Title: Finding patterns in policy questions

Magda Osman[1,2], Nick Cosstick[1,2]

[1] The Centre for Science and Policy, University of Cambridge,
[2] Judge Business School, University of Cambridge

Correspondence: m.osman@jbs.cam.ac.uk, n.cosstick@jbs.cam.ac.uk

Abstract
To help advance exchanges between science and policy, a useful first step is to examine the questions which policy professionals pose to scientists. The style of a question indicates what the asker is motivated to know, and how they might use that knowledge. Therefore, the aggregate pattern of typical policy inquires can help scientists anticipate what types of information policy audiences desire. A dataset (n = 2972) of questions from policy professionals collected over 10 years (2011–2021)—by the Centre for Science and Policy at the University of Cambridge—was classified into one of seven classes. In the main, the most popular questions posed by policy professionals—within the public and private sectors—were those whose answers inform how to achieve specific outcomes—whether directly, or by providing a causal analysis which is instrumental to this process. Moreover, this seems to be a general aspect of policy professionals' inquiries, given that it is preserved regardless of the policy issue considered (e.g., Artificial intelligence, Economy, or Health). Thus, maximizing the usefulness of the information that policy professionals receive when engaging with scientists requires informing how to achieve specific outcomes—directly, or by providing a useful causal analysis.

Introduction
Knowledge exchange activities between science and policy are driven by a need to address practical issues [1]. Several studies have highlighted barriers to effective translation from scientific evidence to policy [2],[3]. In particular, scientists and policy professionals often have different motivations and goals, which limit their collaboration [4]. Whilst the prospects for bringing these two communities' motivations and goals into complete alignment are poor, scientists might reasonably gain a greater understanding of the motivations and goals of policy professionals—and how the evidence they generate feeds into, or helps to achieve, them. Examining the questions which policy professionals pose to scientists is instrumental to achieving this greater understanding. The 'style' of a question—the structure of the information sought (see below)—provides valuable indicators of what the asker is motivated to know, and what they might use that knowledge for [5], [6]. Indeed, science-policy exchanges can often involve framing policy issues through a particular style of inquiry that articulates professionals' policy goals and the means of achieving them [7]. In light of this, the aim of this study is to examine the style—and subject matter—of questions which policy professionals pose to scientists, in order to expose any underlying patterns in these evidence requests. An understanding of the underlying patterns can potentially aid the integration of scientific evidence into policymaking through co-production (Wyborn et al., 2019). Co-production requires a mutual understanding between scientists and policy professionals (Nickulina et al., 2019), which, in turn, requires clarity regarding any subject matter under discussion and the structure of the information sought. Furthermore, the integration of scientific evidence into policymaking is aided by tailoring evidence to the structure of information sought by policy professionals. To this end, the discussion will contain some advice for evidence tailoring.

Before considering the substantive lessons to be drawn from the literature on questions, a terminological confusion must be addressed. Within this domain, a universal agreement regarding the meanings of certain relevant terms has not yet been reached. This might cause confusion, as the same term can be deployed to refer to distinct features of questions and/or answers. For example, Pomerantz [5] uses the term 'content' to refer solely to the subject matter that a question/answer concerns. By contrast—as noted by Pomerantz [5]—Graesser, McMahen, and Johnson [8] use the term 'content' to refer both to a question's subject matter and its style. When using antecedent studies to evidence the arguments in this paper, such terminological issues are disregarded in favour of the underlying point being made.

To begin, it is useful to consider two contrasting features of questions: subject matter and style. The subject matter of a sincere question indicates the information that the inquirer is interested in attaining, thereby indicating the kind of content which would be appropriate for a sincere answer [5]. For example, sincerely asking "what is net zero?" implies that one wants to know about the net-zero emissions goal. The style of a question—the structure of the information sought—indicates the understanding of the inquirer [9], and what kind of answer is expected [5],[10],[11]. Asking a sincere question implies that the inquirer has enough of an understanding of the issue from which to build the question and interpret the answer, but does not know enough to make seeking the answer superfluous [9]. For instance, sincerely asking "what is net-zero?" requires that the inquirer has at least heard of the term 'net-zero' but does not have a complete understanding of its referent. Furthermore, the style of this question indicates that sincere answers should be structured as definitions. By contrast, sincerely asking "what do we need to do to achieve net-zero?" requires that the inquirer has a basic understanding of what net-zero is but not a complete understanding of how to bring it about. Moreover, the style of this question indicates that sincere answers should outline the procedure(s) which will bring about net-zero.

Earlier work in psychology [8],[12],[13], linguistics [14], and information science [5] provided the foundation for the types of analyzes that have been used to develop taxonomies of questions. It is important to highlight here that question stems—such as "why…?", "how…?", "when…?", and "what…?" —have not been the basis on which a taxonomy is developed, because they are polysemous [10],[15]. The ambiguity of question stems makes their application highly context-specific, hence why question classification systems have generally focused on question styles [10].

The most practical approach to taxonomizing questions has been to classify questions according to their style. Lehnert [11] was the originator of this approach. Graesser, Person, and Huber [10] later generated a simpler taxonomy of questions by style (the 'GPH Taxonomy')—these questions were later grouped by the length of the expected answer by Graesser, McMahen, and Johnson [8] (see Table 1). For example, "what does X mean?" was given as part of the abstract characterization ("abstract specification") of the 'definition' style of question. Definition questions invite answers which specify the details—usually as long descriptions—that characterize a phenomenon or event. In contrast, "what caused some event to occur?" was given as the abstract characterization of the 'causal antecedent' style of question. Causal antecedent questions invite answers which outline the factors that brought an event about. Taxonomies of questions can be used to investigate applied scientific problems. In order to improve outcomes in a variety of domains, such taxonomies are used to understand how agents approach the task of structuring a problem or dilemma, what types of solutions they are expecting, and how their inquiries could be improved [6].

The GPH Taxonomy (see Table 1) has proved fairly popular. It has been successfully applied within the education sector [6],[8],[16],[17]. It has also been used as a foundation of, and supplement for, arguments made by other education researchers [18],[19],[20]. Moreover, it has played an applied, foundational,

and/or supplemental role in studies analyzing web search strategies [21],[22],[23], consumer health-related inquiries [24], interpersonal exchanges [25], and interview settings [26].

Through an analysis of the frequency of questions generated, this taxonomy has been used to determine what types of questions are most likely to appear in a particular domain. Such information feeds into proposals regarding what improvements are necessary to support an effective evidence exchange process. In the education domain—where the GPH Taxonomy has been used most often—it has aided in identifying the types of inquires made by students, so that they can then be encouraged to formulate different styles of questions which enable a more substantive understanding of a topic [6]. For instance, often the efforts have been to shift students away from verification-style questions ("did X occur?") to analytical questions—such as causal consequence-style questions ("why did X occur?")—to develop deeper understanding.

Table 1. Taxonomies of question styles

The GPH Taxonomy			Taxonomy of Policy Questions		
Super Ordinate Category	Style Category	Abstract specification	Super ordinate category	Style Category	Abstract specification
Short Answer	Verification	Did X occur?	Bounded Answers	Verification/ Qualification	Is it the case that X is here? Did X event occur? Are Xs more inclined towards y? Is X a viable version of Y?
	Disjunctive	Is X or Y the case?		Comparison	What are the strengths and weaknesses of X? What are the costs and benefits of implementing X?
	Concept Completion	Who? What? What is the referent of a noun argument slot?		Forecasting	Which areas would you foresee improving in the next 10 years? How likely is it that X will be popular in the future?
	Feature Specification	What attributes does X have?			
	Quantification	How may are there of X?			
Long Answer	Definition	What does X mean?	Unbounded Answers	Example/ Explanation	Which X is more like Y? What would be a case where Y is like X? How does X work?
	Example	What is an example of X?			
	Comparison	How is X similar to Y?			
	Interpretation	What concept can be inferred from X?			
	Causal Antecedent	What event caused X?		Casual Analysis (Antecedents or Consequences)	What are the barriers that will prevent X from occurring? What are the effects of X if it is implemented now?
	Causal Consequence	What are the consequences of X?			
	Goal Orientation	What are the motives behind X's actions?		Instrumental/ Procedural/ Enablement	How can we use X to make Y better? What would need to be incorporated to ensure that X is achieved? In what way can we measure X so that it can later be used to support y?
	Instrumental/ Procedural	What plan can allow for X to be achieved?			
	Enablement	What resource allows X to perform their action?			
	Expectational	Why did X event not occur?			
	Value Judgement	What value does the responder place on X?		Explaining/ Asserting Value Judgments	How should the infrastructure available be used to produce x? How should X respond to y?
	Assertion	The inquirer makes a statement indicating lack of knowledge			
	Request/Directive	The inquirer wants the responder to perform an action			

The theoretical underpinning of work analyzing the quality of questions has largely been informed by the 'Grasser-Person-Huber (GPH) Scheme' [6], [10]. It proposes that there are three dimensions on which a question needs to be assessed. Firstly, style ("content"): the structure of the information sought. Secondly, question-generation mechanism: the psychological processes—goals, plans, and knowledge—which bring about a question. The GPH Scheme lists four question-generation mechanisms: reducing, or correcting, a knowledge deficit; monitoring common ground; social

coordination of action; and control of conversation and attention. The scheme holds that these categories are orthogonal to the style categories, since—in theory—a style of question might be motivated by any question-generation mechanism. For example, an inquirer might ask "what are the consequences of academic freedom?" to address a deficit in their knowledge. Alternatively, the same question might be asked to monitor the extent to which they share common ground with the responder. The GPH Scheme's final dimension of assessment is 'degree of specification': the extent to which the information sought is made clear. A highly specific question is clear regarding what information is sought. Whereas, an under-specified question requires that the responder make inferences about what details are relevant to the inquirer.

Within cognitive psychology, associations have been made between the effective generation of questions and problem-solving ability, as well as the learning of complex material [6],[27],[28],[29],[30]. Within social psychology, improvements in ability regarding interpersonal exchanges has also been shown to be the result of asking good questions [31], that can increase one's likability [25],[32]. Many of the efforts to improve cognitive functions (e.g. problem solving, critical thinking, memory, and text comprehension), by improving questioning, are based on two factors. Firstly, increasing the specificity of the question to ensure that the responder has the best chance of providing answers that are directly applicable. Secondly, encouraging 'deep-reasoning questions': those which direct the inquirer to ask questions that invite a causal analysis [6]. In essence this involves considering the cause-effect relationships between variables to start examining the underlying structures that enable inferences to be made about what brings about observable outcomes [33],[34].

To date, there has been no empirical work examining the styles of questions that policy professionals pose to scientific experts—including the types of questions that are asked, and the frequency by which they are asked. Once this is understood, it can be used to improve science-policy exchanges. Improvements can be made to the articulation of policy questions so that the value of the answers provided is maximized. Furthermore, scientists might find it easier to adapt their communication, in order to focus on the evidence that policy audiences want from them. To address this deficit, the present study analyzed policy questions that have been compiled by the Centre for Science and Policy (CSaP), at the University of Cambridge. CSaP is a knowledge brokerage which creates opportunities for public policy professionals and academics—primarily scientists—to learn from each other. This is achieved through CSaP's policy fellowship programs, as well as workshops, seminars, conferences, and professional development activities. Regarding the main policy fellowship program, public policy professionals—as well as those from the private sector—initially submit a set of questions (typically between five to seven) to academics that shape the engagement with them. This is done to develop an evidence base which informs the types of policy issues that they aim to address. Thus, CSaP has accumulated a database of policy questions submitted by over 400 policy fellows over 10 years.

The database was used to examine two properties of policy questions: 1) What frequent styles of questions are posed to expert scientists? 2) Is there a relationship between the subject matter and style of questions posed to expert scientists? By answering these questions, it is possible to build up a profile of what evidence policy professionals invite scientific experts to provide, as well as what that evidence is applied to.

Methods

At the time the analysis of the questions was conducted there were a total 4319 questions for the period of 05-09-2011 to 23-09-2021. These were generated from a total of 443 different policy fellowships taken up at CSAP. The questions were then cleaned. In particular, this involved removing statements and duplicate questions (posed by the same policy fellow over consecutive trips to Cambridge). Once this filtering had been applied, there were a total of 2927 questions from 409 policy fellows that were submitted for analysis. Each question contained the following details: 1) a unique

number to identify it (1 to 2927), whether the policy fellow was from a public or private sector organization (public, private), 2) the year that the question was submitted, and 3) the word length of the question.

To ensure an appropriate classification system was applied to the questions, the questions were classified using an iterative approach [35]. First, the GPH Taxonomy was applied to all 2972 questions. This initial attempt to classify the policy questions served two purposes: to identify categories that are applicable to policy questions from the original taxonomy, and to identify new categories where needed. From this, a second taxonomy was developed, which included categories from the GPH Taxonomy as well as some new ones.

Table 2. Frequency (%) of questions per style category (* indicates categories that were omitted in the development of the revised taxonomy)

Superordinate Category	Subordinate Class	Abstract Specification	Frequency (%) of Questions Coded
Short Answer	Verification	Did X occur?	172 (5.88)
	Disjunctive*	Is X or Y the case?	4 (0.14)
	Concept Completion*	What?	5 (0.17)
	Feature Specification*	What attributes does X have?	15 (0.51)
	Quantification	How may are there of X?	58 (1.98)
Long Answer	Definition*	What does X mean?	23 (0.79)
	Example	What is an example of X?	102 (3.48)
	Comparison	How is X similar to Y?	58 (1.98)
	Interpretation*	What concept can be inferred from X?	21 (0.72)
	Causal Antecedent	What event caused X?	267 (9.12)
	Causal Consequence	What are the consequences of X?	234 (7.99)
	Goal Orientation*	What are the motives behind X's actions?	18 (0.61)
	Instrumental/Procedural	What plan can allow for X to be achieved?	219 (7.48)
	Enablement	What resource allows X to perform their action?	545 (18.62)
	Expectational*	Why did X event not occur?	5 (0.17)
	Value Judgement	What value does the responder place on X?	78 (2.66)
	Assertion*	The inquirer makes a statement indicating lack of knowledge	12 (0.41)
	Request/Directive*	The inquirer wants the responder to perform an action	4 (0.14)

The development of this revised taxonomy started from the principle that the question style categories deployed in a policy-specific taxonomy need to be useful to policy professionals. This was inferred from the percentage of questions in the database which used some style, x. Categories for which $x < 1\%$ were not carried over from the GPH Taxonomy to the revised taxonomy (see Table 2). (Where possible, the revised taxonomy subsumed questions from these omitted categories into other style categories). In addition, the taxonomy included several other categories of questions not present in the original taxonomy to reflect the kinds of questions that were frequently occurring—such as those that invited the inquirer to make forecasts.

This 'Revised GPH Taxonomy' (see Table 3) was then used to classify all 2927 questions, and two independent coders were used to validate the taxonomy. Each coded a subset of questions (n = 1224), the results of which are presented in Table 4. Applying a stringent process for agreement, with only

exact matches recorded, both coders agreed on (n = 582) 47.55% of the questions. However, the coders identified the issue that the differences between some of the Revised GPH Taxonomy's categories are superficial. For example, instrumental/procedural and enablement have superficially different subjects and predicates, yet both drive at the same idea: they seek to identify things which can be used to achieve some goal. This violates the principle of 'qualitative parsimony' (Lewis, 1973): categories should not be inflated beyond necessity. When taking this into account, along with the related point that some of the categories were broad enough to significantly span the specification of others, the next step was to determine how many of the questions coded revealed matches based on feasible overlaps. This identified an addition 331 matches between coders, which increased the total level of agreement to 74.59%. Categories with superficial differences were merged to achieve mutual exclusivity. The resulting 'Taxonomy of Policy Questions' (see Table 1) was then used to analyze the 2927 questions that are reported in the results section.

Table 3. The Revised GPH Taxonomy of question styles

Sub-Ordinate Category	Super Ordinate Category	Abstract Specification
Verification/Forced Choice (Y/N)	Short Answer	Is it the case that X is here?
Quantification	Short Answer	How many are there of X?
Qualifying Quantified Possibilities	Short Answer	What reasons are there for needing to do X?
Value Judgments	Short Answer	What value does the answerer place on an idea or advice? Which X is best?
Example	Long Answer	What is an example or instance of the instance/event/behaviour?
Comparison	Long Answer	What are the costs and benefits (or strengths and weaknesses) of X?
Explanation (open & vague)	Long Answer	What and how would X work?
Causal Consequence	Long Answer	What are the consequences of an event or state, given a set of other states or interventions?
Causal Antecedent	Long Answer	What state or event causally led/leads to an event or state or outcome of an intervention?
Forecasting	Long Answer	What will happen by time scale X?
Explaining Possibilities	Long Answer	How is it that X could be used to achieve Y?
Explaining Value Judgments	Long Answer	Why do you think X might be best if Y is used?
Instrumental /Procedural /Enablement	Long Answer	What instrument/plan/strategy allows an agent to accomplish a goal? What methods of measurement are needed/could be used to detect X?

The aim of the content analysis was to examine whether subjects lent themselves to particular question styles. This involved looking at the domains of the organizations to which the policy fellows belonged, to narrow down the subjects that informed the content analysis. There were seven common types of policy subject: Artificial Intelligence (AI), Economics and Finance, Education, Environment, Defense and Security, Health, and Technology/Manufacturing. From this, several associated topics were identified [35]. From this, each question was coded as "1" if a key subject, or associated terms for

that subject, appeared at least once in the question. For some questions, multiple associated terms were found. To avoid skewing the data in such cases, the question was still coded as "1" to reflect that it was associated with a key subject (regardless of how many other associated terms were present in that question). Trends in subjects over time were not analyzed because the policy interests/positions of the fellows is not controlled for given that policy fellow that takes up a fellowship can come from any department, in any given year. Consequently, some years the data is skewed towards different subjects by virtue of the interests/positions of the fellows at the time.

Table 4. Frequency (%) of questions when classified according to the Revised GPH Taxonomy

Sub-Ordinate Category	Super Ordinate Category	Frequency of Questions Coded (n =2927)	Subset of Questions Coded by Coder 1 (n = 1224)	Subset of Questions Coded by Coder 2 (n = 1224)
Verification/Forced Choice (Y/N)	Short Answer	297	124	101
Quantification	Short Answer	58	20	7
Qualifying Quantified Possibilities	Short Answer	142	51	5
Value judgments	Short Answer	98	38	185
Example	Long Answer	164	69	28
Comparison	Long Answer	99	32	34
Explanation (open & vague)	Long Answer	274	129	176
Causal Antecedent	Long Answer	325	110	49
Causal Consequence	Long Answer	215	96	73
Forecasting	Long Answer	158	63	76
Explaining Possibilities	Long Answer	628	304	181
Explaining Value Judgments	Long Answer	161	64	19
Instrumental /Procedural/Enablement	Long Answer	308	124	176

Results

A general point concerning the statistical analysis of the dataset was that—due to the nature of the dataset—the analyzes were ran to gather a general impression of the pattern of findings rather than to develop firm conclusions. Inferential statistics were used with caution, given that in many cases the data violated basic assumptions for running the test (e.g. independence).

To start with, while there is an uneven distribution of policy professionals by whether they belonged to private (20%) or public sector (80%) organizations, a simple analysis indicated that there is no significant difference between the two groups by frequency of class of questions, χ^2 (6, N = 2927) = 10.16, p = .12, Cramer's V = .06. Given this, for the remainder of the analyzes performed, we collapsed across the two groups of policy professionals. Generally, there were more questions that invited unbounded answers (76%) than bounded (24%), χ^2 (1, N = 2927) = 890.96, p < .001. Running a further analysis indicated there were significant differences in the distribution of questions by the seven sub-ordinate categories, χ^2 (6, N = 2927) = 110.54, p < .001, where the most frequently generated type of question was instrumental/procedural (see Table 2).

Table 5. Summary: frequencies (%) and mean word length (SD) of questions according to superordinate and subordinate categories of the Taxonomy of Policy Questions

		Bounded Answers			Unbounded Answers			
	Number of questions	Verification/ Qualification	Comparison	Forecasting	Example/ Explanation	Casual Analysis (Antecedents or Consequences)	Instrumental /Procedural	Explaining/ Asserting Value judgments
Mean word length (SD)	2927	22.58 (15.67)	22.35 (15.71)	20.24 (8.22)	14.93 (9.70)	20.00 (9.39)	22.70 (12.88)	21.87 (11.24)
% Public	2342 (80.01)	343 (14.65)	79 (3.37)	132 (5.64)	348 (14.86)	455 (19.43)	734 (31.34)	251 (10.72)
% Private	585 (19.99)	97 (16.58)	20 (3.42)	27 (4.62)	90 (15.38)	84 (14.36)	201 (34.56)	66 (11.28)
% of total sample	2927	440 (15.03)	99 (3.38)	159 (5.43)	438 (14.96)	539 (18.42)	935 (31.95)	317 (10.83)

Bounded (*M* = 22.53, *SD* = 15.66) questions are longer than unbounded questions (*M* = 20.42, *SD* = 11.49), *t* (2925) = 3.68, *p* < .001, *d* = .15, *BF* = .05, but the effect sizes and Bayes factor indicate that this is a weak difference. Looking at the average word counts for each of the seven classes of questions, example/explaining class of questions seems to be the outlier (*M* = 14.93, *SD* = 9.70). Applying the Bonferroni correction, when compared against the other unbounded answer types, the phrasing of example/explaining questions were significantly shorted than causal analysis questions, *t* (975) = 8.27, *p* < .005, *d* = .53, BF_{13}= 1.11, instrumental/procedural, *t* (1371) = 11.27, *p* < .005, *d* = .65, BF_{25}= 1.23, and explaining/asserting value judgments, *t* (753) = 9.08, *p* < .005, *d* = .67, BF_{16}= 2.41, verification/qualification, *t* (877) = 8.67, *p* < .005, *d* = .58, BF_{15}= 5.16, Forecasting, *t* (595) = 6.15, *p* < .005, *d* = .57, BF_z= 1.94, comparison, *t* (535) = 6.04, *p* < .005, *d* = .67, *BF7*= 3.45. Overall, the word length of the question doesn't appear to be indicative of the types of answers it invites.

There were two main ways in which the content of the questions was examined. The first was the frequency with which the seven key subjects appeared in the questions. The content analysis applied to identify the presence of any of the seven key subjects meant that approximately two thirds of the questions were coded by the seven main subjects (n=1786/2927, 61%). Given that a question could contain multiple subjects, when taking this into account (n = 951/2972, 32.59 %), the most commonly occurring subjects—occurring once in each question—were AI (n =222/951, 23.34%) and Technology/Manufacturing (n = 205/951, 21.56%). A total of 595 questions had a combination of two subjects present, with the most common pairing being Environment and Economy/Finance (n = 180/595, 30.25%). A total of 193 questions had three subjects in combination, with the most common triple being AI, Environment, and Economy/Finance (n = 32/193, 16.58%). A total of 42 questions had four subjects in combination, with the most common quadruplet being AI, Technology/Manufacturing, Environment, and Economy/Finance (n = 7/42, 16.67%). A total of four questions contained five subjects, and one question contained six subjects, none contained all seven subjects.

The next way in which the content of the questions was examined was how often the various question styles appeared within the set of questions for each subject. Since multiple subjects sometimes appeared within the same question, independence was violated, which prevented any categorical inferential analyzes. Nonetheless, it was possible to determine an overall impression of the most common class of questions for which different subjects appeared in. All questions were classified (n = 1786) that were coded by subject into the seven different classes of questions, and this was then repeated for questions where the subject appeared only once in each question (n = 951) [35]. Classifying the questions by subject and by style on these two sets provided a basis for determining consistency in any patterns detected. Looking across both classification methods, the most common question-style for all seven main subjects, was instrumental/procedural questions (average 34%). This may not be a surprise given the base rate of this class of question. Where subjects differed was the second most common question class that they appeared in. When classifying all questions coded by subject,

the second most common question class for six of the subjects was causal analysis (average 20%), with the exception of AI which was verification/qualification (16.67%). When classifying questions coded by subject appearing only once in a question, then the second most common class was causal analysis (average 21%) for AI, Environment, and Defence/Security. For Economics/Finance, Education, and Technology/Manufacturing the second most common class of question was example/explanation (average 19%), and for Health the second most common class of question was verification/qualification (23.14%). Overall, the indication from the examination of content by question class is that all seven subjects most commonly appeared in instrumental/procedural questions, thereafter the subjects appeared commonly in other unbounded questions-styles (e.g. causal analysis, example/explanation) with few appearing commonly in bounded question-styles (i.e. verification/qualification).

Discussion

From a database of 2927 policy questions, that were classified according to a taxonomy that has its roots in research on the psychology of questions, we find that: 1) the two most frequent question-styles invite answers that address causal-analytic and instrumental/procedural matters; 2) regardless of the policy subject, the most common answer that policy professionals invited informed instrumental/procedural questions. This indicates that the common questions that policy professionals present in exchanges with scientists are deep-reasoning based questions, that aren't just for reducing or correcting knowledge deficits, but to presented knowledge for specific purposes, such as informing what policy interventions could be taken.

This has clear prescriptive implications for scientists who wish to participate in the co-production of policy—and specifically the integration of scientific evidence into policymaking. By tailoring their evidence to these most common policy question styles, scientists might reasonably hope to maximise their chances of success. The abstract specifications of these styles can be used for this purpose. For example, scientists might ask themselves: is there an obvious policy goal that this research might help to achieve? However, it might be necessary to update evidence tailoring to meet the specific interests of any policymakers they engage with.

The fact that the two most frequently generated classes of questions were causal-analytic (e.g. understanding mechanisms) and instrumental/procedural (e.g. interventions) reveals important information regarding the main motivations and interests of policy professionals. Inviting answers that expose cause-effect relationships between variables is also key to examining the underlying structures that enable inferences to be made about what brings about observable outcomes [33],[34]. This aligns closely with work in cognitive psychology that examines causal reasoning, which has shown consistent improvement in the way a causal-analytic representation can impact decision-making [34],[36],[37], problem solving [38],[39], moral reasoning [40],[41], perception [42],[43], interpretation of statistical information [44],[45], and evidential reasoning [46],[47]. Recently, the application of causal-analytic approaches has been extended to policymaking [48],[49],[50],[51],[52]. This work suggests that, in order to interpret the effects of a policy intervention, what is first needed is to decompose the context in which the intervention is introduced into its causal factors (i.e. the variables that will support as well as inhibit the efficacy of the intervention). Achieving this requires formulating questions that concern the mechanisms which can bring about change in a desired direction and what outcomes need to be measured to determine the causal link between the intervention and the outcome. Thus, there is a clear relationship between causal analysis and instrumental/procedural/enablement reasoning: the former is a means to the latter. Scientists wishing to engage with policymakers might keep this framework for interpreting the effects of a policy intervention in mind.

As the results show, professionals do invite answers that are causal-analytic in nature, but they are half as popular as questions which invite answers that inform how to achieve specific outcomes (i.e.

instrumental/procedural). This finding can be contextualized in light of the relationship between causal analysis and procedural reasoning: the most popular questions posed by policy professionals––within the public and private sectors—were those whose answers inform how to achieve specific outcomes—whether directly, or by providing a causal analysis which is instrumental to this process. Given the importance of causal analysis in determining the potential success of policy interventions, scientists might consider framing their answers via causal-analytic terms. Furthermore, depending on the reception to their instrumental/procedural questions, policy professionals may consider increasing the number of causal analysis-style questions they ask.

Policy professionals' preference for asking instrumental/procedural questions is also relevant to several of the academic literatures on policymaking. It is consistent with several characterisations of 'evidence-based policy', which give a (greater or lesser) role to instrumental rationality in policymaking (Schwandt, *The Centrality of Practice to Evaluation;* Sanderson, *Making Sense of 'What Works')*. It also provides some backing for the claim that scientists and policy professionals are divided by goals and motivations ([4]: Hetherington and Philips, Nathan Caplan, *The Two Communities Theory and Knowledge Utilization, 1979*). Which, in turn, might explain the poor fit between researcher's assembling and packaging of information and policy professionals' practical needs (Head, *Reconsidering evidence-based policy: Key issues and challenges)*.

Other insights from the analysis of the questions suggest that more unbounded (long, open ended) questions were generated than bounded (short, closed) questions. Given that there were fewer classes of bounded question-styles to unbounded question-styles, one might think it inevitable that fewer bounded questions would be identified. However, previous work provides evidence which suggests that, independent of the number of categories of questions corresponding to short vs. long answer types—depending on the domain—more short answer types are generated than long answer types. Domains in which this seems to hold include: education [6], interviewing [26], and health inquiries [24]. Thus, the fact that more unbounded than bounded question-styles were generated may reflect the properties of the domain in which the questions were asked—the policy domain—rather than the taxonomic structure used to classify the questions. In the main, policy professionals tended to ask questions directed towards detailed answers. Often this meant steering away from constraining answers to provide an estimate about a future outcome (forecasting), verifying a particular understanding of an issue (verification/qualification), or outlining the strengths/weaknesses costs/benefits of a particular issue (comparison). The prescriptive lesson for scientists wishing to participate in the co-production of policy is that less constraint might be required to address policymakers' needs.

A final point to consider concerns important barriers that limit potential science-policy exchanges, thereby retarding the co-production of policy. Since co-production requires a mutual understanding between scientists and policy professionals (Nickulina et al., 2019), understanding the needs and goals of one's audience is an important barrier which must be surmounted. Splitting this task into understanding the subject matter under discussion and the structure of the information sought might aid its completion. Another potential barrier is that scientists have concerns about the possible blurring of lines between acting in the capacity of providing expertise versus advocating [4],[53],[54]. The findings from this study suggest that this may be warranted, given that the most common type of answer which policy professionals invited from scientists was one that involved suggestions for interventions (e.g. methods of measurement, plans of action, types of instruments) that serve particular goals. While addressing questions of this type may lead to more impact for the scientific knowledge provided, it may potentially draw scientists into making recommendations, rather than presenting policy professionals with factors to consider.

Data availability

The datasets generated and/or analyzed in the study are available in the Open Science Framework repository (https://osf.io/c978s/files/).

References

[1] Hoppe, R. Rules-of-thumb for problem-structuring policy design. *Policy Design and Practice* **1(1)**, 12-29 (2018).

[2] Singh, G. G. *et al.* A more social science: barriers and incentives for scientists engaging in policy. *Frontiers in Ecology and the Environment* **12**, 161–166 (2014).

[3] Weichselgartner, J. & Kasperson, R. Barriers in the science-policy-practice interface: toward a knowledge-action-system in global environmental change research. *Global Environmental Change* **20**, 266–277 (2010).

[4] Hetherington, E. D. & Phillips, A. A. A scientist's guide for engaging in policy in the United States. *Frontiers in Marine Science* **7**, (2020).

[5] Pomerantz, J. A linguistic analysis of question taxonomies. *Journal of the American Society for Information Science and Technology* **56**, 715–728 (2005).

[6] Graesser, A. C. & Person, N. K. Question asking during tutoring. *American Educational Research Journal* **31**, 104–137 (1994).

[7] Jenkins, W. I. *Policy Analysis: A Political and Organisational Perspective.* (M. Robertson, 1978).

[8] Graesser, A. C., McMahen, C. L. & Johnson, B. K. Question asking and answering. In Gernsbacher, M. A. *Handbook of Psycholinguistics* 517–538 (Academic Press, 1994).

[9] Miyake, N. & Norman, D. A. To ask a question, one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior* **18**, 357–364 (1979).

[10] Graesser, A. C., Person, N. & Huber, J. Mechanisms that generate questions. *Questions and Information Systems* **2**, 167–187 (1992).

[11] Lehnert, W. G. *The Process of Question Answering: A Computer Simulation of Cognition*. (Laurence Erlbaum Associates, 1978).

[12] Graesser, A. C., Lang, K. & Horgan, D. A taxonomy for question generation. *Questioning Exchange* **2**, 3–15 (1988).

[13] Graesser, A.C., & Black, J.B. *The Psychology of Questions*. (Erlbaum, 1985).

[14] Liddy, E. D. Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science and Technology* **24**, 14–16 (1998)

[15] de Villiers, J., & Roeper, T. *Handbook of Generative Approaches to Language Acquisition*. vol. 41 (Springer, 2011).

[16] Graesser, A.C., Ozuru, Y., & Sullins, J. What is a good question? In McKeown, M.G. & Kucan, L. *Bringing Reading Research to Life.* (Guilford Press, 2010).

[17] Veerman, A., Andriessen, J. & Kanselaar, G. Collaborative argumentation in academic education. *Instructional Science* **30**, 155–186 (2002).

[18] Silva, M. & Cain, K. The use of questions to scaffold narrative coherence and cohesion. *Journal of Research in Reading* **42**, 1–17 (2019).

[19] Mason, J., Chen, W. & Hoel, T. Questions as data: illuminating the potential of learning analytics through questioning an emergent field. *Research and Practice in Technology Enhanced Learning* **11**, 12 (2016).

[20] Taboada, A., Tonks, S. M., Wigfield, A. & Guthrie, J. T. Effects of motivational and cognitive variables on reading comprehension. *Read Writ* **22**, 85–106 (2009).

[21]Ulyshen, T. Z., Koehler, M. J. & Gao, F. Understanding the connection between epistemic beliefs and internet searching. *Journal of Educational Computing Research* **53**, 345–383 (2015).

[22]Lavender, K., Nicholson, S. & Pomerantz, J. Building bridges for collaborative digital reference between libraries and museums through an examination of reference in special collections. *The Journal of Academic Librarianship* **31**, 106–118 (2005).

[23]White, M. D. & Iivonen, M. Questions as a factor in web search strategy. *Information Processing & Management* **37**, 721–740 (2001).

[24]White, M. D. Questioning behavior on a consumer health electronic list. *The Library Quarterly: Information, Community, Policy* **70**, 302–334 (2000).

[25]Huang, K., Yeomans, M., Brooks, A. W., Minson, J. & Gino, F. It doesn't hurt to ask: question-asking increases liking. *J Pers Soc Psychol* **113**, 430–452 (2017).

[26]White, M. D. Questions in reference interviews. *Journal of Documentation* **54**, 443–465 (1998).

[27]Pedrosa-de-Jesus, H., Moreira, A., Lopes, B. & Watts, M. So much more than just a list: exploring the nature of critical questioning in undergraduate sciences. *Research in Science & Technological Education* **32**, 115–134 (2014).

[28]Portnoy, L. B. & Rabinowitz, M. What's in a domain: understanding how students approach questioning in history and science. *Educational Research and Evaluation* **20**, 122–145 (2014).

[29]Almeida, P. A. Can I ask a question? The importance of classroom questioning. *Procedia - Social and Behavioral Sciences* **31**, 634–638 (2012).

[30]Glaubman, R., Glaubman, H. & Ofir, L. Effects of self-directed learning, story comprehension, and self-questioning in kindergarten. *The Journal of Educational Research* **90**, 361–374 (1997).

[31]Hilton, D. J. Conversational processes and causal explanation. *Psychological Bulletin* **107**, 65–81 (1990).

[32]Yeomans, M., Brooks, A. W., Huang, K., Minson, J. & Gino, F. It helps to ask: The cumulative benefits of asking follow-up questions. *J Pers Soc Psychol* **117**, 1139–1144 (2019).

[33]Kluger, A. N. & Malloy, T. E. Question asking as a dyadic behavior. *Journal of Personality and Social Psychology* **117**, 1127–1138 (2019).

[34]Lagnado, D. A. *Explaining the Evidence: How the Mind Investigates the World*. (Cambridge University Press, 2021).

[35]Osman, M. *Future-Minded: The Psychology of Agency and Control*. (Macmillan International Higher Education, 2014).

[36]Osman, M. Basic data set, *Open Science Framework,* (2022), https://osf.io/c978s/files/.

[37]Hagmayer, Y., Meder, B., Osman, M., Mangold, S. & Lagnado, D. Spontaneous causal learning while controlling a dynamic system. *The Open Psychology Journal* **3**, (2010).

[38]Osman, M. Positive transfer and negative transfer/antilearning of problem-solving skills. *Journal of Experimental Psychology: General* **137**, 97–115 (2008).

[39]Hester, K. S. *et al.* Causal analysis to enhance creative problem-solving: performance and effects on mental models. *Creativity Research Journal* **24**, 115–133 (2012).

[40]Marcy, R. T. & Mumford, M. D. Social innovation: enhancing creative performance through causal analysis. *Creativity Research Journal* **19**, 123–140 (2007).

[41]Lagnado, D. A. & Gerstenberg, T. Causation in Legal and Moral Reasoning. In Waldmann, M.R. *The Oxford Handbook of Causal Reasoning* 565–602 (Oxford University Press, 2017).

[42]Osman, M. & Wiegmann, A. Explaining Moral Behavior. *Experimental Psychology* **64**, 68–81 (2017).

[43]Bechlivanidis, C. & Lagnado, D. A. Time reordered: causal perception guides the interpretation of temporal order. *Cognition* **146**, 58–66 (2016).

[44]Bechlivanidis, C. & Lagnado, D. A. Does the 'why' tell us the 'when'? *Psychol Sci* **24**, 1563–1572 (2013).

[45]Tešić, M., Liefgreen, A. & Lagnado, D. The propensity interpretation of probability and diagnostic split in explaining away. *Cognitive Psychology* **121**, 101293 (2020).

[46]Tubau, E. Enhancing probabilistic reasoning: the role of causal graphs, statistical format and numerical skills. *Learning and Individual Differences* **18**, 187–196 (2008).

[47]Lagnado, D. A., Fenton, N. & Neil, M. Legal idioms: a framework for evidential reasoning. *Argument & Computation* **4**, 46–63 (2013).

[48]Liefgreen, A., Pilditch, T. & Lagnado, D. Strategies for selecting and evaluating information. *Cognitive Psychology* **123**, 101332 (2020).

[49]Osman, M. *et al.* Learning from behavioural changes that fail. *Trends in Cognitive Sciences* **24**, 969–980 (2020).

[50]Joyce, K. E. & Cartwright, N. Bridging the gap between research and practice: predicting what will work locally. *American Educational Research Journal* **57**, 1045–1082 (2020).

[51]Game, E. T. *et al.* Cross-discipline evidence principles for sustainability policy. *Nat Sustain* **1**, 452–454 (2018).

[52]Cartwright, N. & Hardie, J. Predicting what will happen when you intervene. *Clin Soc Work J* **45**, 270–279 (2017).

[53]Cartwright, N. & Hardie, J. *Evidence-Based Policy: A Practical Guide to Doing It Better*. (Oxford University Press, 2012).

[54]Gluckman, P. Policy: the art of science advice to government. *Nature* **507**, 163–165 (2014).

[55]Weingart, P. Scientific expertise and political accountability: paradoxes of science in politics. *Science and Public Policy* **26**, 151–161 (1999).

Legend

1. Table 1. Taxonomies of questions

This table sets out the GPH Taxonomy on the left, and the Taxonomy of Policy Questions on the right. The Taxonomy of Policy Questions was generated by applying the GPH Taxonomy to the dataset, and moderating it—as explained in the methods section—in order to better capture the data.

2. Table 2. Summary: frequencies (%) and mean word length (SD) of questions according to superordinate and subordinate categories of the Taxonomy of Policy Questions

This table sets out all of the question styles within the Taxonomy of Policy Questions. For each question-style, it shows: its mean word length (SD), the percentage of questions of this style which came from public policy professionals, the percentage of questions of this style which came from private policy professionals, and the percentage of questions of this style within the total sample.

Author contributions statement

Magda Osman is responsible for the conception of this project and design of this study. Regarding the analysis, she did half of the coding (of policy questions). Furthermore, she is responsible for all of the data analysis which was done on the coded dataset. She also led on the interpretation of the results, and produced the first draft of this manuscript.

Nick Cosstick did half of the coding (of policy questions)—which was part of the analysis of the database. He also made substantial revisions to the initial draft of this paper, focusing on the introduction and the discussion—the latter of which contains the interpretation of the results.

Both authors have approved the submitted version of this manuscript. Furthermore, both authors have agreed to be personally accountable for their own contributions, and to ensure that questions related to the accuracy or integrity of any part of the work—even ones in which they were not personally involved—are appropriately investigated, resolved, and the resolution documented in the literature.

Competing interests statement
Both authors declare that they are employed by the Centre for Science and Policy (CSaP), at the University of Cambridge.

CSaP's policy fellows team collected the raw data used in this study. Furthermore, CSaP's director, Rob Doubleday, read a draft of this manuscript, but did not edit it in any way. This exhausts CSaP's involvement—as an institution—in the generation of this work.

This research was funded by a grant from the Capabilities in Academic Policy Engagement (CAPE) Collaborations Fund, to the project Tools to Structure Policy Makers' Questions Seeking Academic Expertise—a collaboration between the University of Cambridge and the Government Office for Science.

Wordcount
Introduction + Results + Discussion = 3896 words

Introduction + Results + Discussion + Data availability + Author contributions statement + Competing interests statement = 4220 words