ISSN: 2690-0726

Research Article

Journal of Educational & Psychological Research

How Do Different Groups Judge the Quality of Research Questions Which Inform Evidence-Based Policymaking?

Magda Osman* **and Nick Cosstick**

Judge Business School, University of Cambridge, UK

***Corresponding Author**
Magda Osman, Judge Business School, University of Cambridge, UK

Submitted: 2024, May 27; **Accepted**: 2024, Jun 20; **Published**: 2024, Jun 25

Citation: Osman, M., Cosstick, N. (2024). How Do Different Groups Judge the Quality of Research Questions Which Inform Evidence-Based Policymaking? *J Edu Psyc Res, 6*(2), 01-15.

Abstract

Science-policy co-production depends on successfully coordinating exchanges between different researchers and policymakers—acknowledging that they may vary in their interpretation of the problem and the questions that need addressing. In the UK, 'Areas of Research Interest' (ARI) are questions generated by government departments, agencies, and public bodies to invite responses from external experts, such as researchers. There are two broad aims, to communicate the information needs of government departments and to initiate a co-productive process. But are such questions assessed in the same way by policymakers and researchers? The present study examines the properties of questions to understand whether there is agreement across different groups (i.e. public N = 383, academia N = 182, public administration N = 211) regarding the types of questions which are judged to be better than others. The study presented participants with seven types of questions (Instrumental/Procedural, Causal Analytic, Verification/Qualification, Explanation/Example, Explaining/Asserting Value Judgments, Comparisons, and Forecasting) on the same topic (i.e. climate change) that varied in length (i.e. long vs. short), and that presented as either posed by policy professionals or researchers. Participants were required to assess questions based on quality of communication, neutrality, and overall goodness. The findings show that assessments were unaffected by proposer, sample, and demographics (e.g. age, gender, level of education). Of the seven types of questions investigated, Instrumental/Procedural type questions were rated the best. The implications of these findings are considered with respect to co-production between academia and policy.

Keywords: Co-Production, Questions, Public Policy, Evidence-Based Policy, Expertise

1. Introduction

The focus of this study is how the quality of research questions (designed to inform evidence-based policymaking) is judged by policymakers and researchers (as well as the general public). Following Nurse review's recommendation that UK government departments should maintain "'statements of need', in terms of the most important research questions confronting" them, departments, agencies, and public bodies started to generate 'areas of research interest' (ARI) documents [1]. The immediate aim of these documents (and the questions contained therein) is to communicate the evidential needs of government departments [2,3]. For Boaz and Oliver, ARIs allow researchers (and other stakeholders) to "better understand" how these institutions "think about" the problems they face, so that the researchers might better understand how to help solve these problems [3]. However, the documents themselves indicate greater ambitions. As of 12 November 2023, a majority of the ARI documents [https://www.gov.uk/government/collections/areas-of-research-interest] indicate that engagement and collaboration are desiderata in this context. On this basis, ARI questions are generated in the hopes of initiating some sort of co-productive process between policymakers and academic researchers—two different expert groups. The basic logic underpinning this study is that, if the questions are appraised in the same way, then this provides a common ground for policymakers and researchers from which co-production can take place. The assumption is that such a common ground is more likely to produce fruitful exchanges and outputs than co-productive activity which proceeds from divergent judgments. Furthermore, the extent of any alignment between these expert groups is made easier to assess via comparison with non-experts too.

1.1. The Nature of ARI Questions

Nurse's recommendation was for government departments to maintain lists of their most important "research questions", and a majority of the ARI documents [https://www.gov.uk/government/collections/areas-of-research-interest] (as of 12 November 2023) use this wording [1]. However, Oliver et al, argue that they are not, in fact, research questions. Often, academics describe ARI

questions as "poorly written research questions" [2]. This ignores the constraints under which these questions are generated [3]. For example, if a government department is interested in a particularly sensitive policy area (perhaps, for example, linked to defence capabilities), this might lead to the generation of unclear or vague questions—since they have two goals: signal an information requirement regarding the sensitive area and prudence. For this reason, they prefer the term 'research needs', since, they believe, it "helps to give the impression that there is a process attached to them, that they are valued, and broader than research questions". They are clearly right that these questions are generated to achieve multiple aims, only one of which is to signal information requirements—their epistemic mission. This must be considered in any study of ARIs. Yet, even paradigmatic research questions are often generated to achieve multiple aims, including non-epistemic aims (e.g. capture the epistemic mission of a research program and maximize chances of research funding). Furthermore, a research question should ideally be clear and specific. Yet, we don't see this as a necessary condition for qualifying as a research question: many questions generated to capture an epistemic mission will not meet this condition. Thus, we don't see the need to abandon the term used in most of the ARI documents. [This is an issue regarding which, we believe, reasonable people can disagree. Moreover, different terminology may be suitable for different projects.]

1.2. Co-Production

The term 'co-production' was first used in the public administration literature to refer to the way in which those in different organisations contribute inputs in the production of a public service, such as education, or good—in particular, where service/good *users* also aid in production [4-11]. Later work on the co-production of knowledge—or, more weakly, information—was arguably in keeping with this original characterization, since knowledge and information can be thought of as public goods [12-15]. [Stiglitz characterizes knowledge (in terms of information) as an impure public good, since it is not fully non-excludible: people can, to some extent, be excluded from the use of knowledge/information, by secrecy [15]. However, the knowledge/information created by academics, as opposed to government scientists, is generally publicly available.], [Yet, in focusing on goods rather than services, it arguably moved away from the emphasis of the original literature [4].] Lemos and Morehouse went beyond this, characterizing co-production in terms of iterative interaction between information producers and users, which results in "an actual re-shaping of both groups' perceptions, behaviour, and agendas that occurs as a function of their interaction" [16]. The ideal result, concerning knowledge production, would be two-way information exchange and the adaptation of research to fit the needs of its users (including the required information, its understandability, its timeliness, and its accessibility). Thus, this characterization—the iterative-interactive' characterization —champions a model of communication which moves beyond the one-way transfer of information from research producer to research user, to the cooperative shaping of agendas, research, and/or policy [16-19]. The iterative-interactive lens has proved particularly popular in humanities and social-science research concerning climate change [14,16,17,20-22]. Indeed, it was found to be "by far the most widely used" in Bremer and Meisch's review [17].

The official guidance document on writing and using ARIs highlights several aims for ARI documents. One aim fit with the iterative-interactive characterization: "to foster a culture of using research and innovation within the department that sustains a *continuous dialogue with producers of research".* [Our emphasis.] Furthermore, as noted in the previous section, a majority of the ARI documents indicate the aim for engagement and collaboration beyond a one-way information transfer. Yet, the iterative-interactive characterization appears too strong as a model for collaboration started by the formulating of research questions. To weaken it slightly, we can think of iterative interaction between research producers and users which involves a two-way information exchange, resulting in the re-shaping of both groups' perceptions, behaviour, agendas, and/*or* knowledge production.

What factors influence the success of such a co-productive process? One clear factor is communication. Policymakers and researchers would engage in this process via some sort of verbal and/or written communicative interaction. The communicative interaction of multiple agents counts as a discourse when they (a) all bring their own presuppositions to the table as their *assumed* common ground, with (b) additions made to the *collective* common ground as the discourse proceeds, but (c) only if each agent's contributions are fully understood [23]. Without (c), the agents' beliefs regarding their (assumed) common ground will diverge and the accumulation of collective common ground (b)—mutual understanding—will slip through their grasp. There are two necessary conditions for reaching full understanding regarding the contributions made to a discourse [23]. Firstly, mutual effort: in specifying one's points and attempting to understand those of one's interlocutor. Secondly, 'grounding': the contributor and partner(s) must believe that they have understood the information contained in the contribution sufficiently for the current purpose. The development of common ground is a requirement for cooperative action, especially the action of contributing to a discourse [23,24]. In the context of science-policy interaction, meeting the conditions for a discourse means being part of a two-way information exchange in which both parties develop an accurate, shared account of the information. This makes the desired results easier to attain.

The amount of effort which agents collectively expend in their discourses is governed by the 'principle of least collaborative effort': "participants try to minimize their collaborative effort—the work that both do from the initiation of each contribution to its mutual acceptance" [23,25]. Agents' communicative behaviour accords with this principle because the phrasing of a contribution often requires collaboration. (E.g. the contributor may simply be ignorant of the precise phrasing(s) that their partner(s) would accept.) This means that effective communication can be attained despite any constraints on agents' effort.

Successful iterative-interactive co-production is a useful framework for the interaction of researchers and policymakers

for a myriad of reasons. Of salience are those which concern (i) differences between the views and roles of these two groups, (ii) constraints upon policymakers' decisions, and/or (iii) constraints upon policymakers' power [26]. An example of (i) is the difference between researchers' and policymakers' views regarding what counts as 'good' evidence [26]. An example of (ii) is the cognitive constraints on policymakers (*qua* human agents)—both in terms of limited working-memory capacity and their use of heuristics to make difficult decisions under time and resource constraints [27-42]. An example of (iii) is the insight from the modern policy-studies literature that policymakers do not straightforwardly 'control' the policy process—the policy problem is often not defined clearly, the opportunities for the application of evidence are often unclear, and single moments of authoritative choice are rare [26].

To the extent that researchers and policymakers can move beyond their differences for the benefit of public policy (and research relevant to real-world problems), it requires the kind of two-way information exchange that (communicatively successful) iterative-interactive co-production promotes. Such iterative collaboration is also required for policymakers to explain—and researchers to understand—the constraints upon their decisions and power. Thus, while such a process is not *sufficient* for a well-functioning system of collaboration between researchers and policymakers, the literature suggests that it is *necessary.*

The importance of co-production to science-policy interaction is illustrated by considering its adoption as a model in tackling the science and policy around climate change. The complexity of the problem posed by climate change is such "that neither decision makers nor scientists working alone can specify what science products are needed, how they should be developed, and how they should be applied" [18]. In contrast, co-production allows researchers and policymakers (along with key stakeholders) to cooperatively specify "the scope and context of the problem", identify important research questions, determine relevant methods and evidence, and assess the practical value and applications of the research [16, 18, 43].

1.3. Questions: Type vs. Theme

Research questions play several roles in science-policy co-production. Firstly, they can be generated to inform one community about the needs of the other, potentially setting the stage for co-productive activity. Secondly, they can be generated with the hope of *initiating* co-productive activity. Thirdly, researchers and policymakers can *identify* them via co-productive activity. Additionally, co-productive activity might also be used to amend or edit the research questions (generated to inform and/or initiate co-production). A research programme in psychology has generated several iterations of a taxonomy for categorising questions by type/style: the structure of the information sought [44, 45]. A question's type is distinct from its theme [46, 47]. For example, the questions 'how does inflation work?' and 'how can inflation be decreased?' both have the same theme: inflation. Yet, they ask for information on this topic which is structured in different ways—the former asking for an *explanation* of inflation, the latter for an *instrumental account* of lowering it [46]. Recently, Graesser et al.'s taxonomy of question types has been adapted in the policymaking context. Osman and Cosstick utilized a dataset of 2927 questions posed by over 400 policymakers to researchers over a 10-year period to generate a refined version of Graesser et al.'s taxonomy: the 'taxonomy of policy questions' [46]. When the dataset of policy questions was categorized by type and theme, they found that—regardless of the policy theme—the most frequent question type deployed by policymakers was the 'Instrumental/Procedural' type. In other words, policymakers typically ask questions which invite practical solutions. By contrast, researchers typically ask questions which invite answers which explain the factors associated with certain mechanisms or outcomes and/or forecast possible outcomes which might follow from certain interventions [48].

1.4. Hypotheses

This study adds to the literature on these topics by investigating how the quality of research questions (designed to inform evidence-based policymaking) is judged. As noted, UK government departments, agencies, and public bodies ARIs to inform others about their information requirements, and to initiate co-productive exchange between academia and policy. Yet, we do not know what types/styles of question are judged to be good, nor whether different groups of people—with varying expertise in devising research questions—agree (or not) regarding what makes a good research question. Four hypotheses were generated regarding these issues.

1. Overall, there will be greater favourable judgments towards the shorter compared to the longer research questions.
2. Overall, there will be greater favourable judgments towards research question types that request answers that explain concepts (Explanation/Example type) and the mechanisms behind concepts (Causal Analytic type) compared to the remaining five other types of questions (Forecasting, Comparison, Explaining/Asserting Value Judgments, Verification/Qualification, and Instrumental/Procedural). [Explaining/Asserting Value Judgments does not explain concepts or their mechanisms, as it is focused on soliciting advice: 'How should the infrastructure available be used to produce X? How should X respond to Y?' (Osman and Cosstick, 2022a).]
3. Participants will agree on which types of research questions are judged to be good independent of who (policy professionals/academic researchers) has posed the research questions.
4. Group and individual differences will not reliably predict favourability of the research questions above and beyond differences based on the length and type (based on the seven types of questions as classified by the Taxonomy of Policy Questions) of question.

The hypotheses were pre-registered (see supplementary document 'What are the factors that determine the types of questions that people commonly think make for a good question for scientists to answer?') to protect against cherry-picking findings. Hypothesis

1 reflects findings regarding human agents' limited working-memory capacity [28-37] . Hypothesis 2 reflects theoretical work which postulates that human agents are intrinsically geared towards seeking (causal) explanations [49-53]. Hypotheses 3 and 4 are more exploratory, but generally assume that, if there are fundamental aspects of the way research questions are appraised, then extraneous details regarding the source (i.e. the proposer) and group differences would not influence the general patterns of appraisals of questions. Thus, hypothesis 3 reflects a standard assumption of no difference between populations—plus the intuition that question types/styles are the crucial factor in judging the quality of questions. Whereas, hypothesis 4 reflects the assumption that that length (due to limited working-memory capacity) and type (due to work in the psychology of questions) are the important properties of questions [44-46].

2. Research Design

Basic Details: This study was based on an online experiment which utilized real examples of research questions that are published by UK government departments (e.g. Ministry of Justice and the Department for Transport), agencies (e.g. Food Standards Agency and the Health and Safety Executive) and public bodies (e.g. the National Archives). [https://www.gov.uk/government/collections/areas-of-research-interest] A list of all 18 government departments, agencies, and public bodies plus the research questions compiled from them (for the periods 2017 to 2021) be found in the supplementary document 'Aris Complete set FIRST and SECOND RATER 25_03_2022'. The details regarding how the 2105 ARIs were then filtered to generate the 14 questions that were included in the final online experiment are presented in subsection 2.2.

The current experimental set up involved three independent variables: question type/style, question length, proposer of the question (i.e. policymaker or researcher). The experimental design was mixed, with between and within participant manipulations. Participants—of which there were three samples: Public, Professional in Academia, and Professional in Public Administration/Public Policy—were randomly assigned to one of four conditions, in which question length and proposer was manipulated to be able to compare the impact of length of question (i.e. short vs. long) and proposer of the question (i.e. policymaker or researcher) on the key dependent variables. For each of the seven types of questions, participants were asked to make judgments which formed the main dependent variables, of which there were a total of 10. The online study was fully randomized: the order of the presentation of the seven questions was randomized for each participant, as was the order of presentation of the 10 assessments (see Figure 1). The online study was implemented in mid-July 2022 and all data collection was completed by the end of July 2022.

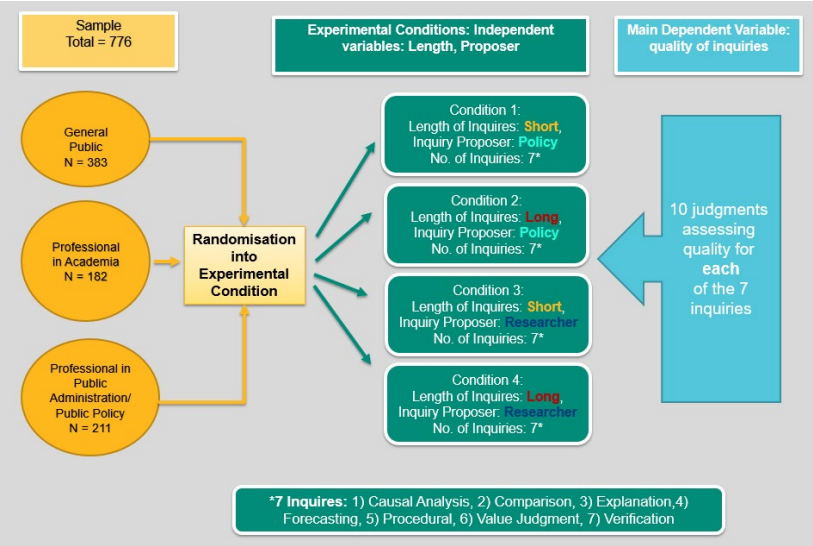

Figure 1: Schematic of Experimental Set Up

2.1. Study Sample

Participants: The present study included a total of 776 participants, of which there were three different samples: Public [Included to determine whether there judgment patterns in experts versus non-experts.] (Total N = 383 [originally 400; 17 were excluded for not completing the experiment]), Professionals in Academia (N = 182 [originally 200; 18 were excluded for not completing the experiment]), Professionals in Public Administration/Public Policy (Total N = 211 [originally 220, nine were excluded for not completing the experiment]). The demographic details, by sample, are presented in Table 1. The study was presented via Qualtrics (https://www.qualtrics.com/uk/) an online platform for hosting experiments—and used a crowdsourcing system (Prolific; https://www.prolific.co/) to recruit participants. The process of participant recruitment was volunteer sampling via Prolific Academic. The inclusion criteria were that participants were born and currently reside in the UK, that they fell within the age range 18–80, and that their first language is English. The residency was important

given that the questions referred to policy issues specific to the UK. The details of the study were posted on Prolific Academic and participants that were interested in taking part where then assigned to take part.

All participants were financially compensated for their time (2.30 USD) for 15 minutes. When taking part in the study, participants were asked to provide responses to three demographic questions (age, gender, and education level), these are summarized in Table 1 for each sample. The study received ethical approval from the Judge Business School University of Cambridge ethics board (Code: 22-24, 25-5-2022).

The rationale for recruiting participants from the aforementioned three samples is that they could reasonably be expected to provide participants with varying degrees of experience in devising research questions. In this respect, the present study would be the first of its kind to investigate whether there is agreement regarding the properties of questions which hold value, across different groups with different degrees of expertise. For this reason, the main sample were the Public—who might exhibit a greater variability in the assessment of the questions compared with Professionals in Academia and Professionals in Public Administration/Public Policy. Prolific Academic allows for inclusion criteria based on profession. For Professional in Academia, the inclusion criteria were that participants should work in either higher education or a research institute. For the Public Administration/Public Policy sample, the inclusion criteria were that participants should work in policy, public policy, or public administration.

		Public N = 378	**Professionals in Academia N = 182**	**Professionals in Public Administration/ Public Policy N = 211**
Age	Please indicate your age in the box below, or else select "prefer not to say"	*M* = 42.25 *SD* = 13.42 *range* = 18–74	*M* = 41.35 *SD* = 11.37 *range* = 26–76	*M* = 42.82 *SD* = 10.90 *range* = 26–67
Education	Please indicate your highest level of education in the box below, or else select "prefer not to say"	GCSE/A Level = 120 College = 38 Undergrad = 160 Postgrad = 57 Prefer not to say = 4	GCSE/A Level = 17 College = 11 Undergrad = 63 Postgrad = 90 Prefer not to say = 1	GCSE/A Level = 47 College = 22 Undergrad = 102 Postgrad = 39 Prefer not to say = 1
Gender	Please indicate the gender you identify with, or else select "prefer not to say"	Men = 193 Women = 181 Prefer not to say = 5	Men = 93 Women = 89 Prefer not to say = 0	Men = 101 Women = 107 Prefer not to say = 3
Attitude to Climate Change	I care about the environment', scale of 1 completely disagree–to–10 completely agree	*M* = 8.05 *SD* = 1.97	*M* = 8.21 *SD* = 1.92	*M* = 8.08 *SD* = 1.94
	I regularly take action to reduce my carbon footprint', scale of 1 completely disagree–to–10 completely agree	*M* = 6.31 *SD* = 2.42	*M* = 6.55 *SD* = 2.13	*M* = 6.46 *SD* = 2.37
	Scientific evidence points to a warming trend in global climate, scale of 1 completely disagree–to–10 completely agree	*M* = 8.42 *SD* = 2.07	*M* = 8.79 *SD* = 1.91	*M* = 8.23 *SD* = 2.26
	Human activity is responsible for the continuing rise in average global temperature, scale of 1 completely disagree–to–10 completely agree	*M* = 8.04 *SD* = 2.18	*M* = 8.39 *SD* = 2.08	*M* = 7.97 *SD* = 2.25
Total Climate Change Score (out of 40)		*M* = 30.50 *SD* = 7.86	*M* = 31.94 *SD* = 7.12	*M* = 30.73 *SD* = 7.46

Table 1: Basic Demographic Questions, and General Attitudes Towards Climate Change

2.2. Dependent and Independent Variables

Independent variables: Question type, Question length. To generate the research questions that would be used as materials for the online experiment, the criteria were as follows. To test for hypothesis 2, all research questions had to concern the same theme, with at least one question falling into each of the seven types (styles; Instrumental/Procedural, Causal Analytic, Explaining/Asserting Value Judgments, Forecasting, Comparison, Explanation/Example, and Verification/Qualification) based on the Taxonomy of Policy Questions [46, 48]. Of the seven themes (i.e. Employment, Transport, Business and Economy, Environment, Education, Energy, and Health), only one generated enough questions from each of the seven types: Environment. In addition, to test for hypothesis 1, example questions needed to either be short (i.e. 10 to 15 words) or long (i.e. 30 to 38 words). The mean word length was analysed to determine questions that were on the shorter end of the distribution of words, and the longer end (Table 1). Given these criteria, out of the 2105 ARIs, 441 contained terms that were associated with climate change, sustainability, or the environment. From the 441, questions were selected if they were within the range of 10–15 words, or 30 to 38 words, this narrowed the number of possible questions down to 155, and from those, there had to be a question from each of the seven subordinate categories, of which there were only 2–5 of each in some categories to choose from. This can be summarized in a selection criterion which incorporated those questions that were as closely related to each other as possible, where any or all of the key terms (i.e. climate, environment, sustainable, and/or net zero) were referred to. This criterion was used to generate the final set of 14 questions presented in Table 2.

Type of Question	Short question	Long question
Causal Analytic	How do climate risks interact with socio-economic factors and vulnerabilities?	How will agriculture affect the resilience to climate change of surrounding habitats and communities—for example, water availability, flooding, land use change, chemical harm on ecosystem functions related to climate resilience?
Comparison	What are the risks and opportunities for the UK across economic, social, and environmental dimensions?	What are the barriers and opportunities for commercialising advanced nuclear technology in the UK and overseas, considering public/consumer attitudes to new uses of nuclear (e.g. industrial heat, desalination, hydrogen production)?
Explanation/ Example	What are the present weather and climate risks globally and within the UK?	What pressures will there be on global natural resources—especially energy, food, water and critical elements—in the short, medium, and long term, and with what strategic policy implications for the UK in a changing world?
Forecasting	Might other future trends in society impact the justice system, for example climate change and technological advances?	On global megatrends (e.g. digitalisation, decarbonisation, demographics, new modes of transport), how can we measure and monitor the impacts of new technologies and emerging industries? How can we identify persistent under-adoption of technologies?
Instrumental/ Procedural	How can we assess and mitigate systemic risks involving environmental factors?	How can we most effectively implement nature-based solutions, such as tree planting and peatland restoration, to address climate change, support progress to net zero carbon emission, reduce biodiversity loss and prevent poverty?
Explaining / Asserting Value Judgments	What is the fate of hydrogen in the environment, and its effect on climate and the ozone layer?	To what extent have international climate finance programmes, covering themes such as technical assistance, cities, forestry, decarbonisation and storage, and private finance, achieved their objectives and contributed to wider departmental and global climate goals?
Verification / Qualification	Are consumer perceptions of carbon emissions for different journeys accurate?	Does local generation lower the costs of moving to a net-zero emissions economy, or are the necessary electricity system upgrades required to electrify of heat and transport so dramatic that the system upgrades are required in any scenario?

Table 2: Materials for the Main Experimental Set Up; Context = Environmental Issues Concerning Anthropogenic Climate Change

Independent Variable: Proposer. To test hypothesis 3, the experimental conditions varied such that half of the participants in the study were informed that the questions were proposed by policymakers, and the rest were instructed that the questions were proposed by researchers. In this way, the design of the experiment was set up so that participants were randomly allocated to one of four conditions. Condition 1: research questions posed by policymakers (short questions, 10–15 words long); Condition 2:

research questions posed by policymakers (long questions, 30–38 words long); Condition 3: research questions posed by academic researchers (short questions, 10–15 words long); Condition 4: research questions posed by academic researchers (long questions, 30–38 words long).

Dependent Variables. Graesser et al. have proposed the Grasser, Person and Huber (GPH) scheme, which includes two general features of questions that enable an overall assessment regarding whether a question is presented in a manner that makes it valuable [44]. First, type/style ("content"): the structure of the information sought. Second, "question-generation mechanism": the psychological processes—goals, plans, and knowledge—which bring about a question. The GPH Scheme also lists four specific properties that consider the question-generation mechanisms: (1) reducing, or correcting, a knowledge deficit; (2) monitoring common ground; (3) social coordination of action; and (4) control of conversation and attention.

Until now, no analysis of research questions has been based on the GPH scheme in the domain of policy questions that have a research component to them. With this in mind, the overall objective was to present participants with an online survey comprised of judgment probes that invite them to consider a research question from the perspective of Grasser et al.'s question-generation mechanisms [44]. From Grasser et al.'s work, the following 10 judgments assessing core dimensions of questions that indicate their quality were used to form the main dependent variable assessing the quality of questions, grouped according to communication quality (judgments 1–5), neutrality (judgments 6 and 7), and overall goodness (judgments 8–10) (Table 3) [44]. The correspondence between the item (i.e. judgments 1–10) and three basic dimensions (i.e. neutrality, communication quality, and overall goodness) were determined statistically, which will be discussed in more detail in subsection 2.3.

Dimension	Assessment type	Judgment
Communication Quality	1) Reducing, or correcting, a knowledge deficit	To what extent is this question presented in a way that helps to increase knowledge of the topic being referred to? Scale: 0 (highly unsuccessful in advancing knowledge) – 100 (highly successful in advancing knowledge)
Communication Quality	2) Reducing, or correcting, a knowledge deficit	To what extent is this question presented in a way that helps to reduce a knowledge gap in the topic being referred to? Scale: 0 (highly unsuccessful in reducing knowledge gaps) – 100 (highly successful in reducing knowledge gaps)
Communication Quality	3) Monitoring common ground	To what extent does this question enable the answerer to know what is needed to address the question? Scale: 0 (highly unsuccessful in communicating the answer needed) – 100 (highly successful in communicating the answer needed)
Communication Quality	4) Monitoring common ground	To what extent does this question communicate essential information needed to answer it? Scale: 0 (highly unsuccessful in communicating critical) information) – (100 highly communicating critical information)
Communication Quality	5) Monitoring common ground	To what extent is this question phrased in a way that signals what the questioner wants as an answer? Scale: 0 (no signal as to an expected answer) – 100 (strong signal as to an expected answer)
Neutrality	6) Social coordination of action	To what extent is this question phrased in a neutral way for the answerer to address? Scale: 0 (not at all neutral) – 100 (highly neutral)
Neutrality	7) Social coordination of action	To what extent is this question phrased in a way that could pressurize the answerer to give an answer they don't want to give? Scale: 0 (not at all pressurising on the answerer) – 100 (highly pressurising on the answerer)
Overall goodness	8) General judgment 1	To what extent is this question phrased in a way that is persuasive as to the importance of the topic? Scale: 0 (not at all persuasive) – 100 (highly persuasive)
Overall goodness	9) General judgment 2	To what extend is this question worth answering? Scale: 0 (not at all necessary to answer) – 100 (extremely necessary to answer)
Overall goodness	10) General judgment 3	To what extent is this question good? Scale: 0 (not at all good) – 100 (exceptionally good)

Table 3: Set of Judgments Participants Are Invited to Make for Each of the Seven Questions Presented

In addition, to examine hypothesis 4, the online experiment required that participants provide responses to a number of questions that concerned demographic details (e.g. gender, age, and educational level) and their general attitudes towards climate change. The items selected to assess general attitudes towards climate change were based on a scale developed by Bissonnette and Contento and Sinatra et al., and used by Osman and Thornton (Table 1) [54-56]. The aim was to examine the impact of group differences based on demographics, as well as general interest in the theme to determine the extent to which personal interest and motivations to respond to anthropogenic climate change would influence the way in which questions were assessed.

In summary, the main manipulations in this online experiment were designed to test four hypotheses. These hypotheses were pre-registered, which is important given that this study is the first of its kind to examine the impact of different features of questions on the assessment of their quality. To achieve this, the core manipulations were based on questions that have been published and designed to invite interest from researchers to provide evidence that would inform policymaking. Therefore, the materials are based on actual questions that are currently of interest to policymakers in the UK. To this end, to assess how good the questions are, each participant was presented with one example from each of seven types of questions, based on a newly developed taxonomy. These examples varied in length, with some being short and others long. We investigated the extent to which the assessments of the questions may be biased by who is posing the questions (policymakers versus researchers). For each of the seven types (styles) of questions, participants were invited to assess them based on 10 types of assessment that concerned neutrality, quality of communication, and overall goodness. Finally, the aim was to also look at the extent to which demographics determine differences in the way questions are assessed, along with personal interests in the topic area of the questions.

2.3. Method of Analysis

Assessments of quality of questions: The simplest way to generate an overall quality score of the research questions was based on summing judgments across all 10 assessment types (Table 3) and doing this separately for every question (total score out of 1000). Note that for items (assessment types) 5 [The reason why item 5 responses needed reversing is that, based on the original formulation of this item, strongly signalling the intended response implies biasing an answer rather than presenting a question in a neutral tone as to not steer the recipient to respond in a given direction.] and 7, the responses were reversed to ensure that all responses were in the same direction.

In addition, to determine the extent to which the items corresponded to the three dimensions (quality of communication, neutrality, and overall goodness), the responses from all participants (N = 766) were collapsed across all seven question types (styles; Table 2) to determine an aggregate score for each of the 10 items for each participant (range of score of each item 0 to 700). The 10 items were then correlated with each other. Correlations between items that were => r .5; large effect size) were treated as corresponding to the same dimension; in fact, many of the items correlated at large effects (e.g. r = .75 to r = .92) [57]. For each of the three dimensions presented in Table 3, the items corresponding to each were highly correlated with one another but not the remaining items. Therefore, the groupings of items were conceptually and statistically determined.

Classification of Item type: Given that the seven question items were classified according to the Taxonomy of Policy Questions, there are two levels of classification (superordinate, subordinate) [46, 48]. Thus, the superordinate level of comparison was based on collapsing scores (averaging across) assessments for items that fell under the category of Bounded (i.e. Verification/Qualification, Comparison, and Forecasting) and Unbounded (i.e. Instrumental/Procedural, Explanation/Example, Explaining/Asserting Value Judgments, and Causal Analytic). In this way each judgment of quality (i.e. quality of communication, neutrality, overall goodness) could be compared for items belonging to either one of the superordinate categories. The subordinate level of comparison was based on all seven question types (styles) examined based on dimensions of quality, which enables a fine-grained level of analysis that could also investigate the hypotheses being tested.

The results from inferential analyses (presented in section 3) are only reported generally when they reached a level of moderate to large effect sizes [57]. The reason for this is based on the various concerns regarding statistical reporting of *p*-values, and a better indication of how to interpret the findings is gained from effect sizes [58-62]. Determining the effect size is a way to quantify the degree to which results from experiments deviate from a null hypothesis relative to a population. Also, reporting effect sizes gives a better indication of the practical value of an experiment [63]. The main inferential analyses used were Analysis of Variance (effect size ranges for $\eta 2$ are: 0.06 = moderate effect, 0.14 = large effect), t-tests (effect size ranges for *d*' are: 0.5 = moderate effect, 0.8 = large effect) and regressions (effect size range for β weights are: 0.1 to 0.5 = moderate effect, and > 0.5 = large effect).

3. Results

Superordinate types of questions: To examine hypotheses 1 and 3, starting with the superordinate category (Bounded versus Unbounded), we looked at each dimension of quality (neutrality, quality of communication, and overall goodness; Table 3) to determine the extent to which assessments varied by length, by proposer, and by sample. Looking at Figure 2, the general patterns suggest that there do not appear to be substantive differences in the way the three samples assessed the questions. This was borne out in the analysis, which did not reveal any effects that reached moderate to large effect sizes for each of the three dimensions examined. Comparisons between different dimensions were not conducted—because the scoring of each dimension differed by the number of items that were included in it, there would be uninteresting differences given that a total score for neutrality was 200, a total score for quality of communication was 500, and a total score for overall goodness was 300. Therefore, the presentation of

the analyses based on the superordinate category of questions is separated by each dimension of quality.

For assessments of neutrality, the main between-subject effects (sample [public, public administration/public policy, academia], question length [short, long], proposer of the question [Policy Professional, Researcher]) did not generate any moderate or high effect sizes. Examining within subject effects, there were also no main effects regarding type of question between Unbounded (N = 775, *M* = 108.91, *SD* = 33.37) and Bounded questions (N = 775, *M* = 112.80, *SD* = 35.56), *F* = 98.4, *η2* = .03 (see Figure 2).

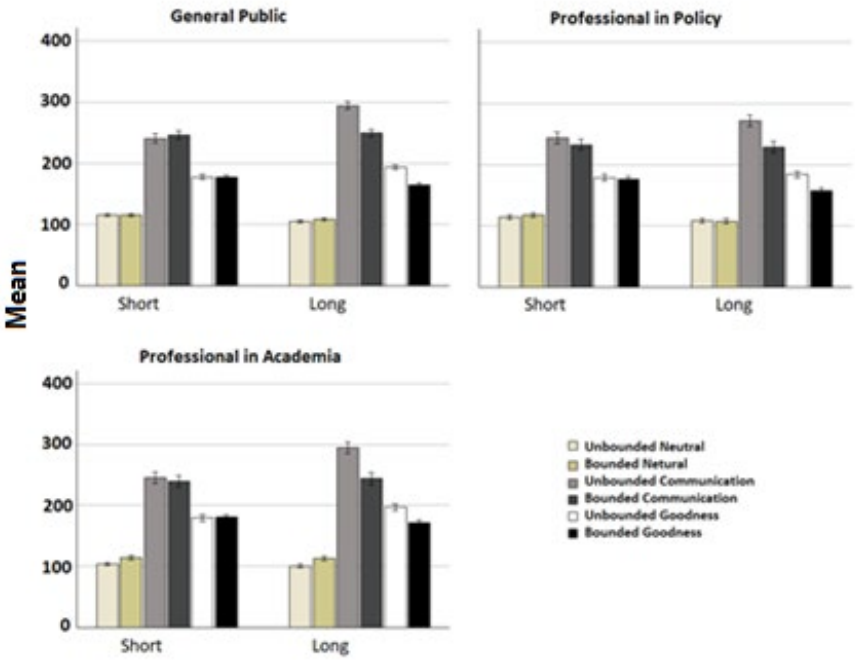

Figure 2: Mean (+/-1 SE) Scores for Unbounded and Bounded Items, by Sample, by Length of Item and by the Three Different Dimensions of Quality (Neutrality, Quality of Communication, and Overall Goodness)

The same analysis was conducted for assessments based on quality of communication. Again, no between-subject effects reached a level of moderate or high effect sizes. There was a large effect of type, suggesting that Unbounded questions (N = 775, *M* = 264.65, *SD* = 96.11) were judged more favourably based on quality of communication than Bounded questions (N = 775, *M* = 241.32, *SD* = 94.55), *F* = 116.76, *η2* = .13. There was also an interaction between type and length, *F* = 82.07, *η2*= .10 (See Figure 2). For short Unbounded (N = 402, *M* = 242.62, *SD* = 96.29) and Bounded questions (N = 402, *M* = 240.69, *SD* = 96.37), there appeared to be little difference in assessments of quality of communication. But long Unbounded questions (N = 373, *M* = 288.00, *SD* = 90.30) were judged better than Bounded questions (N = 377, *M* = 241.99, *SD* = 92.68).

The same analysis was conducted for assessments of overall goodness (the third dimension of quality). As with the other dimensions of assessment, no between-subject effects reached a level of moderate or high effect sizes. There were moderate effects of type—again, in line with the same pattern found for assessments based on quality of communication. For overall goodness, Unbounded questions (N = 775, *M* = 184.83, *SD* = 50.82) were (on the whole) judged to be better than Bounded questions (N = 775, *M* = 171.26, *SD* = 51.14), *F* = 78.98, *η2* = .09). There was also an interaction between type and length, *F* = 69.12, *η2* = .08 (see Figure 2). For short Unbounded (N = 402, *M* = 178.50, *SD* = 53.62) and Bounded questions (N = 402, *M* = 177.53, *SD* = 51.98), there appeared to be little difference in assessments of overall goodness. But long Unbounded questions (N = 373, *M* = 191.65, *SD* = 46.72) were assessed better than long Bounded questions (N = 377, *M* = 164.51, *SD* = 49.40). Also, there was a substantive difference between short and long Bounded questions—where shorter questions were judged to be better than longer questions.

Overall, we find that for all three dimensions of quality, there was no substantive differences based on the proposer of the questions, supporting hypothesis 3. There were also no substantive differences based on the sample, which also further suggests that there was general agreement across participants. This is an important finding, given the different types of expertise and experience that we could reasonably expect the samples had with devising research questions Therefore, regardless of the differences, the results show that there are general indicators based on the phrasing of the questions that were used by all three samples to provide similar judgments.

Regarding assessments of neutrality, it appears that there was no overall difference—so neutrality did not discriminate between types of questions as a measure. Quality of communication and

overall goodness did, however, discriminate between types of questions. For both dimensions, Unbounded questions were judged to be better. In particular, long Unbounded questions were judged overall better than long Bounded questions. There was only partial support for hypothesis 1. The findings indicate that length played a role, its impact on the assessment of questions depending on the type. Generally, long Unbounded questions were judged better both on quality of communication and overall goodness. However, short Bounded questions were judged as better on overall goodness than long Bounded questions.

Subordinate types of questions: The next step was to examine hypotheses 1 and 3 by considering the subordinate category, thus the analyses now consider the seven different types/styles of questions (Table 2). To enable comparison between these types, the overall quality score was used. This means that each participant generated a quality score of each question (from a range 0 to a total of 1000), the mean scores across participants by sample and for each question and by length are presented in Figure 3. Looking at Figure 3, the general patterns suggest that there do not appear to be substantive differences in the way the three samples assessed the questions, this was borne out in the analysis which did not reveal any effects that reached moderate to large effect sizes for each of the three dimensions examined.

Consistent with the patterns reported for the superordinate item analysis, when looking at all seven subordinate items, the main between-subject effects (sample [public, public administration/ public policy, academia], question length [short, long], proposer of the question [Policy Professional, Researcher]) did not generate any moderate or high effect sizes. However, there was a main effect concerning type of item, $F = 88.66$, $\eta 2 = .10$, and an interaction between type and length, $F = 59.94$, $\eta 2 = .07$ (see Figure 3). Therefore, pairwise comparisons were conducted to determine where the differences were located. Taking effect sizes into account, the only comparison that reached a moderate effect size (df = 775, $d' = 0.54$) was between Explanation/Example ($M = 80.13$, $SD = 14.64$) and Verification/Qualification ($M = 491.25$, $SD = 165.37$). While there were differences between items based on assessments of overall quality, no others reached the threshold of moderate effect size. For example, Instrumental/Procedural ($M = 559.23$, $SD = 165.97$) and Verification/Qualification, (df = 775, $d' = 0.41$), and Causal Analytic ($M = 548.16$, $SD = 157.43$) and Verification/Qualification, (df = 775, $d' = 0.36$).

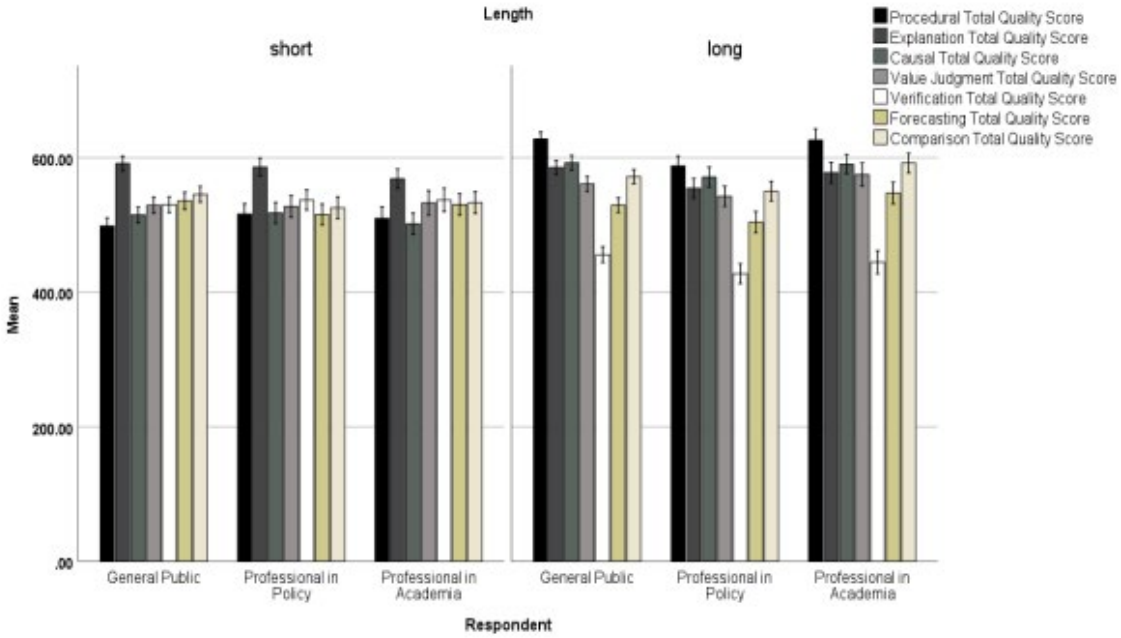

Figure 3: Mean (+/-1 SE) Scores of Overall Quality Score (Out of A Total Of 1000) For Each of the 7 Questions, by Sample, and by Length of Item

Independent t-tests, revealed that, for Instrumental/Procedural questions, long (N = 374, $M = 616.47$, $SD = 164.93$) was judged better than short (N = 402, $M = 505.98$, $SD = 147.12$, $d' = .71$), but for Verification/Qualification questions, short (N = 402, $M = 533.79$, $SD = 161.51$) was judged better than long (N = 374, $M = 445.53$, $SD = 157.27$, $d' = .55$). While just missing the threshold, for Causal Analytic questions, long (N = 374, $M = 586.18$, $SD = 144,45$) was judged better than short (N = 402, $M = 512,80$, $SD = 160.73$, $d' = .48$).

Overall, based on the analyses of subordinate items, we find support for hypothesis 3, indicating that the type of proposer made little difference to assessments of items, and that the pattern was also the same across samples. Here also there was partial support for hypothesis 1. For the Instrumental/Procedural type, and to some extent the Causal Analytic type, longer questions were judged better than shorter questions, bearing out the pattern based on superordinate analyses of which both items belong to. However, for the Verification/Qualification type, shorter questions were judged better than longer questions— again, bearing out the

findings based on the superordinate analyses. Therefore, the impact of length depends on the type of questions; it is not a *general* indicator of quality. Moreover, the findings did not fully support hypothesis 2. Explanation/Example questions were judged better than Verification/Qualification questions, and to some degree so were Instrumental/Procedural and Causal Analytic type questions. Yet, for most of the questions the overall assessment of quality did not lead to vast differences—with the aforementioned exception.

Regression analyses: To test hypothesis 4, regression analyses were performed based on predictors (age, gender, level of education, length of question, proposer of question, sample, and overall attitudinal score of climate change). The predictors were regressed on to the overall quality of assessment of each of the seven questions. For all items, there was a positive relationship between attitude towards climate change and overall assessments of the items; Explanation/Example (β = 8.4), Causal Analytic (β = 7.8), Instrumental/Procedural (β = 6.7), Comparison (β = 6.6), Explaining/Asserting Value Judgments (β = 5.3), Forecasting (β = 5.3), Verification/Qualification (β = 4.2). That is, the items were likely to increase in favourable assessments overall as attitudes towards addressing climate change increased.

For Explaining/Asserting Value Judgments items, and Forecasting items, there was a negative relationship between age—such that as age increased, overall assessments of the items decreased. Finally, for Causal Analytic and Instrumental/Procedural questions there was a positive relationship between length and overall assessments, such that long items were judged more favourably than short items, but the reverse was found for Verification/Qualification items.

4. General Discussion

This study had two main aims: 1) to examine what properties of research questions are judged to be good, and 2) to explore whether different groups of people—with varying expertise in devising research questions—agree (or not) regarding what makes a good research question. Moreover, the empirical study, designed to address these aims, is the first of its kind to use genuine examples of research questions that are published by UK government departments.

The present study showed that the length of the question also impacted assessment of quality (hypothesis 1 was unsupported). Longer-worded research questions were judged better than shorter-worded ones, though this was only true for Unbounded types—specifically Instrumental/Procedural and Causal Analytic questions. In contrast, shorter-worded Verification/Qualification research questions were judged better than longer-worded ones. Taken together, these findings are somewhat counterintuitive; given a wealth of cognitive-psychological research regarding limited working-memory capacity and the extensive use of heuristics, one might predict that their judgments of questions will exhibit cognitive miserliness—resulting in a preference for more succinct communication. What the present findings imply is a more nuanced sort of heuristic: where a detailed answer is required, the questions need to be more detailed; where a short answer is required, the questions need to be more succinct (the 'detail heuristic'). Therefore, the contextual information requirement is a cue to the quality of the different possible questions one might pose, so the length of the question should correspond to the contextual information requirement.

Unbounded research questions were judged better (based on quality of communication and overall goodness) than Bounded research questions. More specifically, when compared against each other, of the seven types/styles (see Table 2) of research questions that were examined, Explanation/Example type research questions were assessed as the best overall, and Verification/Qualification type research questions faired the least well in overall score of quality (partial support for hypothesis 2). Overall, participants were not significantly biased by the proposer of the research questions, and so there were no differences in assessment of quality based on whether the questions were thought to come from policy or academia (support for hypothesis 3). Given the range of expertise, and the differences in age, gender, educational background, and attitudes towards taking action in response to anthropogenic climate change, only the latter seems to have any corresponding relationship to assessment of quality of questions (support for hypothesis 4). Rather than group differences, the findings revealed that the more involved in actively addressing anthropogenic climate change, the better the overall score of quality of questions—with the closest correspondence being for Explanation/Example and Causal Analytic research questions. This seems to indicate that the more invested in climate change people are, the more favourably they will judge questions where explanation of situations related to climate change, as well as impacts on the environment and efforts to encourage sustainability are sought—along with questions that consider understanding of the mechanisms, consequences, and antecedents.

The remainder of this discussion focuses on three issues, the first is the implications of these findings for policy studies, especially concerning co-production; the second is the implications for science-policy interaction; and finally, a discussion of the limitations of this work along with future directions that could be taken.

To begin, the finding that participants were not biased by the proposer of the research questions seems important for policy studies and advice regarding science-policy interaction generally—and co-production in particular. Cairney and Kwiatkowski advise that researchers should adapt their framing of evidence to the cognitive biases of policymakers (*qua* human agents) [27]. The absence of proposer bias means an absence of bias in the other direction: from the responder (academic or policymaker), aimed at the proposer. Thus, it is one less thing to have to adapt to and mitigate—and one less thing to work into policy-studies models on science-policy interaction.

The detail heuristic is of particular interest when considering the literature on communication and co-production. Its use does not contradict the principle of least collaborative effort. Instead, it

provides a more detailed account of what this principle means in practice—in the context of questioning and answering as part of a discourse. The collaborative effort to establish that the questioner is searching for detailed information (and for the answerer to understand the nature of this request) and for the answerer to provide this information (and for the questioner to understand it) is greater than that needed in the case of a search for less detailed information. It also provides an important contextual amendment to advice—offered in the policy studies literature—directed at researchers who wish to engage with policymakers. Cairney and Kwiatkowski argue that researchers ought to tailor their approach to the cognitive and contextual constraints placed on policymakers [27]. For example, researchers should synthesize the evidence they wish to convey "concisely to minimize its cognitive burden". Taken as a general point that one should not overcomplicate (even complex) information, it is clearly reasonable. However, our results indicate that it should not be taken to mean that less information—or less complicated information—is always (or even generally) better. The context of the information requirement informs the kind of answer that a question will invite, which in turn also implies a particular preference for how detailed the answer should be.

Relatedly, Osman and Cosstick found that, regardless of the policy subject, the most common question type/style deployed by policymakers was the Instrumental/Procedural type [46]. They related this to co-production in several ways. Firstly, researchers hoping to engage in the co-production of policy might reasonably hope to maximize their chances of success by tailoring their evidence to the question types most commonly deployed by policymakers. Secondly, since co-production requires a mutual understanding between researchers and policymakers, understanding the needs and goals of one's interlocutor might be more easily achieved by splitting this task into two: ascertaining (i) the general subject under discussion and (ii) the structure of the information sought [64]. The findings of this study are also of relevance to these points. It is an interesting question as to whether researchers would do better to tailor their evidence to the question types most commonly *deployed* by policymakers or those which policymakers typically *judge* to be higher in overall quality.

Here we present some speculations regarding practical implications based on what has been discussed. First, clearly, if researchers are asked a specific question by policymakers, then they should tailor the presentation of their evidence to its type. Yet, if researchers are just presenting to policymakers, then the correct method of evidence tailoring seems more of an open issue—given that the researcher has a responsibility to communicate their finds accurately, while also deciding which may be of most critical policy relevance. Our intuition about this open question works backwards from what such researchers hope to achieve: persuading policymakers that their *evidence* is of high quality. If researchers tailor the presentation of their evidence—or even their investigations that produce evidence—to the information types corresponding to the question types policymakers most typically *deploy*, then a potential rationale might be as follows. 'The common deployment of this question type reveals a preference for it that might be interpreted as a judgment that it is of higher quality, and if that is how they judge questions of this type, then they might judge evidence tailored to this question type more highly'. If, on the other hand, they tailor their evidence to the question types policymakers most commonly *judge* to be higher quality, then a potential rationale might be as follows. 'Policymakers typically judge questions of this type to be higher quality, so they might judge evidence tailored to this question type more highly'. Both rationales are highly speculative—there is no current evidence to support the leap from question quality to evidence quality. Yet, the former rationale contains an extra leap of judgment: the leap from a revealed preference to a formal judgment of quality. Given that formal judgments of evidence quality more closely resemble formal judgments of question quality (*qua* formal judgments) than revealed preferences for certain question types, the former rationale seems more questionable. The stronger rationale could be used to improve the chances of any researcher gaining the interest of policymakers. Posed as a simple heuristic (the 'tailoring heuristic'), tailoring evidence to information types corresponding to question types that policymakers (generally) judge to be of higher quality is more likely to persuade them that one's evidence is of higher quality (than tailoring it to other question types). Again, this 'heuristic' is highly speculative; thus, more work is needed to test whether formal judgments of question quality (or revealed preferences) are related to formal judgments of evidence quality.

The results of this study further indicate that the tailoring heuristic may be too simple. Professionals in Policy judged long Instrumental/Procedural higher in overall goodness than the other types, but not in the case of short questions—where they preferred Explanation/Example questions. Therefore, researchers engaged in evidence tailoring *may* do better to consider the complexity of the information that policymakers are interested in—judged by things such as the context of their meeting and the kinds of questions the policymakers have previously deployed. This also means that the structure of the information sought—and, therefore, the appropriate way of tailoring one's evidence—may be a function of the complexity of the information which policymakers are seeking. This is an empirical matter that can be addressed by investigating the relationship between different properties of questions and answers to explore how much a good answer, as judged by people, corresponds to how good the question is judged to be.

These results are also relevant to work on the relationship between researchers and policymakers. The 'two communities' theory holds that researchers and policymakers constitute two distinct communities that are poorly connected, motivated by different incentives, operate under different rules, and suffer from communication problems [65, 66]. This theory has been charged with inaccuracy, vagueness concerning the mechanism by which communal differences function in the disruption of research utilization, evidential inadequacy (relying on surveys or case studies rather than systematic tests), ignoring important nuances regarding these two communities, and prescriptive inadequacy [66-70]. The results from this study support a novel line of criticism.

Researchers and policymakers show remarkable similarity regarding their assessments of (superordinate or subordinate) categories of questions. Thus, any differences in connectivity and/or rules are not so great as to lead to radically different preferences. Furthermore, the similarity of their preferences suggests that questioning activity will not typically generate major communication problems.

4.1. Limitations

One question hanging over this study is whether the samples had enough expertise to be able to effectively assess the specific research questions posed to them. A current research programme has applied the principles of optimal experimental design to questioning (and other types of information-search activity) [71]. This leads them to characterize the 'normative goodness' of a question in terms of its expected epistemic utility–which has many possible measures. The question hanging over this study can be characterized in these terms: do the samples' judgments of questions—particularly those related to overall goodness (and particularly general judgment 3; Table 3)—approximate to reasonable measures of those questions' expected epistemic utility? For the moment, this question cannot be answered, due to the black-boxed nature of the samples' judgments. Thus, further work would be needed to understand the relationship between the three dimensions of assessment (communication quality, neutrality, and overall goodness) and formal measures of questions' goodness.

Another issue is that ARIs are developed for multiple purposes, only one of which is to signal an information requirement [2, 3]. Thus, it would be wrong to conclude, on the basis of this study, that ARIs ought to be revised to maximize their question quality. Question quality is only one criterion for judging ARIs; it must be balanced against other relevant criteria. However, the results of this study can help policymakers assess, and improve, how ARIs 'do' by this criterion. More broadly, they can help human agents assess, and improve, how research questions (in general) 'do' by this criterion.

A final issue is that the detail heuristic requires further empirical testing to verify its status as a true heuristic. This would include exploring its potential presence in a range of different policy research questions, as well as academically driven research questions for comparison. More theoretically, it would concern whether a precise mental model, not simply an explanation via redescription, can be generated for the detail heuristic [72]. (Though, the need for such specification has been questioned [73].)

The work proposed above may provide some insights here—especially given Meder et al.'s emphasis on finding simple heuristics which approximate to specific measures of expected informational utility [71].

5. Conclusions

This study utilized real examples of research questions—published by UK government departments—to investigate which types of questions are judged to be good. In addition, this study investigates whether different groups of people—with varying expertise in devising research questions—agree regarding what makes a good research question. To date there has been no empirical work investigating these issues, and so the findings from this study provide novel insights into both. The results indicate that, across all those that took part in the study, overall judgments were not biased by the proposer of the research questions. Another key finding was that there was considerable overlap in the way participants appraised the different types of research questions, to the extent that group differences, or demographics did not produce differences what was judged as more or less a good question.

So, overall, what type of question is a good question? The findings here suggest that given there are two broad types (Bounded: closed-type questions; Unbounded: open-ended-type questions), the Unbounded questions were judged than Bounded questions. The basis on which "good" was determined three measures: quality of communication, neutrality, and overall goodness. Another property of the question that informed the way people appraised its goodness was length. Where the intuition is that brevity is preferred over verbosity, the evidence provided a more complex picture. For Unbounded questions—specifically Instrumental/Procedural and Causal Analytic questions—longer questions were judged better than shorter ones. For bounded questions—specifically Verification/Qualification questions—shorter questions were judged better than longer ones. This suggests that people employ a type of heuristic regarding how good a question is in line with the implied level of detail needed to answer it 'detail heuristic'. While further work is needed to determine the presence of this heuristic, at the very least our findings suggest that, on the whole, people adopt a nuanced approach to assess question quality. Furthermore, where length is used as a cue, it interacts with the type of questions posed, and has practical implications. Questions which require a specific answer are judged to be good when they are styled in a *succinct way.* Questions which require a comprehensive response are judged to be good when they are styled in a way that provides *sufficient detail* to enable a relevant comprehensive response. Overall, the results from the present study are suggestive of the need for advice—aimed at helping researchers and policymakers understand the needs and goals of policymaking—to focus on the question types that policymakers typically judge to be of higher quality. However, question quality is only one criterion for judging research questions and needs to be balanced against the other important criteria. Furthermore, the major issue with the study is the open question of whether subjects' assessments of question quality approximate to reasonable measures of expected epistemic utility.

References

1. Nurse, P. (2015). Ensuring a Successful UK Research Endeavour: A review of UK research councils.
2. Oliver, K., Boaz, A., & Cuccato, G. (2022). Areas of research interest: joining the dots between government and research at last?. *F1000Research, 11*(1509), 1509.
3. Boaz, A., & Oliver, K. (2023). How well do the UK government's 'areas of research interest'work as boundary

objects to facilitate the use of research in policymaking?. *Policy & Politics, 51*(2), 314-333.
4. Ostrom, E. (1996). Crossing the great divide: Coproduction, synergy, and development. *World development, 24*(6), 1073-1087.
5. Ostrom, E., & Nevin, G. (1976). On dissemination. Administrative Report No. *6, Workshop in Political Theory and Policy Analysis, Indiana University*.
6. Bish, P., & Neubert, N. M. (1976). A preliminary inquiry into citizen contributions to community safety and security. *In Workshop in Political Theory and Policy Analysis, Indiana University.*
7. Percy, S. L. (1978). Conceptualizing and measuring citizen co-production of community safety. *Policy Studies Journal, 7*, 486-493.
8. Whitaker, G. P. (1980). Coproduction: Citizen participation in service delivery. *Public administration review, 240*-246.
9. Sharp, E. B. (1980). Toward a new understanding of urban services and citizen participation: The coproduction concept. *Midwest Review of Public Administration, 14*(2), 105-118.
10. Kiser, L. L., & Percy, S. L. (1980, April). The concept of coproduction and its implications for public service delivery. *In Annual Meeting of the American Society for Public Administration, San Francisco* (pp. 13-16).
11. Parks, R. B., Baker, P. C., Kiser, L. L., Oakerson, R., Ostrom, E., Ostrom, et al. (1982). Coproduction of public services. In: Rich, R. C. (ed.) *Analyzing urban-service distributions* (pp. 185–199). Lexington Books.
12. Jasanoff, S. (ed.). (2004). *States of knowledge: The co-production of science and social order*. Routledge.
13. Bandola-Gill, J., Arthur, M., & Leng, R. I. (2023). What is co-production? Conceptualising and understanding co-production of knowledge and policy across different theoretical perspectives. *Evidence & Policy, 19*(2), 275-298.
14. Cash, D. W., Borck, J. C., & Patt, A. G. (2006). Countering the loading-dock approach to linking science and decision making: comparative analysis of El Niño/Southern Oscillation (ENSO) forecasting systems. *Science, technology, & human values, 31*(4), 465-494.
15. Stiglitz, J. E. (1999). Knowledge as a global public good. *Global public goods: International cooperation in the 21st century, 308*, 308-325.
16. Lemos, M. C., & Morehouse, B. J. (2005). The co-production of science and policy in integrated climate assessments. *Global environmental change, 15*(1), 57-68.
17. Bremer, S., & Meisch, S. (2017). Co-production in climate change research: reviewing different perspectives. *Wiley Interdisciplinary Reviews: Climate Change, 8*(6), e482.
18. Beier, P., Hansen, L. J., Helbrecht, L., & Behar, D. (2017). A how-to guide for coproduction of actionable science. *Conservation Letters, 10*(3), 288-296.
19. Bucchi, M. (2008). Of deficits, deviations and dialogues: Theories of public communication of science. *In Handbook of public communication of science and technology* (pp. 71-90). Routledge.
20. Dilling, L., & Lemos, M. C. (2011). Creating usable science: Opportunities and constraints for climate knowledge use and their implications for science policy. *Global environmental change, 21*(2), 680-689.
21. National Research Council. (2001). Climate change science: An analysis of some key questions. *National Academies Press*.
22. Wyborn, C., Datta, A., Montana, J., Ryan, M., Leith, P., Chaffin, B., ... & Van Kerkhoff, L. (2019). Co-producing sustainability: reordering the governance of science, policy, and practice. *Annual Review of Environment and Resources, 44*, 319-346.
23. Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science, 13*(2), 259-294.
24. Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association .
25. Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1-39.
26. Cairney, P. (2020). *Understanding public policy: Theories and issues* (2nd ed.). Bloomsbury Academic.
27. Cairney, P., & Kwiatkowski, R. (2017). How to communicate effectively with policymakers: combine insights from psychology and policy studies. *Palgrave Communications, 3*(1), 1-8.
28. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review, 63*(2), 81-97.
29. Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences, 24*(1), 87-114.
30. Cowan, N. (2005). Working memory capacity. *Psychology Press*.
31. Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why?. Current directions in psychological science, 19(1), 51-57.
32. Baddeley, A. (2001). The magic number and the episodic buffer. *Behavioral and Brain Sciences, 24*(1), 117-118.
33. Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological science, 15*(9), 634-640.
34. Jefferies, E., Ralph, M. A. L., & Baddeley, A. D. (2004). Automatic and controlled processing in sentence recall: The role of long-term and working memory. *Journal of memory and language, 51*(4), 623-643.
35. Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: a reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1235-1249.
36. Chen, Z., & Cowan, N. (2009). Core verbal working-memory capacity: The limit in words retained without covert articulation. *Quarterly journal of experimental psychology, 62*(7), 1420-1429.
37. Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(1), 37-55.

38. Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.
39. Gigerenzer, G., & Selten, R. (2001). Rethinking rationality.
40. Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology, 3*(3), 430-454.
41. Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review, 80*(4), 237-251.
42. Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
43. Kates, R., Clark, W. C., Corell, R., Hall, M., Jaeger, C. C., et al. (2000). Sustainability Science. Research and Assessment Systems for Sustainability. In *Program Discussion Paper 2000-33. Environment and Natural Resources Program, Belfer Center for Science and International Affairs*. Kennedy School of Government, Harvard University Cambridge.
44. Graesser, A. C., Person, N., & Huber, J. (2013). Mechanisms that generate questions. In *Questions and information systems* (pp. 167-188). Psychology Press.
45. Graesser, A. C., McMahen, C. L., & Johnson, B. K. (1994). Question asking and answering. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 517–538). Academic Press.
46. Osman, M., & Cosstick, N. (2022). Finding patterns in policy questions. *Scientific Reports, 12*(1), 20126.
47. Pomerantz, J. (2005). A linguistic analysis of question taxonomies. *Journal of the American Society for Information Science and Technology, 56*(7), 715-728.
48. Osman, M., & Cosstick, N. (2022). Do policy questions match up with research questions? No.1. *Centre for Science and Policy, University of Cambridge Working Paper series*.
49. Gopnik, A. (1996). The scientist as child. *Philosophy of science, 63*(4), 485-514.
50. Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn*. William Morrow & Co.
51. Weiner, B. (1985). " Spontaneous" causal thinking. Psychological bulletin, 97(1), 74-84.
52. Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive development, 7*(2), 213-233.
53. Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. Monographs of the Society for Research in Child Development, 72(1), vii–ix.
54. Bissonnette, M. M., & Contento, I. R. (2001). Adolescents' perspectives and food choice behaviors in terms of the environmental impacts of food production practices: application of a psychosocial model. *Journal of nutrition education, 33*(2), 72-82.
55. Sinatra, G. M., Kardash, C. M., Taasoobshirazi, G., & Lombardi, D. (2012). Promoting attitude change and expressed willingness to take action toward climate change in college students. *Instructional Science, 40*, 1-17.
56. Osman, M., & Thornton, K. (2019). Traffic light labelling of meals to promote sustainable consumption and healthy eating. *Appetite, 138*, 60-71.
57. Cohen, J (1988). *Statistical power analysis for the social sciences* (2nd. ed.). Lawrence Erlbaum Associates.
58. Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology, 36*(1), 1-2.
59. Trafimow, D. (2019). A frequentist alternative to significance testing, p-values, and confidence intervals. *Econometrics, 7*(2), 26.
60. Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology, 45*, 19-27.
61. Trafimow, D., & Marks, M. (2015). *Editorial. Basic and Applied Social Psychology, 37*, 1-2.
62. Trafimow, D., & Marks, M. (2016). Editorial. *Basic and Applied Social Psychology, 38*, 1-2.
63. Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102*(4), 989-1004.
64. Nikulina, V., Lindal, J. L., Baumann, H., Simon, D., & Ny, H. (2019). Lost in translation: A framework for analysing complexity of co-production settings in relation to epistemic communities, linguistic diversities and culture. *Futures, 113*, 102442.
65. Caplan, N. (1979). The two-communities theory and knowledge utilization. *American behavioral scientist, 22*(3), 459-470.
66. Newman, J., Cherney, A., & Head, B. W. (2016). Do policy makers use academic research? Reexamining the "two communities" theory of research utilization. *Public Administration Review, 76*(1), 24-32.
67. Kalmuss, D. (1981). Scholars in the courtroom: Two models of applied social science. *The American Sociologist*, 212-223.
68. Bogenschneider, K., & Corbett, T. J. (2010). Family policy: Becoming a field of inquiry and subfield of social policy. *Journal of Marriage and Family, 72*(3), 783-803.
69. Wingens, M. (1990). Toward a general utilization theory: A systems theory reformulation of the two-communities metaphor. *Knowledge, 12*(1), 27-42.
70. Sabatier, P. (1978). The acquisition and utilization of technical information by administrative agencies. *Administrative science quarterly*, 396-417.
71. Meder, B. M., Crupi, V., & Nelson, J. D. (2024). What makes a good question? Prospects for a comprehensive theory of human information acquisition. In I. Cogliati-Dezza, C. Wu & E. Schulz (Eds.), The drive for knowledge: The science of human information-Seeking. Cambridge University Press.
72. Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review, 103*(3), 592-596.
73. Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582-591.