Research Councils (the Warry Report)'. Swindon: Research Council UK.

Watermeyer, R. 2019. *Competitive Accountability in Academic Life: The Struggle for Social Impact and Public Legitimacy*. Cheltenham: Edward Elgar.

Watermeyer, R. and J. Chubb. 2018. 'Evaluating "impact" in the UK's Research Excellence Framework (REF): liminality, looseness and new modalities of scholarly distinction'. *Studies in Higher Education*: 1-13.

Watermeyer, R. and A. Hedgecoe. 2016. 'Selling "impact": peer reviewer projections of what is needed and what counts in REF impact case studies. A retrospective analysis'. *Journal of Education Policy* 31 (5): 651-665.

Watermeyer, R. and M. Tomlinson. 2021. 'Competitive accountability and the dispossession of academic identity: Haunted by an impact phantom'. *Educational Philosophy and Theory*: 1-15.

Weinstein, N., J. Wilsdon, G. Haddock and J. Chubb. 2019. 'The Real-Time REF Review: A Pilot Study to Examine the Feasibility of a Longitudinal Evaluation of Perceptions and Attitudes Towards REF 2021'. Working Paper: University of Cardiff and University of Sheffield. https://eprints.whiterose.ac.uk/id/eprint/147915/, last accessed 08/05/2025.

White, P.R.R. 2021. 'The Language of Attitude, Arguability and Interpersonal Positioning: The Appraisal Website', http://www.languageofevaluation.info/appraisal/, last accessed 07/05/2025.

Williams, B. 2015. 'Sustainable development: the impacts of UK university research'. *Environmental Scientist* 24 (3): 8-12.

Williams, K. 2020. 'Playing the fields: Theorizing research impact and its assessment'. *Research Evaluation* 29 (2): 191-202.

Williams, K. and J. Grant. 2018. 'A comparative review of how the policy and procedures to assess research impact evolved in Australia and the UK'. *Research Evaluation* 27 (2): 93-105.

Williams, K., S. Michalska, E. Cohen, M. Szomszor and J. Grant. 2023. 'Exploring the application of machine learning to expert evaluation of research impact'. *PLoS One* 18 (8): e0288469.

Wilsdon, J., L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill, R. Jones, R. Kain, S. Kerridge, M. Thelwall, J. Tinkler, I. Viney, P. Wouters, J. Hill and B. Johnson. 2015. 'Metric Tide: report of the independent review of the role of metrics in research assessment and management'. HEFCE.

Wooldridge, J. and M.B. King. 2019. 'Altmetric scores: An early indicator of research impact'. *Journal of the Association for Information Science and Technology* 70 (3): 271-282.

Wróblewska, M.N. 2019. 'Impact evaluation in Norway and in the UK: A comparative study, based on REF 2014 and Humeval 2015-2017'. *ENRESSH working paper* 2019.

---. 2021. 'Research impact evaluation and academic discourse'. *Humanities and Social Sciences Communications* 8 (58).

Xu, X. 2017. 'An analysis of stance and voice in Applied Linguistics research articles across Mainland Chinese and British cultures'. Unpublished PhD dissertation, Coventry University.

Xu, X. and H. Nesi. 2019. 'Differences in engagement: A comparison of the strategies used by British and Chinese research article writers'. *Journal of English for Academic Purposes* 38: 121-134.

Yang, A., S.-Y. Zheng and G.-C. Ge. 2015. 'Epistemic modality in English-medium medical research articles: A systemic functional perspective'. *English for Specific Purposes* 38: 1-10.

Yaqub, O., D. Malkov and J. Siepel. 2023. 'How unpredictable is research impact? Evidence from the UK's Research Excellence Framework'. *Research Evaluation* 32 (2): 273-285.

Appendix A: List of impact case studies included in Sample A

The following tables include an overview of all ICS that are included in Sample A, from which sub-samples B and C are mostly drawn. Sample A includes all identifiable 4* ICS from REF2014 (n=124) and all identifiable 1*/2* ICS in those Units of Assessment where 4* ICS could be identified (n=93), resulting in an overall sample of 217 ICS.

Each Unit of Assessment is represented in one table with the score from the official results spreadsheet; the university and impact type taken from the REF2014 impact case study database; the title abbreviation created for this corpus from the official ICS title; a file number that encodes information about each ICS as explained below; and a running number for a unique but simple identifier for each ICS.

File number:

10000 = UoA (e.g. 250000 for UoA25)

1000 = genre (1=Impact case study, 2=Research Article)[25]

100 = score (1=4*, 2=1*/2*)

10 = university (starting with 1 within each UoA)

1 = case study (starting with 1 for each university within each UoA)

The file numbering column includes hyperlinks to each ICS on the REF2014 database.

Main Panel A

Unit of Assessment 1 Clinical Medicine

Score	University	ICS title abbreviation	Impact type[26]	File number	Running number
4*	Bristol	birth	health	11111	1
		bypass	technological	11112	2
		drugdiscovery	technological	11113	3
		fiveaday	health	11114	4
		hip	technological	11115	5
		leukaemia	technological	11116	6
		peptide	technological	11117	7
		potassium	health	11118	8
		trachea	technological	11119	9

[25] This element was included in a pilot phase where a comparison of ICS and research articles was considered, but this was not further pursued.

[26] The FAQs to the database specify that "Case studies are assigned to a single 'Summary Impact Type' by text analysis of the 'Summary of the Impact' (Section 1 of the Impact case study template). This is an indicative guide to aid text searching and is not a definitive assignment of the impact described." (http://impact.ref.ac.uk/CaseStudies/FAQ.aspx#impact)

	Dundee	cardiology	technological	11121	10
		chemicalsafety	political	11122	11
		diabetes	health	11123	12
		filaggrin	health	11124	13
		informatics	health	11125	14
		spironolactone	health	11126	15

Unit of Assessment 2 Public Health, Health Services and Primary Care

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Bristol	cancer	health	21111	16
		cataract	health	21112	17
		cotdeath	health	21113	18
		HIV	health	21114	19
		IRIS	societal	21115	20
		options	health	21116	21
		smoking	health	21117	22
		suicidefall	societal	21118	23

Unit of Assessment 3 Allied Health Professions, Dentistry, Nursing and Pharmacy

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Nottingham	drugscreening	technological	31111	24
		pharmacy	political	31116	25
	Southampton	donation	political	31121	26
		prescribing	health	31122	27
		strokerehab	health	31123	28
		survivors	political	31124	29
2*	Northampton	injuries	technological	31211	1a
		reintegrate	health	31212	2a

Unit of Assessment 4 Psychology, Psychiatry and Neuroscience

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Birkbeck	developmentalresearch	societal	41111	30
		earlyyears	societal	41112	31
		eyetracking	cultural	41113	32
		pregnancy	societal	41114	33
	East London	e-cigarettes	societal	41121	34
		ecstasy	societal	41122	35
	Stirling	EvoFIT	legal	41131	36
		suicide	health	41132	37
	Swansea	alcohol	political	41141	38
		food items	technological	41142	39

2*	Anglia Ruskin	literacy	societal	41211	3a
		stroke	health	41212	4a
	Chichester	colour	health	41221	5a
		diabetes	societal	41222	6a
	Liverpool Hope	chess	societal	41231	7a
		terrorism	political	41232	8a

Unit of Assessment 6 Agriculture, Veterinary and Food Science

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Aberdeen	climate	environmental	61111	40
		rice	political	61112	41
		windfarms	environmental	61113	42
	Warwick	biopesticide	political	61121	43
		footrot	environmental	61122	44
2*	Canterbury	ladybird	environmental	61211	9a
		welfare	societal	61212	10a
	Hertfordshire	agrimanagement	environmental	61221	11a
		mitigation	environmental	61222	12a

Main Panel B

Unit of Assessment 13 Electrical and Electronic Engineering and Metallurgy

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Imperial College London	computing	technological	131113	45
		healthcare	technological	131115	46
		plant	technological	131118	47
		power	political	131119	48
2*	Central Lancashire	retardant	technological	131211	13a
		toxicity	technological	131212	14a

Unit of Assessment 14 Civil and Construction Engineering

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Cardiff	flood	environmental	141111	49
		waste	environmental	141112	50

Main Panel C

Unit of Assessment 18 Economics and Econometrics

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Bristol	heathcaremarket	political	181111	51
		publicservices	political	181112	52
		schoolpolicy	societal	181113	53

Unit of Assessment 20 Law

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Ulster	amnesty	societal	201111	54
		gender	societal	201112	55
		northernireland	societal	201113	56
2*	Bedfordshire	cyberstalking	legal	201211	15a
		humanrights	legal	201212	16a
	Sunderland	localintegrity	legal	201221	17a
		nationalintegrity	legal	201222	18a

Unit of Assessment 22 Social Work and Social Policy

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	LSE	antibiotics	societal	221111	57
		carefinance	societal	221112	58
		childprotection	societal	221113	59
		fuelpoverty	economic	221114	60
		mentalhealth	societal	221115	61
		riots	legal	221116	62
	Oxford	aids	societal	221121	63
		deprivedareas	economic	221122	64
		immigration	political	221123	65
		parenting	societal	221124	66
	UCL	crime	legal	221131	67
		police	legal	221132	68
	York	budgets	political	221141	69
		childsupport	societal	221142	70
		wellbeing	political	221143	71
		workingage	economic	221144	72
2*	Anglia Ruskin	recovery	societal	221211	19a
		selfhelp	societal	221212	20a
	Liverpool Hope	HEpolicy	societal	221221	21a
		radicalpractice	societal	221222	22a
	Sunderland	exclusion	societal	221231	23a
		samesex	societal	221232	24a

Unit of Assessment 23 Sociology

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	York	adviser	societal	231111	73
		biomedical	societal	231112	74
		sickle	societal	231113	75
2*	Abertay	fear	societal	231211	25a
		publiclife	societal	231212	26a
	Leicester	racism	societal	231221	27a
	Winchester	alienation	societal	231231	28a
		cheating	societal	231232	29a

Unit of Assessment 24 Anthropology

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Queens Belfast	assembly	legal	241111	76
		flag	societal	241112	77

Unit of Assessment 25 Education

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Durham	performance	societal	251111	78
		pupilpremium	economic	251112	79
		threshold	societal	251113	80
	Nottingham	leaders	societal	251121	81
		mathematics	societal	251122	82
		vocational	societal	251123	83
	Sheffield	FE	societal	251131	84
		Literacy	societal	251132	85
2*	Anglia Ruskin	nurses	societal	251211	30a
		playful	societal	251212	31a
	Bedfordshire	marginalised	societal	251221	32a
	Birmingham City	creativity	societal	251231	33a
		earlyyearsED	societal	251232	34a
	Brookes	attainment	societal	251241	35a
		leadership	societal	251242	36a
	Chester	multiprofessional	societal	251251	37a
		professionalpractice	societal	251252	38a
	Derby	guidance	societal	251261	39a
		subject	societal	251262	40a
	Newman	pupilexclusion	societal	251271	41a
		safeguarding	societal	251272	42a
	NTU	inclusiveED	societal	251281	43a
		kinaesthetic	societal	251282	44a
		socialinclusion	societal	251283	45a

Score	University	ICS title abbreviation	Impact type	File number	Running number
	Staffordshire	inclusioninED	societal	251291	46a
		schoolperformance	societal	251292	47a
	Twickenham	rethinkingleadership	societal	2512a1	48a
		sustainablesocieties	societal	2512a2	49a
	West London	socio-emotional	societal	2512b1	50a
		wideningparticipation	societal	2512b2	51a

Unit of Assessment 26 Sport and Exercise Sciences, Leisure and Tourism

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Bristol	outdoor	health	261111	86
		travel	societal	261112	87
2*	Abertay	coach	societal	261211	52a
		intensity	societal	261212	53a
	Cumbria	musclefunction	societal	261221	54a
		psychosocial	societal	261222	55a
	Liverpool Hope	cancerous	health	261231	56a
		rugby	societal	261232	57a
	NTU	betaalanine	societal	261241	58a
		heat	societal	261242	59a
	Solent	backpain	societal	261251	60a
		conditioning	societal	261252	61a
	Sunderland	heritage	societal	261261	62a
		migration	societal	261262	63a
	West Scotland	scottishyouth	societal	261271	64a
		steroid	societal	261272	65a
	York St John	hepa	societal	261281	66a
		motivation	societal	261282	67a

Main Panel D

Unit of Assessment 27 Area Studies

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Exeter	Iraq	societal	271111	88
		jihadism	political	271112	89
	LSE	eurozone	economic	271121	90
		greekPM	political	271122	91
		HEfinance	economic	271123	92

Unit of Assessment 28 Modern Languages and Linguistics

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Cardiff University	devolution	societal	281111	93
		mabinogion	cultural	281112	94
2*	Salford	arabia	cultural	281211	68a
		interpreting	societal	281212	69a

Unit of Assessment 29 English Language and Literature

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Bedfordshire	academicenglish	societal	291111	95
		languagetests	societal	291112	96
	Kingston	hampton	cultural	291121	97
		Leveson	political	291122	98
		military	health	291123	99
	Newcastle	diaspora	cultural	291131	100
		poetry	cultural	291132	101
		reddust	cultural	291133	102
		sevenstories	cultural	291134	103
	Swansea	audiences	cultural	291141	104
		chester	cultural	291142	105
		library	cultural	291143	106
2*	Leeds Trinity	aspiring	cultural	291211	70a
		victorian	cultural	291212	71a
	Liverpool Hope	archives	cultural	291221	72a
		shakespeare	societal	291222	73a
	Newman	accident	cultural	291231	74a
		contemporary	societal	291232	75a
	Twickenham	choices	societal	291241	76a
		huxton	cultural	291242	77a

Unit of Assessment 30 History

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Hertfordshire	bailey	cultural	301111	107
		foundling	cultural	301112	108
2*	Sunderland	historical	societal	301211	78a
		mining	cultural	301212	79a
	Twickenham	Palestine	societal	301221	80a
		sharedworld	societal	301222	81a

Unit of Assessment 35 Music, Drama, Dance and Performing Arts

Score	University	ICS title abbreviation	Impact type	File number	Running number
4*	Goldsmith	afghan	cultural	351111	109
		soundscapes	societal	351114	110
	QMUL	brazil	cultural	351121	111
		Hispanic	cultural	351122	112
		performance	cultural	351123	113
	RNCM	anya	cultural	351131	114
		wind	cultural	351132	115
	Southampton	finnissy	cultural	351141	116
		Georgian	cultural	351142	117
		sirens	cultural	351143	118
2*	Bedfordshire	choreographing	societal	351211	82a
		historiography	cultural	351212	83a
	Salford	brass	societal	351221	84a
		jazz	cultural	351222	85a
	West London	record	cultural	351231	86a
		repertoire	cultural	351232	87a

Unit of Assessment 36 Communication, Cultural and Media Studies, Library and Information Management

Score	University	ICS title abbreviation	Impact type	File number	Running number
4	Cardiff	newscoverage	cultural	361111	119
		vegetative	societal	361112	120
	Leicester	disabled	societal	361123	121
		visitors	cultural	361125	122
	LSE	citizen	political	361131	123
		empowering	societal	361132	124
2*	Aberystwyth	digitalassets	cultural	361211	88a
		professionalism	societal	361212	89a
	Brookes	cinema	cultural	361221	90a
		gaydiscourse	societal	361222	91a
	Glyndwr	antidepressant	health	361231	92a
		broadcast	cultural	361232	93a

Appendix B: List of impact case studies included in Sample B

The following tables include an overview of all ICS that are included in Sample B, which is drawn from Sample A but includes a greater balance of ICS across Main Panels, and only those Units of Assessment where both 4* (n=85) and 1*/2* ICS (n=90) could be identified.

Each Unit of Assessment is represented in one table with the score and university; the title abbreviation created for this corpus; and a file number that encodes information about each ICS as explained below.

File number:

10000 = UoA (e.g. 250000 for UoA25)

1000 = genre (1=Impact case study, 2=Research article)[27]

100 = score (1=4*, 2=1*/2*)

10 = university (starting with 1 within each UoA)

1 = case study (starting with 1 for each university within each UoA)

The file numbering column includes hyperlinks to each ICS on the REF2014 database.

Main Panel A

Unit of Assessment 3 Allied Health Professions, Dentistry, Nursing and Pharmacy

Score	University	ICS title abbreviation	File number
4*	Nottingham	drugscreening	31111
		pharmacy	31116
	Southampton	donation	31121
		prescribing	31122
		strokerehab	31123
		survivors	31124
2*	Northampton	injuries	31211
		reintegrate	31212

Unit of Assessment 4 Psychology, Psychiatry and Neuroscience

Score	University	ICS title abbreviation	File number
4*	Birkbeck	developmentalresearch	41111
		earlyyears	41112
		eyetracking	41113
		pregnancy	41114
	East London	e-cigarettes	41121
		ecstasy	41122

[27] This element was included in a pilot phase where a comparison of ICS and research articles was considered, but this was not further pursued.

	Stirling	EvoFIT	41131
		suicide	41132
	Swansea	alcohol	41141
		food items	41142
2*	Anglia Ruskin	literacy	41211
		stroke	41212
	Chichester	colour	41221
		diabetes	41222
	Liverpool Hope	chess	41231
		terrorism	41232

Unit of Assessment 6 Agriculture, Veterinary and Food Science

Score	University	ICS title abbreviation	File number
4*	Aberdeen	climate	61111
		rice	61112
		windfarms	61113
	Warwick	biopesticide	61121
		footrot	61122
2*	Canterbury	ladybird	61211
		welfare	61212
	Hertfordshire	agrimanagement	61221
		mitigation	61222

Main Panel B

Unit of Assessment 11 Computer Science and Informatics

Score	University	ICS title abbreviation	File number
4*	Cambridge	electronic payments	111111
		iris recognition	111112
		realvnc	111113
		security economics	111114
		ubisense	111115
		xen	111116
	Newcastle	circuits	111121
		computational	111122
		dependable	111123
		middleware	111124
2*	Bangor	nationalgrid	111211
		virtualpatients	111212
	Derby	innovativecloud	111221
		sustainablecloud	111222
	East London	Italian	111231
		securesoftware	111232
	Glyndwr	motordesign	111241
		whitegoods	111242

	Liverpool Hope	magic2vip	111251
		watermarking	111252
	West London	hci	111261
		modeldriven	111262
	West of Scotland	enablingtech	111271
		ICTE	111272

Unit of Assessment 13 Electrical and Electronic Engineering and Metallurgy

Score	University	ICS title abbreviation	File number
4*	Imperial College London	computing	131113
		healthcare	131115
		plant	131118
		power	131119
	Oxford	Atomprobe	131121
		Cellmark	131122
2*	Central Lancashire	retardant	131211
		toxicity	131212

Main Panel C

Unit of Assessment 20 Law

Score	University	ICS title abbreviation	File number
4*	Ulster	amnesty	201111
		gender	201112
		northernireland	201113
2*	Bedfordshire	cyberstalking	201211
		humanrights	201212
	Sunderland	localintegrity	201221
		nationalintegrity	201222

Unit of Assessment 22 Social Work and Social Policy

Score	University	ICS title abbreviation	File number
4*	LSE	antibiotics	221111
		fuelpoverty	221114
	Oxford	aids	221121
		immigration	221123
	UCL	crime	221131
	York	budgets	221141
2*	Anglia Ruskin	recovery	221211
		selfhelp	221212
	Liverpool Hope	HEpolicy	221221
		radicalpractice	221222
	Sunderland	exclusion	221231
		samesex	221232

Unit of Assessment 23 Sociology

Score	University	ICS title abbreviation	File number
4*	York	adviser	231111
		biomedical	231112
		sickle	231113
2*	Abertay	fear	231211
		publiclife	231212
	Leicester	racism	231221
	Winchester	alienation	231231
		cheating	231232

Unit of Assessment 25 Education

Score	University	ICS title abbreviation	File number
4*	Durham	performance	251111
		pupilpremium	251112
	Nottingham	leaders	251121
		mathematics	251122
	Sheffield	FE	251131
		Literacy	251132
2*	Anglia Ruskin	nurses	251211
	Bedfordshire	marginalised	251221
	Birmingham City	creativity	251231
	Brookes	attainment	251241
	Chester	multiprofessional	251251
	Derby	guidance	251261
	Newman	pupilexclusion	251271
	NTU	inclusiveED	251281
	Staffordshire	inclusioninED	251291
	Twickenham	rethinkingleadership	2512a1
	West London	socio-emotional	2512b1

Unit of Assessment 26 Sport and Exercise Sciences, Leisure and Tourism

Score	University	ICS title abbreviation	File number
4*	Bristol	outdoor	261111
		travel	261112
2*	Abertay	coach	261211
	Cumbria	psychosocial	261222
	Liverpool Hope	cancerous	261231
	NTU	betaalanine	261241
		heat	261242
	Solent	conditioning	261252
	Sunderland	heritage	261261

	West Scotland	scottishyouth	261271
		steroid	261272
	York St John	motivation	261282

Main Panel D

Unit of Assessment 28 Modern Languages and Linguistics

Score	University	ICS title abbreviation	File number
4*	Cardiff University	devolution	281111
		mabinogion	281112
2*	Salford	arabia	281211
		interpreting	281212

Unit of Assessment 29 English Language and Literature

Score	University	ICS title abbreviation	File number
4*	Bedfordshire	academicenglish	291111
	Kingston	hampton	291121
		military	291123
	Newcastle	poetry	291132
		sevenstories	291134
	Swansea	chester	291142
2*	Leeds Trinity	aspiring	291211
		victorian	291212
	Liverpool Hope	archives	291221
		shakespeare	291222
	Newman	accident	291231
		contemporary	291232
	Twickenham	choices	291241
		huxton	291242

Unit of Assessment 30 History

Score	University	ICS title abbreviation	File number
4*	Hertfordshire	bailey	301111
		foundling	301112
2*	Sunderland	historical	301211
		mining	301212
	Twickenham	Palestine	301221
		sharedworld	301222

Unit of Assessment 35 Music, Drama, Dance and Performing Arts

Score	University	ICS title abbreviation	File number
4*	Goldsmith	afghan	351111
		soundscapes	351114
	QMUL	Hispanic	351122
	RNCM	anya	351131
	Southampton	finnissy	351141
		sirens	351143
2*	Bedfordshire	choreographing	351211
		historiography	351212
	Salford	brass	351221
		jazz	351222
	West London	record	351231
		repertoire	351232

Unit of Assessment 36 Communication, Cultural and Media Studies, Library and Information Management

Score	University	ICS title abbreviation	File number
4*	Cardiff	newscoverage	361111
		vegetative	361112
	Leicester	disabled	361123
		visitors	361125
	LSE	empowering	361132
		citizen	361131
2*	Aberystwyth	digitalassets	361211
		professionalism	361212
	Brookes	cinema	361221
		gaydiscourse	361222
	Glyndwr	antidepressant	361231
		broadcast	361232

Appendix C: List of impact case studies included in Sample C

The following table includes an overview of all ICS that are included in Sample C, which is a smaller selection from Sample A with balance across sub-corpus parts as a principal aim. Main Panels A and B are combined in this sample for both sampling and analysis purposes.

The ICS title abbreviation refers to the abbreviations introduced in Appendix A for Sample A.

Main Panel	Score	Unit of Assessment	University	ICS title abbreviation
A/B	4*	4	Birkbeck	developmental research
			East London	ecstasy
			Stirling	EvoFIT
			Swansea	food items
		6	Aberdeen	climate
			Aberdeen	windfarms
			Warwick	biopesticide
			Warwick	footrot
		13	Imperial	computing
			Imperial	healthcare
			Imperial	plant
			Imperial	power
	1*/2*	4	Anglia Ruskin	literacy
			Anglia Ruskin	stroke
			Chichester	colour
			Chichester	diabetes
			Liverpool Hope	chess
			Liverpool Hope	terrorism
		6	Canterbury	ladybird
			Canterbury	welfare
			Hertfordshire	agrimanagement
			Hertfordshire	mitigation
		13	Central Lancashire	retardant
			Central Lancashire	toxicity
C	4*	20	Ulster	amnesty
			Ulster	gender
			Ulster	northernireland
		22	LSE	antibiotics
			LSE	fuelpoverty
			Oxford	parenting
			UCL	crime
			York	childsupport
			York	workingage
		25	Durham	performance
			Durham	threshold
			Nottingham	leaders

Main Panel	Score	Unit of Assessment	University	ICS title abbreviation
			Nottingham	vocational
			Sheffield	FE
	1*/2*	20	Bedfordshire	cyberstalking
			Bedfordshire	humanrights
			Sunderland	localintegrity
			Sunderland	nationalintegrity
		22	Anglia Ruskin	recovery
			Liverpool Hope	hepolicy
			Liverpool Hope	radicalpractice
			Sunderland	exclusion
		25	Anglia Ruskin	nurses
			Birmingham City	earlyyearsED
			Chester	multiprofessional
			Newman	safeguarding
			Staffordshire	inclusioninED
			West London	wideningparticipation
D	4*	29	Bedfordshire	academicenglish
			Kingston	military
			Newcastle	poetry
			Swansea	chester
		35	Goldsmith	afghan
			QMUL	hispanic
			RNCM	wind
			Southampton	georgian
		36	Cardiff	vegetative
			Leicester	disabled
			Leicester	visitors
			LSE	citizen
	1*/2*	29	Leeds Trinity	victorian
			Liverpool Hope	archives
			Newman	contemporary
			Twickenham	choices
		35	Bedfordshire	choreography
			Salford	jazz
			West London	record
			West London	repertoire
		36	Aberystwyth	professionalism
			Brookes	cinema
			Brookes	gaydiscourse
			Glyndwr	antidepressant

Appendix D: List of n-grams that are significantly more frequent in either high- or low-scoring impact case studies

This table includes a list of all the n-grams that appeared at significantly different frequencies across high- and low-scoring ICS, as described in sections 6.1.2 and 6.1.3.

It includes the following information:

1. Theme category that an n-gram was seen to represent
2. Word form of the n-gram
3. Number of occurrences in the high- and low-scoring sub-corpus respectively
4. Type of comparison in which this n-gram was significant: overall (all Main Panels combined) or in one of the within-panel comparisons between high- and low-scoring ICS
5. Sub-corpus in which this n-gram appears significantly more frequently (high- or low-scoring ICS) – this is also colour-coded, that is, n-grams that are key for high-scoring ICS are shaded blue
6. Brief description of the use of this n-gram in the sub-corpus in which it appears more frequently, based on concordance lines as explained in section 6.1.4
7. Whether this n-gram was seen as free or restricted editorial choice, or dictated by content, as explained in section 6.1.5.

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
attribution - agency	**by professor**	58	9	overall	high	usually with either "research" or verbs describing researching, e.g. "undertaken", "carried out", "conducted"	restricted
	by the	164	24	A	high	Left collocates: “funded”, “research/study/studies”, “used”, “published”. Right collocates: “health”, “university”, “group”, “department”, “team”, “institute”.	free
	led by	35	0	A	high	followed by name of researcher or institute. To the left: “study”, “project(s)”, “team”, “trial”, “programme”, “research”	restricted

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	led by professor	24	2	overall	high	Left collocates: “team”, “project”, “research group”, “study”… very often together with name of university	restricted
	on our	19	0	overall	high	usually with "findings", "work", "research", "data"	restricted
	our research	61	7	A	high	Some (esp. UoA3 and 4) use this quite a few times. Others only once, e.g. as opening for section 1. “Our research (findings) has/reports/suggests”	restricted
	professor of	27	0	A	high	Usually as part of the format Name (possibly institution), professor of (field)	restricted
	research fellow	50	10	overall	high	mostly in a list of researchers at the end (sometimes start) of Section 2	restricted
	s research	79	24	D	high	mostly with a researcher name, otherwise the name of a research unit; to the right: “(has) contributed”, “demonstrated”, “impacted”, “informed”...	free
	school of	44	10	overall	high	mostly the submitting school, often in the context of describing the researchers' positions or the specialism of the school	restricted
	university of	118	15	A	high	15x a partner university, otherwise submitting institution	free
	at the university	11	31	C	low	actions (or employment!) at a partner university abroad (13x), but mostly (19x) the submitting institution	free
	by dr	5	10	A	low	variety of research words: (data were) collected, conducted, headed, led, carried out, undertaken	restricted
	his work	8	21	C	low	often indicating significance/change: has, is, was; also often followed by "on", i.e. it is used to frame content	restricted
	of dr	3	16	overall	low	more words relating to impact (e.g. "impact", "influence"), less of a focus on research than "by dr"	restricted
	research team	24	26	overall	low	action-focused: have/were; often in passive voice ("have been approached")	restricted

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	the university	26	73	C	low	18x a partner university, otherwise the submitting institution	free
	we have	31	38	overall	low	part of present perfect constructions. Occasionally these are part of testimonials	restricted
	work of	46	45	overall	low	25/45 relate to the work of the researchers, 20 to that of research partners/pathways/impact	free
attribution - link	**cited in**	23	9	C	high	policy documents, guidance, white papers, Lords debate	restricted
	contribution of	18	0	overall	high	specifying what the contribution (of the research/impact/funder…) is.	free
	evidence on	11	0	C	high	provide “evidence on” whatever the research problem is, often to policy makers	restricted
	led to	83	10	A	high	Left: has (also/already/directly) led to; Right: research/observation/discovery/collaboration	free
	led to the	58	20	overall	high	creation/development/establishment/recommendation	free
	of evidence	26	5	overall	high	often "evidence-based"; for; on.	free
	result of	30	0	A	high	as a result of our/her/the research/findings/work/studies	free
	resulting in	30	3	overall	high	Usually making research-impact-link; often with comparative adjectives (e.g. fewer, higher, increased)	free
	the basis	26	10	C	high	served as/formed/was used as/will form the basis of… on the basis of	free
	used the	33	7	overall	high	evidence, findings, research	free
	used to	33	12	C	high	research/documents/evidence has been/was “used to” decide/demonstrate/inform/prioritise/test	free
	using the	36	10	overall	high	evidence, research, results	free
	was used	14	0	C	high	research/report/data “was used” as evidence/to inform/ by stakeholders	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	from the	71	112	C	low	Left: findings/was published/research(ers)/staff from the…; Right: from the research / project / University	free
	has informed	24	27	overall	low	research has informed; 6x in relation to something like debate/thinking, 5x connected with policy	free
	of the research	21	47	C	low	impact, outcomes, results, findings of the research	free
	of this research	20	23	overall	low	claims, findings, impact, results. Difference between findings (of the research activity) and results (a change elsewhere that can be traced back to the research, or at least be claimed to go back to the research)	free
	the research has	52	53	overall	low	Right: informed, influenced, been disseminated - showing link to impact	free
	the work of	35	35	overall	low	15 of 35 relate to the work of the researchers, 20 to that of research partners/pathways/impact	free
	this research has	9	24	C	low	Right: informed, influenced, been disseminated - showing link to impact	free
	this work	21	41	C	low	Right: has informed/influenced/improves	free
	through the	22	45	C	low	through the engagement/identification/work of -> often with abstract nouns	free
	work has	25	48	C	low	often with a name or pronoun (his work has…); been disseminated, had an impact, informed -> creating a link between research(er) and impact, similar to "the research has"	free
	contribute to	17	22	overall	low	8/22 are passives: was/has been asked/invited to “contribute to”…; 2x "will contribute to" at end of a section, looking like future impact	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
attribution - non-specific	**as part of the**	21	32	overall	low	part of a funded project, of the research, of a pathway activity, of an impact outcome… quite varied	free
beneficiaries	**and community**	15	26	overall	low	often beneficiaries	content
	group of	22	24	overall	low	e.g. stakeholders, students, participants, service users, professionals	restricted
	members of	6	26	C	low	members of government/parliaments, community groups, NGOs, and occasionally the research team.	content
	practitioners and	11	27	C	low	practitioners and policy makers/practitioners and (beneficiaries, e.g. teachers, athletes)	content
	working in	9	24	overall	low	ca. 1/3 is people (e.g. professionals) working in a certain environment. Other instances are unspecific or incidental	restricted
date	**after the**	31	9	overall	high	Either referring to a point in the narrative of the impact, or to an external event that had an influence or serves as an anchor for the narrative	restricted
	from to	40	13	overall	high	14x date (from [year] to [year], otherwise quantifying change related to the impact (increasing, reducing…). Low: 12x date, no quantitative change reported using these words	free
	in march	33	6	overall	high	always followed by year (2006-2013)	content
	since the	52	15	overall	high	21x with year ("Since 2009, the…"); otherwise specifying a point in the narrative, e.g. "since the introduction of…", sometimes with a year. Low: only 4x with year (1994-2009); of the 11 remaining ones, only 3 refer to events in the 2000s (others: 1990s, earlier, or causal particle; usually not referring to a point in the narrative of the impact)	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
discourse	**e g**	49	11	D	high	usually in brackets to give examples of stakeholders/beneficiaries or pathways. The significance is that high-scoring ICS give examples, and that they do this with minimum word/space investment and maximum signposting.	free
	in these	31	9	overall	high	in these patients, in these studies etc. - pointing backwards to previously introduced entities	free
	that the	71	22	D	high	Left: comment/ensure/recommend/state/show	free
	and also	22	32	overall	low	no clear pattern of use visible	free
	as well as	26	52	C	low	listing either the researchers' actions towards impact or beneficiaries	free
	case study	15	49	C	low	Right: is/was, describes, examines; includes "this case study"; in 3 cases: refers to a case study approach to research	free
	has a	20	22	overall	low	5x related to impact, otherwise description of some content, occasionally reputation of researcher	free
	in this	77	79	overall	low	16x area, 8x case study, 7x field; research, study, work	free
	in which	79	81	overall	low	40x way(s), 8x manner	free
	interest in	15	24	overall	low	no specific pattern, ranging from research interests to testimonials and describing stakeholders/beneficiaries	free
	is not	18	22	overall	low	2x "not only" - strong; 2x "not possible" - concession, and other instances are more like this (not the case, not clear…) -> occasionally this is hedging the case study content, but partly looks like it's part of the narrative. Using "not" might make things look more negative, apart from "not only".	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	it has	72	63	overall	low	it has had, been, influenced. The 63 instances would need closer examination to determine whether "it" refers to research, impact or a pathway even	free
	of this	38	79	C	low	consequence, example, finding, impact, result of this research/work	free
	that this	20	24	overall	low	to the left: argue/argument, claim; assume/indicate/reveal/show/suggest	free
	the case	27	34	overall	low	18x the case study; otherwise: is the case, make the case for (2x each) and others	free
	the following	4	18	D	low	impact, research, question(s) - metadiscourse organising the text	free
	this case study	2	12	D	low	introducing the topic of the case study or the research (the impact described in this case study is underpinned by…)	free
	to their	25	29	overall	low	no clear pattern of use visible.	free
	which the	8	19	C	low	manner/way/area in which	free
	who had	18	21	overall	low	individuals, students, people - would need a closer look to determine whether these are participants in *research* studies or pathway activities. Relatively more frequent occurrence in section 2 suggests probably more research participants	free
framing	**aim of**	16	19	overall	low	Making the aim of the research or a pathway activity ("day", "book") explicit	free
	area of	6	22	C	low	area of research, or area of (whichever area it is - usually research area but sometimes related to impact topic); area of concern	free
	focus on	36	40	overall	low	no clear pattern of use visible.	free
	focused on	26	30	overall	low	Research questions, some engagement activities	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	focuses on	2	19	C	low	5x with "research", 3x with "this case study"	free
	in relation to	10	22	C	low	specifying - e.g. "particularly in relation to"	free
	in terms of	13	27	C	low	specifying the area of research or impact, e.g. "in terms of coaching methodologies", "national policy"	free
	in which	33	33	D	low	way(s) in which (21x) - demonstrate, identify, investigate - research questions and type of change	free
	nature of	34	34	overall	low	8x with "impact", usually with various content words. No clear pattern beyond this	free
	the area of	3	16	overall	low	expertise, research, work, study in the area of	free
	the concept of	0	10	overall	low	appears across UoAs	free
	the context of	9	22	overall	low	describing the context of the pathway activity or the further research context	free
	the issue of	0	11	overall	low	address, examine, tackle	free
	the way	9	19	D	low	to the right: often with verbs of perception, e.g. perceive, think, interpret; also action verbs (create, select, contribute to). To the left: change, influence	free
	the ways in which	16	23	overall	low	demonstrate, identify, investigate - research questions and type of change	free
	to assess	16	19	overall	low	8x with "impact" or similar words (Change, implementation) - part of the narrative in these cases; otherwise research aim. Coded as "framing" as conveying the intention to assess something	free
	ways in which	19	30	overall	low	in addition to "the ways in which" (separate): research, discover, explore, highlight, suggest, propose - mostly impact-, rather than research-related	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
impact descriptions - awareness	**knowledge and**	15	22	overall	low	knowledge and experience/skill/understanding; occasionally of the researcher/team, but more often enhancing or improving knowledge of beneficiaries	content
	understanding and	10	24	overall	low	increased knowledge and understanding, public understanding - this is predominantly about enhancing the understanding of beneficiaries (or the public as beneficiary)	content
	understanding of the	23	29	overall	low	add/contribute to/inform/enhance/enrich understanding of the; occasionally this is a research aim but mostly an impact (e.g. this has led to greater understanding of…)	content
impact descriptions - capacity	**and professional**	14	25	overall	low	professional practice, development, bodies/associations, practitioners. This is mostly impact on some sort of professional development but partly beneficiaries	content
	and skills	14	35	overall	low	14x official body or government department with "skills" in the title. The Learning and Skills sector too. Employment and skills training. Even though the word itself appears as title of a centre/department/… occasionally, the implication is still that this is to build capacity, especially in conjunction with the other occurrences	content
	professional practice	8	33	overall	low	one ICS ("Professionalism") has 16x -> content word. Counting only one of these instances, the result is still significant. In the other files, left collocates include "enhance", "influence", "inform"	content
impact descriptions - change	**improve the**	29	6	overall	high	improve the lives, quality, response, teaching.	content
	reduction in	34	8	overall	high	Nearly half with % (a 50% reduction in…); all but 6 in MP A; Low: never with %, mostly UoA26 (MP C)	content
	to improve	29	16	C	high	wide range of words following this - diverse variety of improvements	content

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	to establish	13	21	overall	low	4x research aim, otherwise establish e.g. a project/centre/… i.e. impact	content
impact descriptions - public	**and public**	45	12	overall	high	Across UoAs but notably it appears in a large number of MP D ICS, whereas it appears in fewer MP A ones (but multiple times in those). Right collocates: Awareness, understanding, debate, engagement, policy. And "health" - 4 of these are "professor of physical activity and public health" in list of researchers. Discounting these, the difference is still significant.	content
	of policy	26	12	C	high	Right: recommendation, revision, debate	content
	of public	22	11	C	high	Right: public spending, figures, services…	content
	public awareness	26	5	overall	high	Left: enhanced, (variations of) increase and raise; usually followed by "of". Mainly MP A	content
	public health	59	5	overall	high	9 are in UoA2 which does not have low-scoring ICS represented in the sample, and a further 25 are in *one* ICS in UoA3 - still meets significance threshold without the 25	content
	the policy	23	7	C	high	policy change/decision/debate/document…	content
impact descriptions - unspecific	**impacts on**	38	10	overall	high	8x as heading, further 5 at start of section/paragraph; left collocates: direct, immediate, significant	free
	quality of	26	12	C	high	6x “of life” -> content word. Otherwise “of the work/research” or whatever the research subject or intended impact was	content
	an impact on	5	17	overall	low	Used to frame impact at the beginning (of the ICS or the section); nearly always "has/have had" or "have"; to the right: ca. half are vague, e.g. "public discourse" or "the way people think", others are specific e.g. "survival in this	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
						group" (UoA26) and "over 100,000 high-risk patients" (also UoA26)	
	impact of	51	81	C	low	ca. half: impact of the research/work. Others: impact of an intervention, or impact of a problem/state of affairs that caused the need for an intervention (e.g. light pollution)	free
	impact on the	27	29	overall	low	Difference to "an impact on" is usually an adjective: huge, major, practical, measurable (a measurable impact etc.)	free
	the development of	26	28	D	low	7x reference to theoretical/research-related developments, otherwise impact descriptions: new software tools, an in-service training program	restricted
	the impact	160	136	overall	low	usually used numerous times per ICS (not all use them - 66/93); 6x start of section 1, 9x start of section 4; examine/explore/evaluate/maximise the impact; 7x evidence/example of the impact	free
partner	**the national**	41	9	D	high	institutions like National Assembly (Welsh), National Theatre/library/trust	content
	worked with	28	7	overall	high	mostly research partners (Basque government, companies, NGOs…), occasionally target groups/beneficiaries (e.g. refugee writers).	restricted
	centre for	34	33	overall	low	Occasionally the university/research centre carrying out the research/submitting the ICS, but more often other centres for something which were either research collaborators or partners for achieving the impact	restricted
	university of	178	164	overall	low	99/178 partner university rather than submitting university	free
	with the	77	29	A	low	collaboration/involvement/… with the (partner institute of sorts, e.g. Donkey Sanctuary)	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
pathway	**a series of**	27	14	C	high	varied! Right: series of articles/studies, but also problems, recommendations, interventions.	restricted
	and other	29	13	C	high	mostly part of a series of pathway activities/interventions/beneficiaries	free
	by the	217	141	C	high	Left collocates: funded, commissioned, acknowledged, influenced. Right collocates: police, trust, council, commission, foundation, minister. Secondary meaning of "agency" (of the researchers): left collocates: produced, organised; right collocates: department, institute, team	free
	have been	70	21	D	high	wide range of verbs; often enabling (have been inspired/possible/stimulated) or pathway-related (have been reported/visited/discussed)	free
	institute for	31	8	overall	high	Institute for excellence, for Health (NICE), other official institutes (non-university).	restricted
	on how	21	5	C	high	evidence/guidance/report on how (best) to…	free
	report on	19	0	C	high	usually a report by an official body, e.g. UNESCO, WHO; occasionally a report by the researchers, usually then requested by an external source	content
	review of	43	14	overall	high	independent, official, systematic; either an external review citing a research/impact-related source (i.e. a document), or a review of policy or the like (i.e. a call for change)	content
	set up	26	5	overall	high	mostly passive without agent, with a pathway to impact (e.g. website or a collaboration) as subject; if there is an agent: often external sources like the Prime Minister, UK Government	restricted
	as part of	21	21	D	low	as part of a pathway or impact activity	restricted

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	co uk	6	20	overall	low	part of URLs. 9x news outlet sites (e.g. THE, BBC, Guardian), 5x from one submission the researcher's own (no longer maintained) website; others include Eventbrite and a review on Amazon	restricted
	conference in	6	19	overall	low	12x with date (month or year), 6x with place (city or country); reporting on (invited/keynote/other) contribution of the researcher to conferences (academic or professional) and events hosted at the research institution/by the researchers	content
	debate on	16	20	overall	low	public (6x)/international (2x) debate; contribution to/influence on/informed. Occasionally research topic or literal meaning, e.g. "chaired the debate on"	content
	disseminated through	2	13	overall	low	Left: research, work, outcome has been/is regularly disseminated	free
	dissemination of	7	27	overall	low	6x findings/insights, 7x research (+ one ICS had 4 headings starting with "dissemination of")	free
	has been	115	45	A	low	To the left: research/work and all sorts of content words: training, tool, memorandum; researcher names. To the right: has been invited/noticed/picked up/implemented. More varied than in High, but often less pointed ("has been well received"), more impression of agentless passive	free
	has been disseminated	0	11	overall	low	5x research, 3x work has been disseminated; 7x followed by means of dissemination (event, publication), 4x by reach (disseminated to larger groups, disseminated widely across non-academic fields)	free
	has led to	10	22	C	low	has led to a number of journal articles / a book / further research collaborations; invitations to speak, meetings, greater understanding.	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	http www	11	31	C	low	parts of URLs; news outlets, university websites, government websites, YouTube	restricted
	in the project	3	15	overall	low	3x research project, otherwise impact; for one of the research projects, it is unclear whether the claimed impact was part of the research project.	restricted
	involved in	42	44	overall	low	can be the researcher or participants and/or beneficiaries; often "project" or other pathway activities; occasionally research	free
	of the project	13	31	overall	low	mostly pathway projects, occasionally research project (in low)	restricted
	part of the	19	27	D	low	1/2 are "as part of the"; usually part of a pathway project or event, e.g. festival (though those events may be the impact in some UoAs)	restricted
	project has	10	24	overall	low	pathway to impact projects	restricted
	the book	22	44	overall	low	book that has been cited, received, used etc.; often a monograph (i.e. research output) but sometimes a textbook or a book aimed at professionals. -> pointing in direction of one-way dissemination, a theme emerging from the "low" n-grams	content
	the conference	0	8	D	low	pathway activities, e.g. a "Sixth Form Conference" (annual lecture as main impact)	content
	the debate	13	22	overall	low	4x "inform", 6x "contribute". Cluster in UoA20 Law. It is interesting that the difference between High and Low is much bigger in "the debate" than in "debate on". Low has more "the debate" and High has more "debate on". Are High more routinely being specific about the topic of the debate, or more judicious in their use of the word?	content

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	the event	4	17	D	low	In UoA35 (visual or performing arts), often the event was the impact	restricted
	the project	47	49	D	low	pathway project	restricted
	the subject	15	20	overall	low	4x was the subject of newspaper coverage; 6x content word ("subject-based curriculum", across 2 ICS, and "attainment in the subject" in a third ICS); otherwise topic framing (e.g. "workshop on the subject"). In "High", this is never used as a content word	free
	the training	3	23	C	low	usually either training events or a claimed impact on the training of (teachers, coaches…)	content
	through a	24	28	overall	low	various pathway activities, e.g. through a website, survey, focus group	free
	through the	24	29	D	low	was circulated/disseminated/promoted through the	free
	to contribute	11	23	overall	low	Difference between high and low bigger than for "contribute to" - because of "asked to contribute"? 11/23 occurrences in "low" have this pattern; others describe research aims or pointing to type of impact, e.g. government policy or contributing to public debate.	free
	training and	8	24	C	low	usually either training events or a claimed impact on the training of (teachers, coaches…)	content
reach - national	**in the uk**	43	11	D	high	sometimes content-related (noise-pollution/slavery in the UK), but also reach-related (engage communities in the UK). To the right: 6x unspecified wider reach (and beyond/abroad), 8x specified wider reach (Spain/US)	content
	in the uk and	49	14	overall	high	5x abroad, 4x across/around the globe/world, 3x beyond, 10x internationally, 13x with more specific place (Europe, US, South Africa…)	content

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	in the us	30	0	overall	high	great variety of impact in the US	content
	of the uk	27	5	overall	high	1/3 in various ICS in UoA36; 1/3 in UoA22.	content
	the uk	123	74	C	high	not checked due to number of occurrences; appears in 31 of the 37 files	content
	the uk and	24	11	C	high	international(ly), worldwide, overseas; 3x with a specific other country, 3x in non-geographic contexts	content
	the us	26	2	C	high	usually with US government/official body	content
	the world	37	10	D	high	4x the world's largest/leading…; 17x around/across/all over/throughout the world; world premieres. Sometimes: "the world of" (e.g. sex trafficking/private music making)	content
	uk and	27	4	D	high	9x unspecific wider reach; 10x specific wider reach	content
	a national	10	22	C	low	a national level; a national event; other ways of indicating national activity or recognition	content
	an international	19	26	overall	low	5x conference; 4x international experts; other ways of expressing international relevance	content
	national and	23	43	C	low	European (4x); international (25x); local (10x)	content
	nationally and internationally	10	21	overall	low	often at end of a sentence or even paragraph; 7x "both"	content
reach - subnational	**in england**	46	20	C	high	policy, practice, benefits system…	content
	and local	31	33	overall	low	5x authorities, 7x government; 10x national and local	content
	city council	5	18	overall	low	11x the city council where the submitting university is located; 6x Rome, where a key conference in one ICS took place (result is still significant without these)	content
	in local	4	18	overall	low	5x authorities, 4x government, 3x local and national	content
	of england	9	22	overall	low	4x England-wide bodies (e.g. bishop's conference); 1x south-east (from Oxford Brookes), the others are north (or north west or north east). All but one area-specific	content

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
						occurrences are from a university in the area (Sunderland once writes about the University of the West of England)	
	of local	14	26	overall	low	only 4x "authority/ies" -> wider range of types of local involvement, with no clear pattern	content
	the local	18	33	overall	low	4x press (or the like); no other clear pattern visible	content
	the north	6	33	overall	low	Nearly always north east, north of england, north west - from unis located there; north essex from anglia ruskin	content
reach - unspecific	**across the**	32	8	D	high	either conceptual (across the sector/organisation) or geographic (across the UK, EU, world, nation)	restricted
	a number of	18	47	C	low	in relation to dissemination of research and/or recommendations; with people/beneficiaries; some more conceptual co-occurrences, e.g. barriers. To the left: highlighted, identified; resulted in, led to; collaboration with	free
	a range of	69	60	overall	low	left: across/by/from/with a range of; to the right: beneficiaries, activities, outputs (e.g. media platforms), more conceptual terms like "approaches"	free
	within the	26	54	C	low	within the community/UK/group; no other clear patterns of use visible	free
research	**programme of**	23	8	C	high	mostly of research	restricted
	research on	54	34	C	high	Left: adjectives, e.g. comparative, current, funded, influential, innovative; also possessive ([name]'s research). Right: research topic	free
	research programme	30	4	overall	high	Describing overall research leading to impact	restricted
	the population	21	0	overall	high	mainly about populations of people or things being studied, in MP A Section 2 mostly	content

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	the study	29	14	C	high	usually related to research methods or aims, or explanation of the value of the research	free
	body of	14	31	C	low	work, research.	restricted
	in the field	20	26	overall	low	figuratively to mean "in the discipline", apart from one literal example ("coyotes in the field" in UoA6). 14x specified "in the field of", most other instances it is understood that it is the field of the researcher. 7x describing beneficiaries ("professionals working in the field") but usually the research/researcher, occasionally with esteem marker ("distinguished in the field")	free
	of research	75	80	overall	low	Left: 5x area, 14x body, 6x dissemination, 5x programme, 4x strand(s)	free
	research is	0	13	D	low	this/the underpinning/(name)'s research is based on / reported in / central to..	free
	research project	8	19	C	low	2x collaborative, 4x funded; descriptions of research aims, research context (e.g. part of a larger project); occasionally linking the project to effects	restricted
	research projects	14	23	overall	low	18x indicating that the impact in the ICS is based on more than one project (cf. qualitative analysis: this sometimes resulted in a disjointed narrative)	restricted
	the research	84	53	A	low	Right: findings, team, has/was/will	free
	the work	98	86	overall	low	ca. 1/3 is work after the research stage, the others refer to research	free
	this research	41	71	C	low	24x with present perfect (indicating change); otherwise no clear pattern	free
	wellcome trust	19	0	overall	high	often in list of funders	content

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
research - output	**et al**	18	64	C	low	always part of a citation, occasionally in the sentence rather than in brackets. 49/64 occurrences in UoAs 25/26 though	restricted
	journal of	7	18	overall	low	always research journals. They should not feature so prominently in the main text but be listed in Section 3.	restricted
	peer reviewed	12	23	overall	low	18x publications/books/articles/journals, otherwise research, conference, award, evaluation	free
	the paper	6	17	overall	low	7x what happened to the paper (was published, has been utilised), 4x who wrote it, 7x what the paper does	restricted
research - result	**showed that**	29	4	C	high	research/findings/study	free
	that the	62	25	A	low	demonstrate/indicat*/state*/suggest* that the (more tentative than use in High)	free
research - subject	**explored the**	11	21	overall	low	14x research/study/work, otherwise lecture or other output; followed by topic/subject of research	free
	related to	9	21	C	low	specifying the area of research, occasionally relating the research to impact	free
	relationship between	15	21	overall	low	mostly framing research questions	restricted
	research into	12	33	C	low	followed by research area/topic	free
	the impact of	36	63	C	low	Partly making links between research (11x)/work(4x)/ project(3x), but mostly specifying a research area, studying the impact of something on something	free
significance	**the key**	20	8	C	high	actors/drivers, messages, recommendations	free
	up to	30	0	A	high	3x up to date, 7x phrasal verbs with "up" e.g. set up (in order) to -> statistical significance threshold met if phrasal verbs are discounted, but test classed as "not reliable" - term still kept, as the phrasal-verb uses do not appear in	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
						Low and are therefore peculiar to High, usually pathway-related. Otherwise usually followed by a number and often by time (a number of years, or a particular year)	
	were the	29	4	overall	high	were the first/focus/foundation	free
	one of the	20	39	C	low	one of the few/first/most important/main	free
	practice and	37	51	overall	low	No strong collocations - 8x with "policy" (left or right), 6 with "professional".	restricted
	the first	35	33	D	low	21x emphasising novelty (the first conference on…); otherwise structuring the text (The first impact/the second impact)	restricted
	the value	18	23	overall	low	showing the value of the impact topic to beneficiaries	free
significance - change	**a new**	38	6	D	high	to the left: develop/establish/launch; to the right: approach, direction, research - or description of whatever the impact is (a new website, youth theatre…)	restricted
	are now	40	6	overall	high	are now embedded, required, used; often with present participle to form a present continuous verb form, indicating ongoing action.	restricted
Significance - official	**department for**	32	8	C	high	government departments - mainly Education and Work and Pensions	content
	department of	44	14	C	high	mainly Department of Health / Social Development (UK and other governments)	content
	department of health	40	6	overall	high	Some possible sampling bias as UoA 1 Medicine is not included in Low corpus; however, some occurrences are in UoA 26 Sport and Exercise science, where there are 2 high- and 16 low-scoring ICS	content
	for education	21	9	C	high	Department, Minister, Secretary of State	content

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	government and	29	6	overall	high	right: usually another organisation e.g. NGO, a partner university, business/industry.	content
	government policy	41	11	overall	high	across UoAs; appears in headings/lists; often with verbs of change, e.g. influence, affect, change; "impact on"; occasionally with country (Dutch, South Africa, Welsh, UK)	content
	government s	32	2	C	high	strategy, paper…	content
	house of	22	10	C	high	commons, lords	content
	including the	63	23	overall	high	often with official bodies, e.g. government, WHO	free
	of health	34	12	C	high	usually Department of Health or similar institution in other countries	content
	policy makers	33	3	D	high	influenced/informed//impacted policy makers (6x); 9x link between beneficiaries and policy makers	content
	prime minister	32	0	overall	high	10x content - still significant without these	content
	produced by	12	0	C	high	documents, research, reports produced by government or university departments or other official bodies	restricted
	secretary of	19	7	C	high	state	content
	the department	57	13	C	high	mostly government departments (Health, Education, Work and Pensions, Transport…), but sometimes the researching department. 6x listed as funder, esp. in Section 2 - "commissioned by"	content
	the department for	28	3	C	high	government departments - mainly Education and Work and Pensions	content
	the department of	23	10	C	high	mainly Department of Health / Social Development (UK and other governments)	content

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	the european	31	0	A	high	commission, food safety authority, parliament, registry of quality outcomes, union.. -> political and regulatory institutions	content
	the government	52	13	C	high	announced changes or made changes, or funding	content
	the government s	20	0	C	high	strategy, paper…	content
	the uk s	29	0	overall	high	This is sometimes with an official body or guidelines ("joint committee", "standards of care"), sometimes with a superlative (first, growing, largest)	content
	policy and practice	29	31	overall	low	change, contribute to, influence	content
significance - scale	**a major**	26	14	C	high	often with words of change: impact, change; also with scale of contribution, e.g. a major part/element	restricted
	long term	28	5	C	high	various words describing the impact: effect, benefits, performance, investment	restricted
	millions of	18	0	overall	high	usually people or money	content
	more than	32	0	A	high	usually followed by a number, then followed by (potential) beneficiaries. Either scale of the impact or scale of the problem. Occasionally scale of pathways (more than 30 news articles)	restricted
	number of	62	6	A	high	25x a number of, 24x the number of. Difference: "the" is more specific, e.g. reduc* /limig*/ increas* the number of. Interesting that "a number of" is Key for overall Low. This points to high-scoring ICS using more metrics-based comparisons and commenting on numbers, rather than using the generic "a number of" as a quantifier.	free

Theme category	n-gram form	High (no.)	Low (no.)	Main Panel	Key for…	Use in key sub-corpus	Editorial choice?
	the most	35	19	C	high	most significant impact, most deprived area… indicating either the value of the intervention or the need of the beneficiary to be selected	restricted
split: research and pathway	**project was**	11	23	overall	low	3x "research project", but 8 others also relate to research projects -> half of these are research, the other half are pathway.	restricted
	the project	25	62	C	low	split between research and pathway projects, occasionally difficult to decide which one (blurred lines between research and pathway or even impact)	restricted
	this project	24	29	overall	low	split between research and pathway projects, occasionally difficult to decide which one (blurred lines between research and pathway or even impact)	restricted
	work in	43	45	overall	low	no clear pattern of use; can be "work in the field" or "youth work in…", i.e. research or pathway	free
split: significance (for the first time), impact description/content.	**for the**	123	18	A	high	16x “for the first time”; “for the management/treatment of [a condition]”; “for the development of [a sort of impact]”. “Providing evidence for the”; “guide(lines) for the”; “strategies/systems for the”	free

Appendix E: List of n-grams that carry persuasive meaning

This appendix includes the list of n-grams that were assessed as having persuasive meaning, as described in sections 6.1.6 and 6.2.3. The table includes the type of persuasive meaning (credibility, richness, specificity and added value); the word form that carries that meaning; the number of times that that word combination appears in the high- and low-scoring sub-corpora; the type of comparison in which there was a significant difference (all Main Panels combined, or within-panel comparison); and whether the n-gram was key in the relevant high- or low-scoring sub-corpus.

Type of Persuasion	n-gram form	High (number)	Low (number)	Main Panel	Key for…
credibility	**by professor**	58	9	overall	high
	led by professor	24	2	overall	high
	on our	19	0	overall	high
	research fellow	50	10	overall	high
	school of	44	10	overall	high
	contribution of	18	0	overall	high
	led to the	58	20	overall	high
	of evidence	26	5	overall	high
	resulting in	30	3	overall	high
	worked with	28	7	overall	high
	institute for	31	8	overall	high
	including the	63	23	overall	high
	led by	35	0	A	high
	our research	61	7	A	high
	professor of	27	0	A	high
	university of	118	15	A	high
	led to	83	10	A	high
	result of	30	0	A	high
	cited in	23	9	C	high
	evidence on	11	0	C	high
	the basis	26	10	C	high
	s research	79	24	D	high
	of dr	3	16	overall	low
	we have	31	38	overall	low
	has informed	24	27	overall	low
	centre for	34	33	overall	low
	university of	178	164	overall	low
	contribute to	17	22	overall	low
	involved in	42	44	overall	low
	to contribute	11	23	overall	low
	journal of	7	18	overall	low
	peer reviewed	12	23	overall	low

Type of Persuasion	n-gram form	High (number)	Low (number)	Main Panel	Key for…
	by dr	5	10	A	low
	at the university	11	31	C	low
	the university	26	73	C	low
	has led to	10	22	C	low
	et al	18	64	C	low
richness	**number of**	62	6	A	high
	a series of	27	14	C	high
	e g	49	11	D	high
	and also	22	32	overall	low
	a range of	69	60	overall	low
	as well as	26	52	C	low
	a number of	18	47	C	low
specificity	**after the**	31	9	overall	high
	from to	40	13	overall	high
	since the	52	15	overall	high
	up to	30	0	A	high
	long term	28	5	C	high
	focus on	36	40	overall	low
	focused on	26	30	overall	low
	co uk	6	20	overall	low
	focuses on	2	19	C	low
	in relation to	10	22	C	low
	in terms of	13	27	C	low
	http www	11	31	C	low
added value	**are now**	40	6	overall	high
	more than	32	0	A	high
	the key	20	8	C	high
	a major	26	14	C	high
	the most	35	19	C	high
	a new	38	6	D	high
	the value	18	23	overall	low
	one of the	20	39	C	low
	the first	35	33	D	low

Appendix F: Coding manual – Appraisal in impact case studies

This coding manual follows the structure suggested in the appendix to Fuoli and Hommerberg, 2015.

1. Introduction

The texts in this study are the first section (target length: 100 words) of four-page impact case studies submitted to the UK Research Excellence Framework (REF) in 2014. The broad purpose of these texts is to showcase stories of real-world impact based on the research done at the submitting university department, with assessors assigning a score from 1* ("recognised but modest") to 4* ("outstanding"). This score feeds into an overall score for each Unit of Assessment, which in turn is linked to the allocation of significant amounts of research funding. Therefore, the texts are extremely high-stakes.

Specifically, Section 1 "Summary of the impact" is expected to give a succinct overview of the claimed impact and acts like an executive summary, setting the bar and general expectation of the ICS at hand. After reading the summary, assessors may already have formed an initial opinion of the score. The writers therefore have incentive to present their impact in the best possible light (see https://juliebayley.blog/2021/09/02/shiny-vs-authentic-impact/ who calls ICS "competition entries"), while also having to adhere to a small word count. They are challenged to evaluate their story as positive with as few words as possible, trying to align the reader with their evaluation without creating a sales pitch that may seem unconvincing (McKenna 2021: 54).

The aim of this study is to investigate how writers have approached this, and to compare texts that received different ratings (1*/2* versus 4*). The Appraisal system provides a clear framework for identifying and tagging words or phrases that are used in the text as resources of evaluation. The term "resource" is therefore used below to refer to the units of one or more words that can be construed to carry evaluative meaning in a way that is represented in the GRADUATION framework. In particular, given the need to be brief and to avoid sales pitches, it will be interesting to see how evaluation is expressed in this context and where the balance is between overt/inscribed and covert/invoked evaluation. The assessment context may prompt writers to be more explicit in their evaluation, whereas the academic context and word limit may pull in the other direction.

2. General principles

Coding is conducted using the UAM Corpus Tool. This allows for different layers of tagging information in the same text. Segment-level layers allow the coder to apply tags to segments within a text, whereas document-level layers serve to categorise the whole document. The latter can help with post-tagging quantitative analysis by identifying texts as being part of certain sub-corpora. The present analysis uses four separate layers which are coded in separate steps, such that each text appears as uncoded when a new layer is applied to the text:

1) **Graduation_ICS:** a segment-level layer to classify types of evaluative language
2) **Target**: a segment-level layer to tag the target of the evaluation, that is, what is being evaluated
3) **Score**: a document-level layer to facilitate analysis according to whether a text is part of the 1*/2* or the 4* corpus section
4) **Panel**: a document-level layer to facilitate analysis according to which REF Main Panel the text was submitted to, corresponding to discipline

2.1 Graduation

GRADUATION is situated in the Appraisal framework as described in section 3.3 of this thesis. General principles for how GRADUATION tags are applied in this study are listed in this section, and further discussion of the process of analysis is provided in section 7.1 of the thesis. The coding scheme is introduced in more detail in section 3.1 of this coding manual below.

Do not code words that are clearly part of the impact description or of collocations. For example, in “early years”, “early” is not the writer’s choice. Rather, code only those words where the writer may have tried to exert influence.

Similarly, do not code technical terms because the writer has little choice here. This presents a problem because coders may not be embedded in the nine broad disciplines in the corpus and therefore may not know whether something is a technical term. Where they suspect that a technical term may be at work, coders should do a quick internet search and if in doubt, code the term and add a note. The main coder is familiar with the units of assessment in the corpus through deep engagement with the full texts (the texts in the corpus tagged for GRADUATION are only the first section) and through consultancy work in all

of them and therefore feels confident in identifying the vast majority of potential technical terms correctly as such.

Sometimes the problem statement seems apparently neutral but "feels" like the write was trying to exert influence through word choice or placement. In that case, do code. A question to ask is: What else could the writer have written? Would there have been a more neutral choice?

Unitisation: Tag the smallest unit of evaluative meaning, even if there are several words contributing to an evaluative expression. Does each word form part of the same meaning as the other words, or do they represent different resources of GRADUATION in the text? For example, in "very wide recognition", "recognition" is an evaluative term but not part of GRADUATION, but "wide" is GRADUATION:FORCE:QUANTIFICATION:EXTENT:DISTRIBUTION:SPACE, and "very" is GRADUATION:FORCE:INTENSIFICATION:INTENSIFIER. Therefore, given that they perform different function because GRADUATION is so fine-grained, they should be tagged separately.

If two (or more) resources are separated by a comma or coordinating conjunction, they should also be tagged separately as two (or more) instances, even if they receive the same tag.

Unitisation and Classification (see section 3 below) is to be done in the same coding round within each text because they depend on each other, and deciding on classification can help to delineate the item (that is, to unitise). However, they are dealt with separately in the manual because different principles apply.

Discontinuous evaluative expressions: Sometimes evaluative expressions stretch across words that are not part of the same expression, and where the components of the evaluative expression are not by themselves evaluative and therefore cannot be treated as separate units. This can be seen as problematic because the additional words can confound the length of evaluative expressions if this is later quantified. Not all tools provide an option to code these parts as belonging to the same evaluative expression without also including the intervening words. Fuoli points to Carretero and Taboada (2014) who coded the whole expression, and I will follow their decision because of the restrictions in the UAM Corpus Tool. If a tag spans several intervening words that are not part of the expression, a note should be added to specify which words should be considered as tagged.

Multiple function: A word or expression could be read as having, or being part of, more than one evaluative function. It is therefore necessary to decide whether more than one label can be applied in such cases, to allow for both functions to be recorded. In my study, only one label should be applied to each (part-)expression, because multiple tags can lead to double counting and skew the quantitative analysis if there are more labels than coded instances. Especially in a study focusing on GRADUATION, there is less scope for double function than in a study of Attitude, where more fine-grained levels are less easily distinguished especially regarding the Attitude-resources of Judgement (people-focused) and Appreciation (thing-focused) – the options for GRADUATION, especially in the Quantification branch, are clearer. The only situation where a word can be included in more than one tag is in the case of discontinuous evaluative expressions, where one expression may be nested between two components of a different expression. Because the discontinuous expression is specified in a comment to the tag, it should be unambiguous which tag applies to any given word.

Where there may be a discrepancy between the likely intention of the writer and the likely interpretation of the reader, the latter takes precedence.

2.2 Target

To pinpoint what it is that is being put in an evaluative light, a second layer segments all parts of the text such that each word is part of a segment that indicates the type of content. Stretches of text that seem to refer to the same entity being evaluated should be tagged according to the scheme in Figure 22. Such a stretch is likely a clause or entire sentence, but could also be a smaller unit. Section 1 typically describes research and impact, and occasionally activity that led from the research to the impact or otherwise links the impact back to the research. Material describing such activity should be tagged as pathway. Some Sections 1 also include a problem statement which should be tagged as problem. Lengthy background descriptions that cannot be classed as problem statement should be tagged as other. Stretches that include words like *influence* or *implementation* are likely pathway:researcher-led. The label impact is applied to outcomes of such activity.

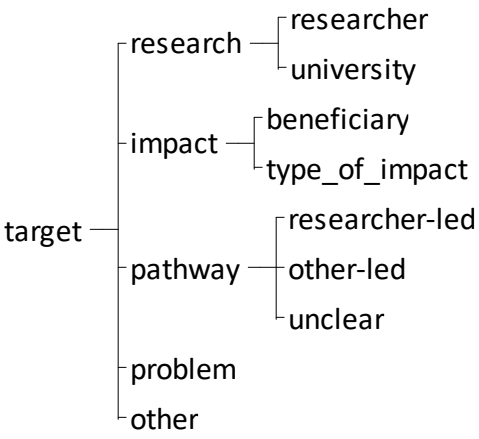


Figure 22: Coding scheme for the "Target" of evaluation in the text

2.3 Score and Panel

The two document-wide layers can be applied according to the file name, which states the Score (1-2 = low or 4 = high) and the Unit of Assessment, from which the Main Panel can be inferred. The coding schemes for these two layers are represented in Figure 23 and Figure 24 respectively.

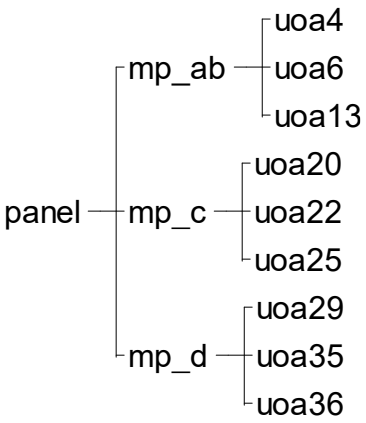


Figure 23: Coding scheme for the Unit of Assessment of a text

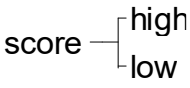


Figure 24: Coding scheme for the scoring bracket of a text

3. Classification

3.1 Overview of the coding scheme

Each instance of GRADUATION should be labelled with the annotation scheme in Figure 25.

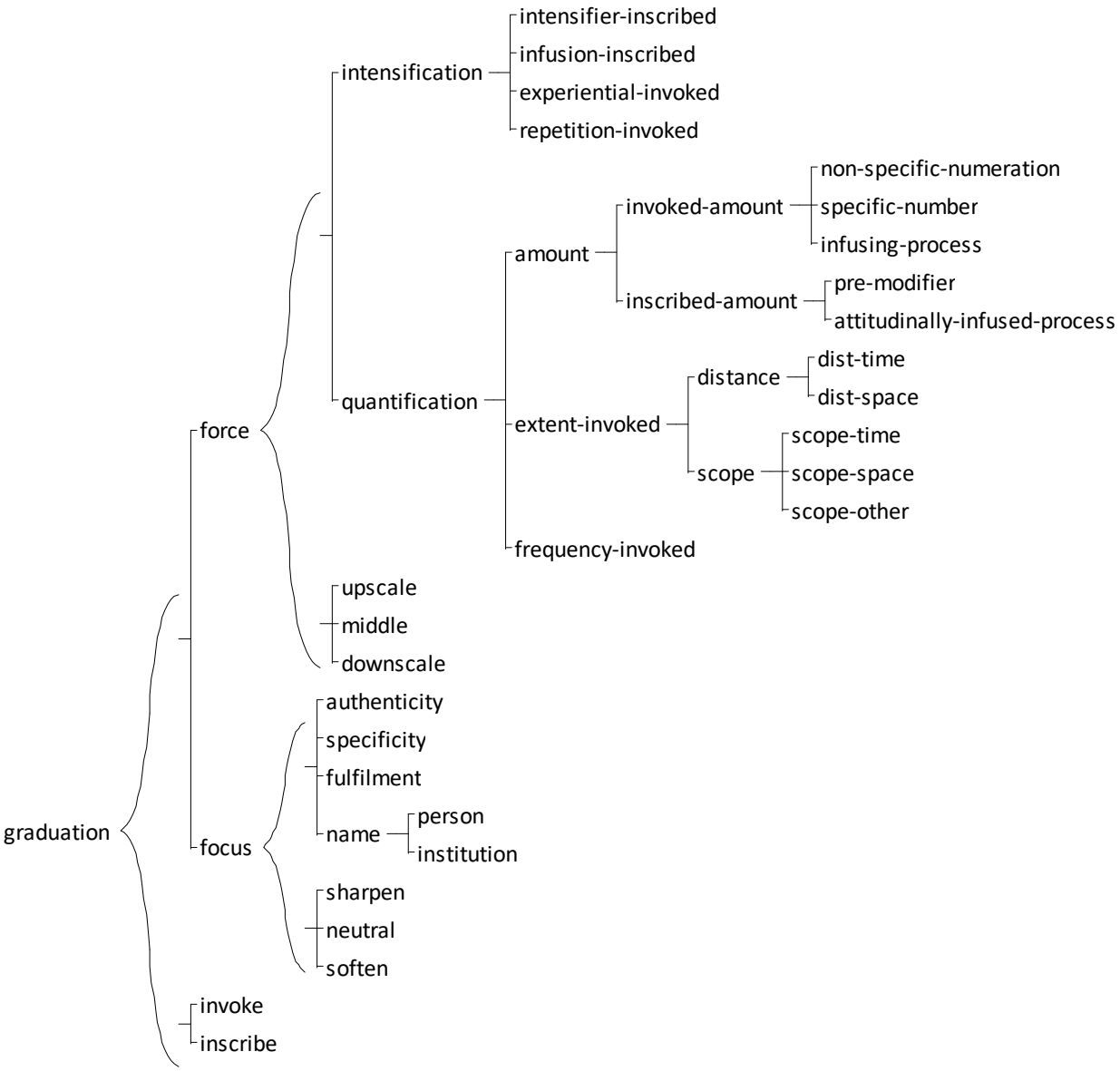


Figure 25: Coding scheme for GRADUATION in impact case studies

3.2 Definition and explanation of categories in the scheme

This section explains and defines the categorisations in the coding scheme represented in Figure 25. These definitions are accompanied by illustrative examples from the literature (Martin and White 2005; Hood 2010; Xu 2017) and the pilot corpora, but not the sub-corpus that was ultimately used for analysis; examples were added throughout the Pilot phase,

which used small sub-corpora from the same pool as the texts included in the final analysis (see thesis section 4.3.3 for the sample and 7.1.3 for the process of piloting and analysis).

3.2.1 FORCE

Expressions of Force receive two labels: one for the type of force (described in this section) and one for the direction. Once the type label has been applied, the direction needs to be indicated through a separate label: does the expression upscale or downscale the attitude? Sometimes there is a GRADUATION resource without indicating a specific direction. In these cases, middle should be selected. Especially for impact case studies, it is not always possible to be certain whether a resource is intended to show upscale or downscale.

1) FORCE:INTENSIFICATION modifies the quality of the attitude and usually applies to adjectives, but also occasionally to nouns and verbs. Note though that pre-modifiers of nouns may also be a quantification – e.g. in *considerable interest*, the pre-modifier increases the amount of interest, compared to *very interesting*, where the pre-modifier intensifies the quality of the adjective. It can be applied both to inscribed resources, where the word that is graded is attitudinal in itself (a and b), and to resources invoking GRADUATION, where the word that is graded is non-attitudinal, that is, experiential/factual (c and d).

   a. INTENSIFIER: this is where a separate word is used to show GRADUATION, e.g. *VERY / SLIGHLTY / SOMEWHAT important*

   b. INFUSION: this is where the GRADUATION is part of the word choice, rather than through adding an additional word. This usually happens where the word is part of a series of words on a scale, e.g. *crucial>important>negligible*. The word then usually has another meaning beyond the scaling and derives its graduating meaning from its position relative to other, similar words. Other examples are: *advantage, powerful, fascinating, innovative, successful, essential, improve*

   c. EXPERIENTIAL applies where there is intensification of a term that does not carry attitude in itself, and therefore the attitude is invoked through the use of an intensifying resource (which may itself be attitudinal). E.g. *REINFORCE understanding*

d. REPETITION: Here, GRADUATION is shown through repetition of words with similar meanings. This also applies where the text lists several examples where one might have been sufficient, with the assumed aim of emphasising the range of activities and thereby intensifying the claim.

Martin and White (2005: 145) point out that only certain verbal groups can be intensified like this: those of affect (e.g. *this greatly troubles me*), attitude (e.g. *improve*), transformation (e.g. *increase, reduce*) and conation (e.g. *help, hinder*). Other types of verbs, such as those of motion or perception, cannot be scaled up or down with grammatical intensification; they can be scaled with lexical adverbs (e.g. *closely observe*).

2) FORCE:QUANTIFICATION modifies the quantity of the attitude, or may invoke attitudinal meaning by adding a quantifying expression which in the context of an impact case study will be understood as evaluative (making an argument for the scale of either the problem or the solution). Most of the different labels here are applied to invoked resources, unless specified.

a. FORCE:QUANTIFICATION:AMOUNT normally applies to a noun and includes number and mass.

i. FORCE:QUANTIFICATION:AMOUNT:INSCRIBED applies to GRADUATION of words with attitudinal meaning, realised through one of two grammatical means:

1. A PRE-MODIFIER, e.g. *GREATER competence, CONSIDERABLE interest, NOT ENOUGH evidence, a SIGNIFICANT problem* (note though that "significant" can also be intensification – use the context to determine whether it is quantitatively or qualitatively significant; see the specific decisions in 3.3 below)

2. An ATTITUDINALLY INFUSED PROCESS, that is, a verb that carries attitudinal or quantitative meaning, e.g. *alleviate* (this indicates downscaling, e.g. of pain), *expand*

ii. FORCE:QUANTIFICATION:AMOUNT:INVOKED applies where the mere addition of quantification can be seen as adding evaluative meaning. This

applies where the modified word is not in itself attitudinal, and can be achieved in the following ways:

1. NON-SPECIFIC NUMERATION usually as pre-modifier, e.g. *many, more, few, a crowd, increase (noun), also, as well as*
2. A SPECIFIC NUMBER
3. An INFUSING PROCESS, that is, a verb that carries quantitative meaning and applies to a non-attitudinal noun, e.g. *BROADENING understanding*. This is similar to the last point above (ATTITUDINALLY INFUSED PROCESS in AMOUNT:INSCRIBED), with the difference being whether or not the noun carries attitudinal meaning in itself or whether the attitude is INVOKED by the addition of the graduation-infused verb.

b. FORCE:QUANTIFICATION:EXTENT and FREQUENCY are probably the most straightforward labels to apply, as they relate to recognisable quantifying expressions such as geographical extent or time (both in length and specific points). The challenge here is to interpret the direction of GRADUATION. Should "national" be seen as upscaling our downscaling in the context of an ICS? The assumption is that all such labels are applied in order to upscale, but the reader may not perceive them as such. In discussions, the coders decided that "England", "Wales" etc., "UK" and "National" etc. should be coded MIDDLE; any unit smaller than that to be coded DOWNSCALE; anything international to be coded as UPSCALE (though note that geographical reach is only one kind of reach assessed in REF). Specific decisions deviating from this rule warrant a comment to explain why a certain word would be perceived differently in that context.

   It is unlikely that an ICS would intend to downscale the extent of its impact (or the problem, or the research basis) but it might happen in low-scoring ICS. If in doubt, MIDDLE is probably more appropriate than DOWNSCALE.

   i. DISTANCE:TIME: e.g. *recent, in 2010* – the closer to 2014 the expression is, the more likely it is to be DOWNSCALE or MIDDLE rather than UPSCALE, but decide for each instance based on context. *2008* is likely UPSCALE

because it was the start of the eligibility period, and if a text highlights that, they might be highlighting the maturity or longevity of the impact.

ii. DISTANCE:SPACE: e.g. *near, far*

iii. SCOPE:TIME: e.g. *long-lasting, short-term, since 2010, over X years, between X and Y, longitudinal, now* (upscale)

iv. SCOPE:SPACE: e.g. *wide-spread,* an enumeration of *countries/cities,* adjectives such as *local/national/international, widely adopted*

v. SCOPE:OTHER: e.g. *every, over*

vi. FREQUENCY: e.g. *often, mainly, annually*

3.2.2 FOCUS

Expressions of FOCUS also receive two labels, and similarly to FORCE, these refer to the type and the direction of the scaling, respectively.

FOCUS type generally refers to the "prototypicality" of the entity. How "real" or how "complete" is it? Its three categories, plus the ICS-specific category of "name", can be SOFTENED or SHARPENED, similar to the scaling of FORCE resources:

FOCUS:SOFTEN is like hedging or down-toning. E.g. it was *sort of* nice.

FOCUS:SHARPEN is the opposite, intensifying the declaration of focus type. E.g. *very real*.

1) FOCUS:AUTHENTICITY refers to the nature of the entity. How representative is it for its category? It can be applied to nouns and adjectives, and especially to non-scalable ones, where the scaling properties of Force do not apply and the degree to which the thing is what it is is expressed through its nature instead. It can also be applied to experiential (non-attitudinal) expressions, where a sharpening is generally seen as positive and a softening as watering down and therefore as negative (Martin and White 2005: 139).

Examples include: *sort of, kind of, more, real, proper, true, genuinely*

2) FOCUS:SPECIFICITY applies when there is a comment on whether the entity is more *general* or more *specific, particular*. It can be applied to nouns, verbs and adjectives. Other examples: *precise(ly), identify, define, clarify*

3) FOCUS:FULFILMENT can be applied to processes (verbs or nominalisations). These resources are used to indicate to what extent the process is (successfully) completed, e.g. *achieve* vs. *attempt*, or the link/reference is strong or weak, e.g. *claim, suggest, apparently*. Sometimes, especially in problem statements, a softened fulfilment can be tagged, e.g. the *need* for something.

a. Examples for FULFILMENT:SHARPEN: *achieve, manage, produced, changed, developed, enabled;* also causal links: *le(a)d to, cause, result in/from*.

b. Examples for "FULFILMENT:NEUTRAL" are: *facilitate, inform, influence, foster, nurture, encourage*

c. Examples for "FULFILMENT:SOFTEN": *attempt, indicate, assist (verb), need (noun), engage (with)*

4) To accommodate names of universities, researchers or centres, we added a new feature (NAME with a distinction into PERSON or INSTITUTION) under FOCUS – it is similar to SPECIFICITY but constitutes a special case which should not skew all the other specificity tagging. NAME will always be SHARPEN:INVOKE.

3.2.3 INSCRIBE vs INVOKE

Finally, each resource of either FORCE or FOCUS should be tagged as either INSCRIBED or INVOKED. Often the coding scheme itself shows whether a resource is INSCRIBED or INVOKED, but for the purposes of researching ICS it will be meaningful to see separately how much INVOKED and INSCRIBED GRADUATION there is in each scoring bracket. Tagging this separately is therefore helpful for analysis.

The distinction is: Does the writer show obviously that they are grading here, or is the grading more covert, simply by the fact that there is a word? For example, *greater* is INSCRIBED GRADUATION, in the sense that the writer clearly indicates that they make a comparison. By contrast, a number or even something categorical such as a country name is covert or INVOKED GRADUATION: the expression itself is not marked as scaling in either direction, although it may be quite clear from the context that it is meant in a scaling way.

FOCUS:SPECIFICITY is inscribed for expressions such as *in particular*, *specifically*, *especially*.

The vast majority of GRADUATION resources in ICS are expected to be INVOKED.

3.3 List of specific decisions made and discussed during training phase

"Since 2008" is coded as UPSCALE because this year was the start of the REF assessment eligibility period, and therefore those studies that include that year (or years shortly before that) are emphasising how long they have been having the influence claimed in the ICS.

"new" is FORCE:INTENSIFICATION:EXPERIENTIAL/INFUSION, depending on whether it modifies an experiential or attitudinal word.

"first" (also: "pioneering", "spearheaded" etc.) is coded as FORCE:INTENSIFICATION:EXPERIENTIAL-INVOKED:UPSCALE:INVOKE.

"a range of" is probably meant as UPSCALE, but because it can be modified in turn (*a GREAT/VARIED/… range of*), if it appears without such modification, it should be coded as MIDDLE.

"significant(ly)" can be FORCE:INTENSIFICATION:INTENSIFIER-INSCRIBED:UPSCALE:INSCRIBE if it pre-modifies a qualitative expression, or it can be FORCE:QUANTIFICATION:AMOUNT:INSCRIBED-AMOUNT:PRE-MODIFIER:UPSCALE:INSCRIBE if it pre-modifies something quantitative.

"distinct and material" is a direct quote from the official REF guidance (HEFCE 2011: 29) and therefore does not carry much editorial evaluation. It would probably have been included by the writer as UPSCALE but may be recognised as an empty phrase by a reader intimately familiar with REF guidance, therefore it is coded as FORCE:INTENSIFICATION:INTENSIFIER-INSCRIBED:**MIDDLE**:INSCRIBE. (This example occurred in Pilot 2 file 31212, see Appendix A for file numbering.)

"reduction" and similar expressions constitute upscaling benefit through the downscaling of a problem. On its own, "reductions" would be DOWNSCALE, but unitisation should include the problem that is being reduced, so that these instances can be coded as UPSCALE which is what this should be understood as.

"informed/influenced" should be tagged as FOCUS:FULFILMENT:NEUTRAL:INVOKE because these words indicate partial contribution. This may of course have been the appropriate way in a

research-to-impact context, but from a perspective of FULFILMENT, the actions of informing and influencing do not express completion of an action (such as implementing policy).

“change(s)” can be either FOCUS:FULFILMENT:SHARPEN:INVOKE if the expression refers to the impact, or FORCE:INTENSIFICATION:INFUSION-INSCRIBED:MIDDLE:INSCRIBE if it refers to the problem statement or research finding.

“lead to” / “led to”, “resulted in/from”, “caused”, “based on” are coded as FOCUS:FULFILMENT:SHARPEN:INVOKE

“key”: FOCUS:AUTHENTICITY:SHARPEN:INVOKE (unless appears in e.g. “key areas”, where it has the meaning of “important”)

“identify”: FOCUS:SPECIFICITY:SHARPEN:INVOKE

“focus”: FOCUS:SPECIFICITY:SHARPEN:INVOKE

“recommendations”: FOCUS:FULFILMENT:SOFTEN:INVOKE