PERSPECTIVE

# Use (and abuse) of expert elicitation in support of decision making for public policy

**M. Granger Morgan**[1]

*Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213*

Edited by William C. Clark, Harvard University, Cambridge, MA, and approved March 18, 2014 (received for review October 22, 2013)

**The elicitation of scientific and technical judgments from experts, in the form of subjective probability distributions, can be a valuable addition to other forms of evidence in support of public policy decision making. This paper explores when it is sensible to perform such elicitation and how that can best be done. A number of key issues are discussed, including topics on which there are, and are not, experts who have knowledge that provides a basis for making informed predictive judgments; the inadequacy of only using qualitative uncertainty language; the role of cognitive heuristics and of overconfidence; the choice of experts; the development, refinement, and iterative testing of elicitation protocols that are designed to help experts to consider systematically all relevant knowledge when they make their judgments; the treatment of uncertainty about model functional form; diversity of expert opinion; and when it does or does not make sense to combine judgments from different experts. Although it may be tempting to view expert elicitation as a low-cost, low-effort alternative to conducting serious research and analysis, it is neither. Rather, expert elicitation should build on and use the best available research and analysis and be undertaken only when, given those, the state of knowledge will remain insufficient to support timely informed assessment and decision making.**

Society often calls on experts for advice that requires judgments that go beyond well-established knowledge. In providing such judgments, it is common practice to use simulation models, engineering–economic assessment, and similar tools. Although such analytical strategies can provide valuable insight, they can never hope to include all relevant factors. In such situations, the community of applied decision analysis has long used quantitative expert judgments in the form of subjective probability distributions that have been elicited from relevant experts. Most such applications have been undertaken in support of decisions being made by private parties (1–4). Sometimes the resulting distributions are used directly, and sometimes they are fitted to formal functions and used in various Bayesian decision models (2, 5).

The use of expert elicitation in public sector decision making has been less common. Several studies have explored issues such as the health impacts of fine particle air pollution (6–12) and of lead pollution (13), the likely nature and extent of climate change (14–16), the various impacts that may result from climate change (17, 18), herbicide-tolerant oilseed crops (19), and the likely cost and performance of various energy technologies (20–24). The Environmental Protection Agency (EPA) has begun to make use of elicitation methods to address uncertain issues in environmental science (25), and those who work in both the Department of Energy and the Food and Drug Administration (FDA) have expressed interest in possibly using the method.

Done well, expert elicitation can make a valuable contribution to informed decision making. Done poorly it can lead to useless or even misleading results that lead decision makers astray, alienate experts, and wrongly discredit the entire approach. In what follows, I draw on relevant literature and 35 y of personal experience in designing and conducting substantively detailed expert elicitations, to suggest when it does and does not make sense to perform elicitations, how they should be designed and conducted, and how I believe the results should and should not be used. In contrast to much of the literature in Bayesian decision-making and applied decision analysis, my focus is on developing detailed descriptions of the state of understanding in some field of science or technology.

## First, Are There Any Experts?

To conduct an expert elicitation, there must be experts whose knowledge can support informed judgment and prediction about the issues of interest. There are many topics about which people have extensive knowledge that provides little or no basis for making informed predictive judgments. For example, the further one moves away from questions whose answers involve matters of fact that are largely dependent on empirical natural or social science and well-validated models into realms in which individual and social behavior determine the outcomes of interest, the more one should ask whether expertise, with predictive capability, exists. For example, given a specified time series of future radiative forcing and other relevant physical variables, in my view, it is reasonable to ask climate scientists to make probabilistic judgments about average global temperature 150 y in the future. I am far less persuaded that it makes sense to ask "experts" questions that entail an assessment of how the stock market, or the price of natural gas will evolve over the next 25 y, or what the value of gross world product will be 150 y in the future.

## The Interpretation of Probability

A subjectivist or Bayesian interpretation of probability (5, 26–28) is used when one makes subjective probabilistic assessments of the present or future value of uncertain quantities, the state of the world, or the nature of the processes that govern the world. In such situations, probability is viewed as a statement of an individual's belief, informed by all formal and informal evidence that he or she has available. Although subjective, such judgments cannot be arbitrary. They must conform to the laws of probability. Further, when large quantities of evidence are available on identical repeated events, one's subjective probability should converge to the classical frequentist interpretation of probability.

Partly as a result of their different training and professional cultures, different groups of experts display different views about the appropriateness of making subjective probabilistic judgments, and have different levels of willingness to make such judgments. Although every natural scientist and engineer I have ever interviewed seemed to think naturally in terms of subjective probabilities, others, such as some experts in the health sciences, have been far

Author contributions: M.G.M. designed research, performed research, and wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission.

[1]E-mail: granger.morgan@andrew.cmu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1319946111/-/DCSupplemental.

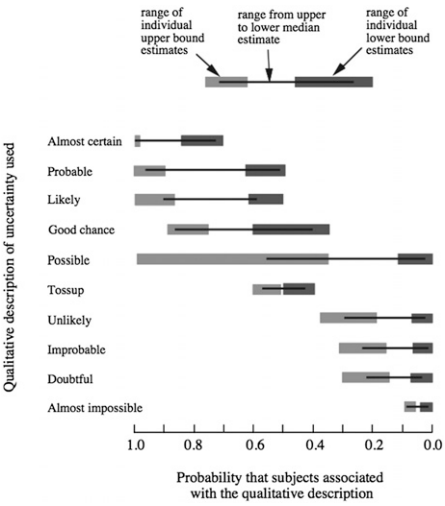


**Fig. 1.** The range of numerical probabilities that respondents attached to qualitative probability words in the absence of any specific context are shown. Note the very wide ranges of probability that were associated with some of these words. Figure redrawn from Wallsten et al. (30).

less comfortable with such formulations. For example, some years ago, my colleagues and I conducted an expert elicitation among a group of different types of health experts in an effort to gain insight about health damages that could result from chronic exposure to submicron sulfate air pollution. One of our experts, an inhalation toxicologist, tried repeatedly to answer our questions to provide a subjective probability distribution on the slope of a health damage function, but simply could not bring himself to provide such answers. After framing our questions in several different ways, and always reaching an impasse, we suspended the elicitation. Some days later the expert came back to us saying he had been thinking about it, that the questions we had been asking made sense, and that he wanted to try again. However, when we did that, he once again found that he could not bring himself to make the necessary quantitative judgments. Although this may be an extreme case, I believe that it also reflects a broader difference among fields.

Fifteen years ago, the Presidential/Congressional Commission on Risk Assessment and Risk Management (29), almost all of whose members were medical professionals, argued that natural scientists should provide probabilistic assessments of exposures, and economists should provide probabilistic assessments of damages, but that health experts should provide only a deterministic treatment of the health damage functions associated with environmental exposures. This reticence to engage in making quantitative subjective judgments has led some to draw an overly sharp distinction between variability and uncertainty—with the claim that only the former should be described in terms of distributions (i.e., with histograms). Although there are certainly situations in which it is important to distinguish variability from uncertainty, there are also many decision contexts in which distinguishing between the two simply adds unnecessary complication.

## Qualitative Uncertainty Words Are Not Sufficient

There is clear evidence that without some quantification, the use of qualitative words such as "likely" and "unlikely" to describe uncertainty can mask important, often critical, differences between the views of different experts. The problem arises because the same words can mean very different things to different people, as well as different things to the same person in different contexts. Fig. 1 summarizes the range of quantitative values that respondents attached to various probability words, independent of any specific context, in a study conducted by Wallsten et al. (30). Wardekker et al. (31) report similar findings in more recent studies undertaken in The Netherlands to improve the communication of uncertainty in results from environmental assessments. Fig. 2 summarizes the range of quantitative values that members of the EPA Science Advisory Board attached to probability words used to describe the likelihood that a chemical agent is a human carcinogen. Such results make a compelling case for at least some quantification when assessing the value of uncertain coefficients or the likelihood of uncertain events. The climate assessment community has taken this lesson seriously, providing mappings of probability words into quantitative values in most assessment reports (34–36).

## Cognitive Heuristics and Bias

We humans are not equipped with a competent mental statistical processor. Rather, in making judgments in the face of uncertainty, we unconsciously use a variety of cognitive heuristics. As a consequence, when asked to make probabilistic judgments, either in a formal elicitation or in any less formal setting, people's judgments are often biased. Two of the cognitive heuristics that are most relevant to expert elicitation are called "availability" and "anchoring and adjustment." These heuristics have been extensively studied by Tversky and Kahneman (37, 38).

Through the operation of availability, people assess the frequency of a class, or the probability of an event, by the ease with which instances or occurrences can be brought to mind. In performing elicitation, the objective should be to obtain an expert's carefully considered judgment based on a systematic consideration of all relevant evidence. For this reason one should take care to adopt strategies designed to help the expert being interviewed to avoid overlooking relevant evidence.

When presented with an estimation task, if people start with a first value (i.e., an anchor) and then adjust up and down from that value, they typically do not adjust sufficiently. Kahneman and Tversky call this second heuristic "anchoring and adjustment" (37, 38). To minimize the influence of this heuristic when eliciting probability distributions, it is standard procedure not to begin with questions that ask about "best" or most probable values but rather to first ask about extremes: "What is the highest (lowest) value you can imagine for coefficient *X*?" or "Please give me a value for coefficient *X* for which you think there is only one chance in 100 that actual value could be larger (smaller)." Having obtained an estimate of an upper (lower) bound, it is then standard practice to ask the expert to imagine that the uncertainty about the coefficient's value has been resolved and the actual value has turned out to be 10% or 15% larger (smaller) than the bound they offered. We then ask the expert, "Can you offer


**Fig. 2.** Results obtained by Morgan (32) when members of the Executive Committee of the EPA Science Advisory Board were asked to assign numerical probabilities to uncertainty words that had been proposed for use with EPA cancer guidelines (33). Note that even in this relatively small and expert group, the minimum probability associated with the word "likely" spans 4 orders of magnitude, the maximum probability associated with the word "not likely" spans more than 5 orders of magnitude, and there is an overlap of the probabilities the different experts associated with the two words.

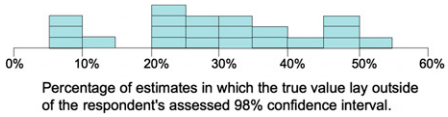


**Fig. 3.** Summary of the value of the surprise index (ideal value = 2%) observed in 21 different studies involving over 10,000 assessment questions. These results indicate clearly the ubiquitous tendency to overconfidence (i.e., assessed probability distributions that are too narrow). A more detailed summary is provided in Morgan and Henrion (39).

an explanation of how that might be possible?" Sometimes experts can offer a perfectly plausible physical explanation, at which point we ask them to revise their bound. After obtaining estimates of upper and lower bounds on the value of a coefficient of interest, we then go on to elicit intermediate values across the probability distribution ["What is the probability that the value of $X$ is greater (less) than Y?"]. If results seem to be unduly scattered, changing the question format may help: "Give me a value of $X$ such that the odds that the true value is greater (less) than 1 in $Z$ (or probability $P$)."

To support such interviews the decision analysis community has developed adjustable probability wheels on which the size of a colored pie-slice portion of a wheel can be adjusted so that respondents can compare their assessments of probability to the size of the slice and adjust it up and down until the size of the slice corresponds to their judged probability (1). Although such aids may be helpful for decision analysts who are dealing with clients with limited numeracy, when we have shown such an aid to an expert in science or technology they have typically toyed with it in a bemused way and then set it aside to give direct quantitative responses.

Only after filling in a number of intervening points in a cumulative distribution function does one finally ask for a median or best estimate, sketch the resulting distribution, and show it to the expert for their assessment and possible revision.

## Ubiquitous Overconfidence

One reason for adopting this rather elaborate procedure is that there is strong evidence that most such judgments are overconfident. A standard measure of overconfidence is the surprise index: the fraction of true values that lie outside an assessor's 98% confidence interval when answering questions for which the true answer is known (e.g., the length of the Panama Canal). Fig. 3 reports a summary of results from 21 different studies involving over 10,000 such assessment questions. Note that none yield the target value for the surprise index of 2% and over half yielded values of 30% or more! Lest the reader infer that such overconfidence is only observed in judgments made by lay respondents, Fig. 4 shows the evolution over time of the recommended values for the speed of light. Similar results exist for other physical quantities.

Calibration is a widely used measure of the performance of someone making subjective probabilistic judgments. Lichtenstein et al. (41) explain that an assessor (judge) is well calibrated "if, over the long run, for all propositions assigned a given probability, the proportion that is true equals the probability that is assigned. Judges' calibration can be empirically evaluated by observing their probability assessments, verifying the associated propositions, and then observing the proportion that is true in each response category." With a few exceptions, such as weather forecasters who make daily precipitation forecasts aided by computer models and receive regular feedback on how well they are performing (42, 43), most people making subjective judgments are not very well calibrated. Fig. 5 shows examples of very poorly calibrated results from clinical diagnosis of pneumonia (44) to very well-calibrated probabilistic precipitation judgments by US weather forecasters (43).

Lichtenstein et al. (41) found that probability judgments tend to be too high when questions are hard, and too low when questions are easy, where "hard" and "easy" questions were classified in terms of the percentage of correct answers made by a reference group. One possible explanation is that assessors partition their responses according to some fixed cut-off value. The hard/easy effect would result if that value remains constant as the difficulty of the question changes. Lichtenstein et al. (41) suggest that the hard–easy effect may result because of ". . .an inability to change the cutoffs involved in the transformation from feelings of certainty to probabilistic responses."

If an assessor is asked a large enough set of questions to make it possible to plot a calibration curve, one might be tempted to simply adjust his or her assessed probabilities (e.g., when the expert says $P = 0.7$, adjust it to 0.8). Kadane and Fischhoff (45) have shown that for assessed probabilities that conform to the basic laws of probability (i.e., are coherent) such a procedure is not justified.

## Developing a Protocol

A primary output of many expert elicitations is a set of subjective probability distributions on the value of quantities of interest, such as an oxidation rate or the slope of a health damage function (for an example of a simple elicitation interview, see *SI Appendix*).

However, often the objective is broader than that—to obtain an expert's characterization of the state of knowledge about a general topic or problem area in which the elicitation of specific probability distributions may be only one of a number of tasks. Either way the development of a good elicitation protocol requires considerable time and care, and multiple iterations on format and question wording. Working with colleagues who are familiar with the domain and its literature one can usually build a much longer list of questions than it is reasonable to have an expert answer in a session of a few hours or the better part of a day. If the objectives of the elicitation have not already been sharply defined, this is the time to do that. A sharp focus

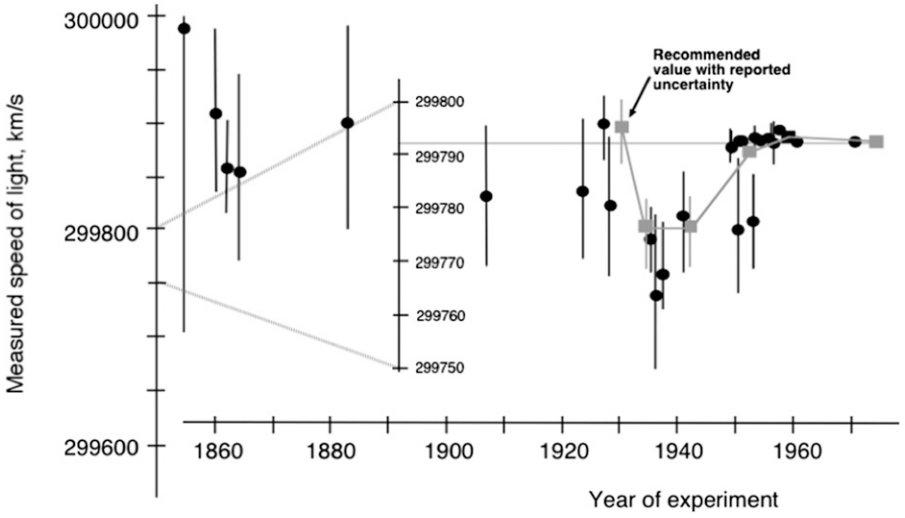


**Fig. 4.** Published estimates of the speed of light. The light gray boxes that start in 1930 are the recommended values from the particle physics group that presumably include an effort to consider uncertainty arising from systematic error (40). Note that for over two decades the reported confidence intervals on these recommended values did not include the present best-measured value. Henrion and Fischhoff (40), from which this figure is combined and redrawn, report that the same overconfidence is observed in the recommended values of a number of other physical constants.

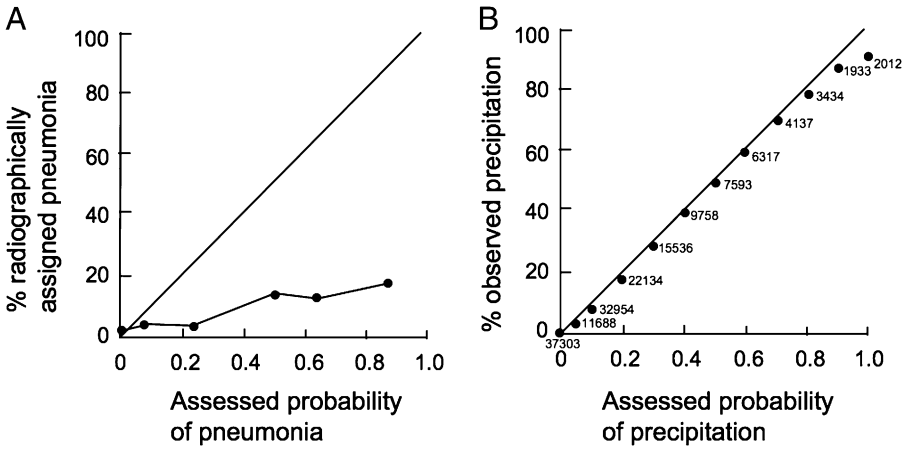


**Fig. 5.** Illustration of two extremes in expert calibration. (*A*) Assessment of probability of pneumonia (based on observed symptoms) in 1,531 first-time patients by nine physicians compared with radiographically assigned cases of pneumonia as reported by Christensen-Szalanski and Bushyhead (44). (*B*) Once-daily US Weather Service precipitation forecasts for 87 stations are compared with actual occurrence of precipitation (April 1977 to March 1979) as reported by Charba and Klein (43). The small numbers adjacent to each point report the number of forecasts.

can help the pruning process and sometimes the pruning can help to sharpen the focus.

Questions that are posed in an expert elicitation should pass what is commonly termed a clairvoyant test. The question, "What will be the price of gasoline next year?" fails such a test. Without specifying the octane, and when and where that gasoline is to be purchased, a clairvoyant cannot provide a precise answer to this question.

The best experts have comprehensive mental models of all of the various factors that may influence the value of an uncertain quantity, as well as which of those factors most contribute to its uncertainty. However, not all of that knowledge may be comparably accessible. Because the objective of an elicitation should be to obtain each expert's best-considered judgment, it is important to help them keep all of those factors in mind as they answer specific questions in an elicitation. To assist in that process, we have often used a variety of graphical aids such as summary tables and influence diagrams to illustrate the relation between key factors that influence the value of interest. For a simple example see pages 4 and 5 of the protocol used in Curtright et al. (20) (available at http://pubs.acs.org/doi/suppl/10.1021/es8014088/suppl_file/es8014088_si_001.pdf).

My colleagues and I have also made frequent use of card-sorting tasks, in which, working iteratively with the group of experts before we visit them, we develop a set of cards, each of which lists a factor that may influence the value of interest (blank cards are included so that an expert can add, modify, or combine factors). After discussing and possibly refining or modifying the factors, the expert is then asked to sort the cards, first in terms of the strength of influence, and then a second time in terms of how much each factor contributes to uncertainty in the value of the quantity of interest. Such an exercise helps experts to differentiate between the strength of influences versus sources of uncertainty, and to focus on the most important of the latter in formulating their probabilistic responses. For an example, see pages 5–7 of the protocol used in Zickfeld et al. (16) (available at www.pnas.org/content/suppl/2010/06/28/0908906107.DCSupplemental/Appendix.pdf).

Similarly, when we have done an elicitation on a future technology, such as carbon capture and geological sequestration for coal-fire power plants, we have taken it apart into component pieces, rather than simply asking for holistic judgments about the entire system (46).

In choosing the questions that will be posed, it is important to draw a clear distinction between questions of fact and questions whose answers largely entail normative judgments. It may be appropriate in some circumstances to ask experts what they believe a specific group's preferences are or will be. However, one should take care to distinguish such questions from those in which, using methods similar to those used in expert elicitation, experts' own value judgments are elicited. An example of the former would be questions of fact, such as the implicit "value of a statistical life" that a specific socioeconomic group can be expected to display in making a well-specified risky decision. An example of the latter would be normative questions about what value of a statistical life society should adopt in making regulatory decisions. Although it may be interesting to learn what value of a statistical life an economist thinks society should adopt, or what level of protection an ecologist thinks society should afford a particular species or habitat, such questions are not about issues of fact, and thus are more appropriately handled as part of an opinion survey.

In most of the elicitations I have conducted, I have involved an excellent postdoctorate or junior colleague, who has not yet established a reputation or a professional stake in the field, but has performed a recent systematic review of the relevant literature. Upon hearing a particular response from an expert, they may observe, "That response would appear to be at odds with work reported by group X." Sometimes the expert will respond "Oh yes, I had forgotten about that" and adjust his or her answer. More often he or she says something more along the lines of, "Yes, I know, but I really discount the work of group X because I have grave doubts about how they calibrate their instrument." When I have described this proactive procedure to some colleagues who work in survey research they have expressed concern that such intervention may inappropriately influence an expert's response. Although I am not aware of literature on this point, in most of the elicitations that I have conducted in areas of natural science, such as air chemistry or climate change, it is my experience that the experts are intimately familiar with and have assessed each others' work, and it is most unlikely that anything I or my colleagues say during an elicitation session will change their judgment once they have considered all relevant evidence. When that is not the case, care should be taken ahead of time to provide literature packets, reviews, and summaries so that all experts come to the questions with a comparable familiarity with available knowledge.

In contrast to political or similar polling, the objective of most expert elicitation is not to obtain a statistically representative sample of the views of a population. Rather, it is to gain an understanding of the range of responsible expert judgments and interpretations across the field of interest. Thus, in selecting the group of experts, care must be taken to include people who represent all of the major perspectives and interpretations that exist within the community. This can typically be achieved by a careful reading of the literature and discussion with experts who can identify the views of their various peers. In the elicitations we have conducted, we have often constructed tables of the experts sorted by background and technical perspective. Because we have always worked with a collaborator who was expert in the field and with the relevant literatures, we have not felt it necessary to use more formal procedures for sorting and selecting participants.

When results from an expert elicitation are to be used as input to regulatory or other public policy decision making (by EPA, FDA, etc.), perceived legitimacy or fairness become especially important (47). In such cases, a more

systematic approach should be used in the selection of experts. Knol et al. (48) outline a number of more formal procedures that they and others have used to select experts. In their expert elicitation of the health impacts from submicron particles (PM2.5) conducted for EPA, Roman et al. (9) used a two-part selection process that used publication counts and peer nomination of experts. The EPA White Paper (25) on expert elicitation provides a discussion of these issues.

There is no right answer to the question, "How many experts are needed for a good elicitation." The answer depends on the nature of the field. If virtually all experts adopt similar basic models of the underlying science, then as few as five or six might suffice. In most cases, because experts will have a diversity of opinions about the underlying science, a larger group will be necessary to obtain adequate coverage of the range of opinions.

When we have published the results from expert elicitations, in most cases, we have identified the experts involved, but have not linked individual experts to specific results (although in a few cases experts familiar with the views of their colleagues have been able to privately identify who said what). In many cases, providing such limited anonymity is important so that experts can provide their considered judgment unconstrained by corporate, political, or other considerations. The EPA White Paper on expert elicitation (25) observes that given "...current norms within the scientific community, experts may be unwilling to participate and share their judgments honestly if they fear a need to defend any judgments that divert from the mainstream or conflict with positions taken by their institutions." Although I agree, I find troubling the extension of this argument made by Aspinall (49) who suggests that an advantage of combining results elicited from several experts "...is that it encourages experts wary of getting involved in policy advice: the structured, neutral procedure, and the collective nature of the result reassures experts and relieves them of the burden of sole responsibility." Experts should be providing their careful considered judgments, and too much anonymity may result in their taking those judgments less seriously.

Writing in the specific context of elicitations done in support of environmental impact assessment, Knol et al. (48) describe a seven-step approach to developing and conducting expert elicitations. Despite the title (which sounds like the authors might be offering a cook book) their treatment is thoughtful and nuanced. It explores many of the issues discussed in the preceding paragraphs, reaching broadly similar conclusions.

Although I have argued that the development of an elicitation protocol should be an iterative process, requiring considerable effort, pretesting and refining, not everyone agrees. For example, Aspinall (49) argues that "the speed with which ... elicitations can be conducted is one of their advantages," and cites a study of the virulence of biological agents conducted in just 2 d "with a few days of preparatory work." I have no doubt that in this case, in a study of a very focused topic with an intensive couple of days of preparation, it was possible to develop a quality study. However, one needs to be careful not to encourage the development of "quick and dirty" expert elicitations.

## Computer Tools to Support or Perform Elicitation

A variety of computer tools have been developed and used in support of expert elicitation (50). Some of these are quite specific to the process of elicitation (51, 52); others, such as tools for constructing influence diagrams and Bayesian belief nets, are much more general in nature. For example, in our own work, we have had experts who chose to perform runs of their own computer models to gain insights before answering specific questions we have posed. Using specialized software tools to summarize literature or construct influence diagram or similar aides can also be very helpful.

Although I have found no published literature that evaluates them, several investigators have developed software to perform the actual elicitation, posing questions to establish ranges and seek probabilistic judgments that allow the construction of probability distributions. In at least one case, the software also supports a card-sorting exercise before performing the elicitation. Such tools might be useful if used in conjunction with a face-to-face elicitation. It is an open question whether experts working on their own will devote the same degree of serious consideration in responding to an automated elicitation system that they clearly do when responding to a well-developed protocol during a face-to-face interview with attentive and technically knowledgeable interviewers sitting with them in their office.

## Uncertainty About Model Functional Form

A few investigators have conducted studies in which the assumptions about the functional form of a set of underlying causal processes are explicitly identified and experts are asked to make judgments about the likelihood that each is a correct description of underlying physical reality. Evans et al. (53, 54) developed and demonstrated such methods in the context of health experts' judgments about low-dose cancer risk from exposure to formaldehyde in environmental and occupational settings. The method used the construction of probability trees that allowed experts to make judgments about the relative likelihood that alternative models of possible pharmacokinetic and pharmacodynamic processes correctly describe the biological process that are involved. Budnitz et al. (55–57) have used a set of deliberative processes designed to support a group of experts in developing a "composite probability distribution [that] represents the overall scientific community." The process they developed is very labor intensive and uses experts as evaluators of alternative causal models and their implications rather than as proponents of one or another model. It would be highly desirable to apply procedures such as those developed and demonstrated by Evans et al. (53, 54) and Budnitz et al. (55, 56) in assessment processes such as that used by the Intergovernmental Panel on Climate Change (IPCC). However, resource constraints and the limited familiarity that most experts have with decision science, probably makes such an effort infeasible.

In contrast to integrated assessment models of climate change that adopt fixed model structures and fixed functional relationships among variables, Dowlatabadi and I (58, 59) populated our integrated climate assessment model (ICAM) with switches which allow the user to explore the implications of a wide range of plausible alternative functional forms. In addition to alternative assumptions about climate science and impacts, ICAM also allows users to explore models that use a variety of different approaches to time preference, and allows a variety of different behavioral responses (e.g., nations may or may not defect from a global carbon tax regime as tax rates become high). In exploring a wide range of alternative model functional forms, it became clear that we could get an enormous variety of answers depending on the range of plausible assumptions we made about the structure of the model and which regional decision maker we considered. Rarely was any emission abatement policy optimal for all regions. Rarely were any results stochastically dominant. We concluded that it is indefensible to use integrated assessment models that have a fixed functional form in an effort to find a single globally optimal climate policy.

Finally, Refsgaard et al. (60) have suggested a variety of different strategies that can be used to explore the implications of what they term "uncertainty due to model structure error."

## Confidence, Second Order Uncertainty, and Pedigree

The nature and quality of the evidence that experts draw on to make probabilistic judgments is often highly variable. In developing guidance on the treatment of uncertainty for IPCC, Moss and Schneider (34) distinguished between the amount of evidence available to

support a judgment and the degree of consensus within the scientific community. When both are high they term the state of knowledge as "well established." When evidence is modest but agreement is high they term the state "established but incomplete;" when the reverse is true they say "there are competing explanations." When both evidence and agreement are low they describe the situation as "speculative."

The IPCC has continued to use a 2D formulation. However, for the fifth assessment (61) the interpretation evolved to (*i*) confidence in the validity of a finding, based on the type, amount, quality, and consistency of evidence (e.g., mechanistic understanding, theory, data, models, expert judgment) and the degree of agreement; and (*ii*) quantified measures of uncertainty in a finding expressed probabilistically (based on statistical analysis of observations or model results, or expert judgment).

Rather than quantify confidence, the guidance document explains that the level of confidence in a probabilistic assessment should be expressed using [one of] five qualifiers: very low, low, medium, high, and very high. The guidance explains that "levels of confidence are intended to synthesize author teams' judgments about the validity of findings as determined through their evaluation of evidence and agreement, and to communicate their relative level of confidence qualitatively." In addition, a mapping is provided between probability words and probability values. Hence, in the IPCC's fifth assessment report (62) one reads statements such as, "In the Northern Hemisphere, 1983–2012 was likely the warmest 30-y period of the last 1,400 y (medium confidence)."

In that statement, the IPCC maps the word "likely" to a probability range of 66–100%. Statements such as this are basically an alternative to reporting a second-order uncertainty, that is, to reporting an assessment of the probability that one's single value assessed probability is correct. For a graphical display, see figure 9.2 in Climate Change Science Program 5.2 (36).

Funtowicz and Ravetz (63) further refined these ideas by introducing a five-element vector to describe uncertain quantities. The elements in their Numeral Unit Spread Assessment Pedigree (NUSAP) characterization of an uncertain quantity are as follows: numeral (typically a best estimate); unit (the units in van der Sluijs et al. in refs. 64 and 65 in which the value is measured); spread and assessment (which are simple and more complete descriptions of uncertainty about the value of the quantity); and pedigree which is intended to "convey an evaluative account of the production process of the quantitative information," typically in the form of a matrix of qualitative values.

Assigning a pedigree to each uncertain quantity is an appealing idea, but implementing it in practice becomes rather complicated. In refs. 64 and 65 several attempts to implement the NUSAP idea in environmental assessments have been made. Assessing and propagating a pedigree matrix of qualitative values through a quantitative model obviously requires one to focus on those variables that have greatest influence on the output of interest. The results become rather complex and, in my view, their utility to decision makers remains an open question.

## Diversity in Expert Opinion

It is common in assessment processes such as those conducted by the IPCC, to convene panels of experts and ask them to produce consensus judgments of the value of key uncertain quantities. In most cases, this is done informally. Whereas the cognitive biases described above certainly operate in such circumstances, there is typically no way to assess their impact or control their influence in such informal settings. In several of the elicitations of individual experts that my colleagues and I have conducted on issues related to climate change, we have obtained significantly wider ranges of values than those reported by the analogous IPCC consensus process.

Fig. 6 compares results from an elicitation of the values of radiative forcing by aerosols with the IPCC fourth assessment (66) with distributions elicited at about the same time from 24 aerosol experts (15). Note that several experts place significant probability outside of the bounds that result if one simply adds the upper bound of the direct and cloud albedo estimates from the IPCC fourth assessment.

Fig. 7 compares results from an elicitation of the values of climate sensitivity (16) with the IPCC fourth assessment (66) that estimated that the "equilibrium climate sensitivity is likely to lie in the range 2–4.5 °C, with a most likely value of about 3 °C." IPCC defined likely as a 0.66–0.90 probability, which in chapter 19 of Working Group II (68) was interpreted as a 0.05–0.17 probability that climate sensitivity is >4.5 °C. Ten of the 14 elicited distributions reported in Fig. 7 placed more than 0.17 of their probability above 4.5 °C.

Without arguing that these results from individual elicitations are more appropriate or informative than IPCC consensus judgments, the difference does suggest that IPCC and similar groups might be well advised to adopt a strategy that uses both approaches. For example, after the experts involved in an assessment team have individually reviewed all of the available evidence, an elicitation of probability distributions for key parameters of interest might be performed with each individual team member. The results could then be used as inputs to the group deliberations in which the team develops their collective assessment.

Oppenheimer et al. (69), Aspinall (49), and the EPA White Paper (25) all argue that an advantage of expert elicitation is that results can clearly display different schools of thought within an expert community. The EPA (25) writes,

> . . .differences in response may result from different paradigms by which the experts view the

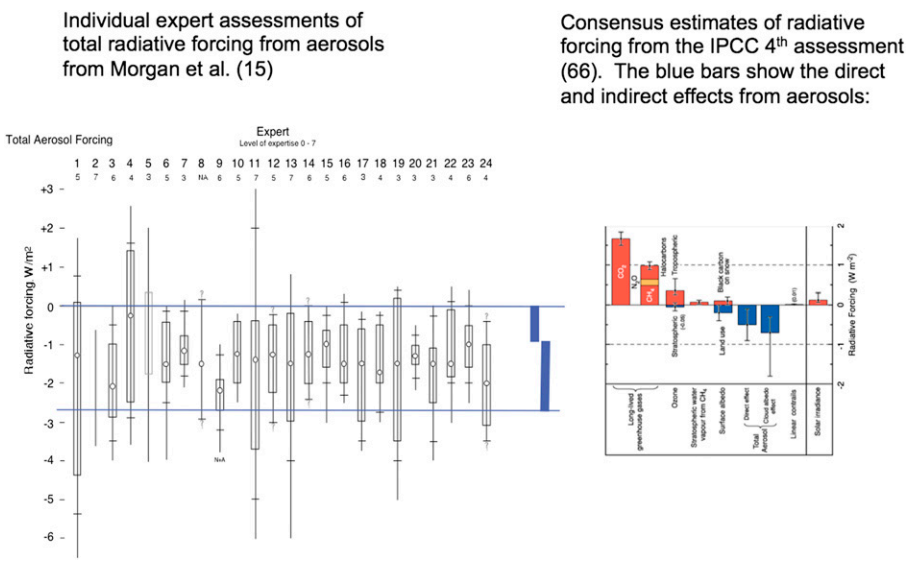


**Fig. 6.** Comparison of individually assessed value of total radiative forcing produced by aerosols (15) (*Left*) with the summary assessment produced by the fourth IPCC assessment (66) (*Right*). Note that many of the individual assessments reported in *Left* involve a wider range of uncertainty than IPCC consensus summary. The summary that was provided in the third assessment (67) included only a portion of the indirect effects and the range was narrower.

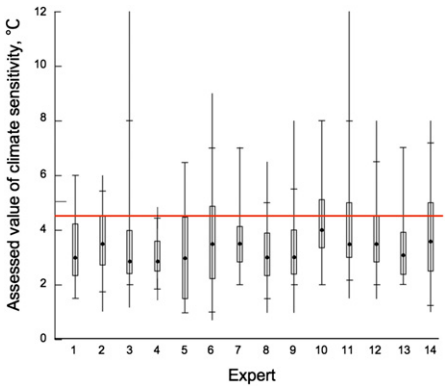


**Fig. 7.** Individual expert assessments of the value of climate sensitivity as reported in Zickfeld et al. (16) compared with the IPCC assessment by Schneider et al. (68) that there is between an 0.05 and 0.17 probability that climate sensitivity is >4.5 °C (i.e., above the red line). The assessed expert distributions place probability of between 0.07 and 0.37 above 4.5 °C.

world and the data. This often is true when the experts come from different disciplinary backgrounds. Experts tend to trust data obtained through methods with which they have direct experience. For example, when one is trying to estimate the relationship between exposure to a substance and increased morbidity or mortality, epidemiologists may tend to find epidemiological data compelling while being more suspect of toxicological studies on animals. Toxicologists may have the opposite preference. In this situation, the variability among the findings represents a spectrum of beliefs and weights that experts from different fields place on the various types of evidence. In such cases, reconciling the differences may be imprudent.

In the context of climate change, Oppenheimer et al. (69) argue that "with the general credibility of the science of climate change established, it is now equally important that policy-makers understand the more extreme possibilities that consensus may exclude or downplay."

Fig. 8 provides a striking example of two quite different schools of thought that existed just under a decade ago within the community of oceanographers on the topic of possible collapse of the Atlantic meridional overturning circulation (AMOC) in the face of global warming. After reviewing literature on paleoclimate change and model simulations, in its 2007 assessment, IPCC Working Group II (68) wrote, "The third line of evidence, not assessed by Working Group I, relies on expert elicitations (sometimes combined with the analysis of simple climate models). These [A]MOC projections show a large spread, with some suggesting a substantial likelihood of triggering a [A]MOC threshold response within this century." However, Fig. 8 was not reproduced in the report.

## Combining Expert Judgments

There is extensive literature on strategies to combine experts' probabilistic judgments, excellent overviews of which can be found in the writings of Clemen and Winkler (70, 71). Clearly, there are circumstances in which combining the judgments of different experts is a sensible thing to do. However, if the experts make very different judgments about the relevant underlying science, or if the uncertain value that is being assessed will be used as an input to a nonlinear model, then it is best not to combine the separate judgments, but rather to run separate analyses to explore how much the difference in expert opinions affect the outcome of interest. For example, in early work on the health impacts of fine-particle air pollution, we found that differences among air pollution experts made relatively little difference in assessments of health impact compared with the wide range of different functional models and views expressed by health experts (6).

Cooke and Goossens (8, 72) have worked extensively on developing and applying methods to assess the quality of expert judgments and support the combining of those judgments. In an approach they and coworkers term the "classical method," experts are asked to make judgments about a number of "seed" questions—questions about quantities in the same general domain as the topic of interest, but for which true values can be found. By performing a product of a calibration score and an information score (a measure of assessed confidence interval), and dropping those experts whose calibration score is lower than a cutoff value, the performance of experts is evaluated, and only those who achieve a high enough score are used to produce a combined distribution. It remains an open question just how diagnostic this procedure is for assessing the quality of expert judgments on complex scientific questions for which answers cannot be known for at least many years in the future. Withholding various numbers of seed questions and treating them as the target quantities of interest has allowed some evaluation of the screening method. On the basis of an examination of 14 studies that used the classical method, Clemen (73) concludes, "the overall out-of-sample performance of Cooke's method appears to be no better than EQ [the use of equal weights on all experts]; the two methods have similar median combination scores, but EQ has less variability and better accuracy." Similarly, Lin and Cheng (74) conclude that although sometimes the performance weight method is superior it does not always outperform EQ. To clarify these issues, Cooke now has plans to extend such assessment to a much larger set of data on seed questions.

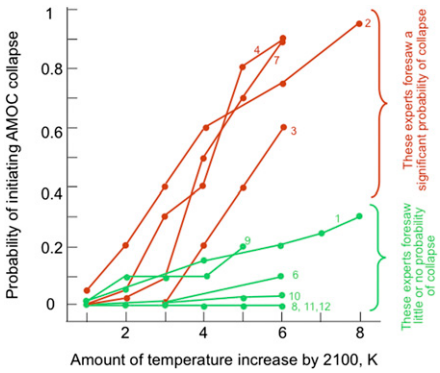


**Fig. 8.** Expert elicitation can be effective in displaying the range of opinions that exist within a scientific community. This plot displays clearly the two very different schools of thought that existed roughly a decade ago within the community of oceanographers about the probability "that a collapse of the AMOC will occur or will be irreversibly triggered as a function of the global mean temperature increase realized in the year 2100." Each curve shows the subjective judgments of one of 12 experts. Four experts (2, 3, 4, and 7 in red) foresaw a high probability of collapse, while seven experts (in red) foresaw little, if any, likelihood of collapse. Collapse was defined as a reduction in AMOC strength by more than 90% relative to present day. Figure redrawn from Zickfeld et al. (18).

While Cooke's method has been used in a number of applications (8, 49), it is potentially problematic in situations, such as the assessment of health damage functions or various quantities in climate science in which different experts make very different assumptions about the nature of the underlying causal mechanisms. As noted above, depending on how the results will be used, combining the judgments of experts (by any procedure) may not be appropriate. It would also be problematic if one were to exclude some experts who represent plausible but poorly represented alternative views about the science. The history of science is replete with examples in which the minority opinion about uncertain science ultimately proved to be correct.

A special case in the literature on combining expert judgments involves the combination of judgments about binary events (either the event happens or it does not). In laboratory studies, Karvetski et al. (75) showed that by eliciting extra judgments to determine how coherent a judgment is, adjusting the resulting set of judgments to make them more coherent, and then weighting those adjusted judgment on the basis of original coherence, a significant improvement in performance could be achieved.

In the early 1950s, a group of investigators at the RAND Corporation developed a strategy to obtain group judgment that they termed the "Delphi method" (76). This method was first used in classified studies conducted for the US Air Force on bombing requirements. When that work was declassified a decade later (76), the method became popular as a strategy for

developing group consensus about the value of unknown parameters or various normative or policy issues. Criticisms of the method soon began to appear. With support from the US Air Force, Sackman (77), another RAND analyst, performed an assessment of the technique, the conclusions of which were highly critical: "Delphi consensus is specious consensus." He recommended that the use of "conventional Delphi be dropped . . . until its principles, methods and fundamental applications can be experimentally established as scientifically tenable." Fifteen years later, after an extensive review conducted for the Dutch government of a much larger body of literature, Woudenberg (78) reached a very similar conclusion, writing, "A Delphi is extremely efficient in obtaining consensus, but this consensus is not based on genuine agreement; rather, it is the result of . . . strong group pressure to conformity."

## Concluding Thoughts and Advice

Some may find it tempting to view expert elicitation as a low-cost, low-effort alternative to doing serious research and analysis. It is neither. Rather, expert elicitation should build on the best available research and analysis and be undertaken only when, given those, the state of knowledge will remain insufficient to support timely informed assessment and decision making.

If expert elicitation is to obtain careful considered judgments from the experts involved, elicitation protocols must be developed through careful iterative refinement. Draft protocols should be pilot tested with quasi experts (such as advanced graduate students or postdoctoral fellows) to assure that question formulations are workable and can be understood. Such iterative refinement is essential because there are always many more things one would like to ask than time and experts' patience will allow. This process of iterative refinement can often take several months or longer. In most cases, true experts are a rare resource that must be conserved and treated with care. A few shoddy studies can sour an entire expert community to participation.

Most of the elicitations my colleagues and I have performed have been conducted using face-to-face interviews in experts' offices where the expert can readily access relevant data and analytic and model results. In many cases, we have prepared concise literature summaries or other materials and have used card sorting and other tasks to encourage experts to systematically identify all relevant factors that may influence a value of interest or contribute to its uncertainty. Although well-informed experts obviously know and have thought about all of these things, it is important to make sure that they do not overlook any of them when they are asked to make quantitative judgments. Indeed, when an answer seems to be at odds with such evidence, it is important to push for explanations and justifications. If it becomes clear that respondents have not thought about some of the relevant evidence, then care should be taken to identify the bounds of their expertise and appropriately limit the use of, and generalizations drawn from, their judgments.

Because experts are human, there is simply no way to eliminate cognitive bias and overconfidence. The best one can hope to do is to work diligently to minimize its influence. It is important to acknowledge this, brief experts on the issue, and design elicitation procedures that work to achieve this objective. Of course, the same cognitive biases arise in the deliberations of less formal consensus panels, but in those cases they are virtually never acknowledged or addressed. The performance of consensus expert panels might be improved if panel members first performed individual elicitations before they begin their group deliberations.

It is tempting to want to combine the judgments of multiple experts to obtain "the" answer. Sometimes this makes sense. However, if different experts base their judgments on very different models of the way in which the world works, or if they produce quite different judgments that will be used as the input to a nonlinear model, then combining judgments does not make sense. It is always important to remember that science is not a matter of majority vote. Sometimes it is the minority outlier who ultimately turns out to have been correct. Ignoring that fact can lead to results that do not serve the needs of decision makers.

**ACKNOWLEDGMENTS.** In my work on expert elicitation, I have benefited from collaboration with, and advice and assistance from, many colleagues including Peter Adams, Ahmed Abdulla, Myles Allen, Deborah Amaral, Inês Azevedo, Aimee Curtright, Hadi Dowlatabadi, Baruch Fischhoff, David Frame, Umit Guvenc, Max Henrion, David Keith, Alan Meier, Samuel Morris, D. Warner North, Stefan Rahmstorf, Anand Rao, William Rish, Edward Rubin, Stephen Schneider, Debra Shenk, Patti Steranchak, Jeroen van der Sluijs, Kirsten Zickfeld, and many others. Much of the work has been supported by grants from the National Science Foundation (NSF) and Electric Power Research Institute. Most recent support has been under NSF Cooperative Agreements SES-0345798 and SES-0949710 and from the John D. and Catherine T. MacArthur Foundation.

**1** Spetzler CS, Staël von Holstein C-AS (1975) Probability encoding in decision analysis. *Manage Sci* 22(3):340–358.
**2** Garthwaite PH, Kadane JB, O'Hagan A (2005) Statistical methods for eliciting probability distributions. *J Am Stat Assoc* 100(470):680–700.
**3** O'Hagan A, et al. (2006) *Uncertain Judgments: Eliciting Experts' Probabilities* (John Wiley & Sons, Hoboken, NJ), 321 pp.
**4** Hora SC (2007) *Advances in Decision Analysis: From Foundations to Applications*, eds Edwards W, Miles RF, Jr, von Winterfeldt D (Cambridge Univ Press, New York), pp 129–153.
**5** DeGroot MH (1970) *Optimal Statistical Decisions* (McGraw-Hill, New York), 489 pp.
**6** Morgan MG, Morris SC, Rish WR, Meier AK (1978) Sulfur control in coal-fired power plants: A probabilistic approach to policy analysis. *J Air Pollut Control Assoc* 28(10):993–997.
**7** Morgan MG, et al. (1984) Technical uncertainty in quantitative policy analysis: A sulfur air pollution example. *Risk Anal* 4(3):201–216.
**8** Cooke RM (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford Univ Press, New York), 336 pp.
**9** Roman HA, et al. (2008) Expert judgment assessment of the mortality impact of changes in ambient fine particulate matter in the U.S. *Environ Sci Technol* 42(7):2268–2274.
**10** Knol AB, et al. (2009) Expert elicitation on ultrafine particles: Likelihood of health effects and causal pathways. *Part Fibre Toxicol* 6:19.
**11** Hoek G, et al. (2010) Concentration response functions for ultrafine particles and all-cause mortality and hospital admissions: Results of a European expert panel elicitation. *Environ Sci Technol* 44(1):476–482.
**12** Cooke RM, et al. (2007) A probabilistic characterization of the relationship between fine particulate matter and mortality: Elicitation of European experts. *Environ Sci Technol* 41(18):6598–6605.
**13** Whitfield RG, Wallsten TS (1989) A risk assessment for selected lead-induced health effects: An example of a general methodology. *Risk Anal* 9(2):197–207.
**14** Morgan MG, Keith DW (1995) Subjective judgments by climate experts. *Environ Sci Technol* 29(10):468A–476A.
**15** Morgan MG, Adams P, Keith DW (2006) Elicitation of expert judgments of aerosol forcing. *Clim Change* 75(1-2):195–214.
**16** Zickfeld K, Morgan MG, Frame DJ, Keith DW (2010) Expert judgments about transient climate response to alternative future trajectories of radiative forcing. *Proc Natl Acad Sci USA* 107(28): 12451–12456.
**17** Morgan MG, Pitelka LF, Shevliakova E (2001) Elicitation of expert judgments of climate change impacts on forest ecosystems. *Clim Change* 49(3):279–307.
**18** Zickfeld K, et al. (2007) Expert judgements on the response on the Atlantic meridional overturning circulation to climate change. *Clim Change* 82(3-4):235–265.
**19** Krayer von Krauss MP, Casman EA, Small MJ (2004) Elicitation of expert judgments of uncertainty in the risk assessment of herbicide-tolerant oilseed crops. *Risk Anal* 24(6):1515–1527.
**20** Curtright AE, Morgan MG, Keith DW (2008) Expert assessments of future photovoltaic technologies. *Environ Sci Technol* 42(24): 9031–9038.
**21** Anadon LD, Nemet GF, Verdolini E (2013) The future costs of nuclear power using multiple expert elicitations: Effects of RD&D and elicitation design. *Environ Res Lett* 8(3):034020.
**22** Anadón LD, Bosetti V, Bunn M, Catenacci M, Lee A (2012) Expert judgments about RD&D and the future of nuclear energy. *Environ Sci Technol* 46(21):11497–11504.
**23** Chan G, et al. (2011) Expert elicitation of cost, performance, and RD&D budgets for coal power with CCS. *Energy Procedia* 4:2685–2692.
**24** Abdulla A, Azevedo IL, Morgan MG (2013) Expert assessments of the cost of light water small modular reactors. *Proc Natl Acad Sci USA* 110(24):9686–9691.
**25** US Environmental Protection Agency (2011) Expert elicitation task force white paper. Available at www.epa.gov/stpc/pdfs/ee-white-paper-final.pdf. Accessed April 25, 2014.
**26** de Finetti B (1974) *Theory of Probability: A Critical Introductory Treatment* (Wiley, New York).
**27** Good IJ (1971) 46656 varieties of Bayesians. *Am Stat* 25(5): 62–63.
**28** Jaynes ET (2003) *Probability Theory: The Logic of Science* (Cambridge Univ Pres, New York), 727 pp.
**29** Presidential/Congressional Commission on Risk Assessment and Risk Management (1997) *Risk Assessment and Risk Management in Regulatory Decision Making* (Washington), Vol 2 of Final Report.
**30** Wallsten TS, et al. (1986) Measuring the vague meanings of probability terms. *J Exp Psychol Gen* 155(4):348–365.
**31** Wardekker JA, et al. (2008) Uncertainty communication in environmental assessments: Views from the Dutch science-policy interface. *Environ Sci Policy* 11(7):627–641.
**32** Morgan MG (1998) Uncertainty analysis in risk assessment. *Hum Ecol Risk Assess* 4(1):25–39.
**33** US Environmental Protection Agency (1996) *Proposed Guidelines for Cancer Risk Assessment* (Office of Research and Development, US EPA, Washington), EPA/600P-92/003C.
**34** Moss R, Schneider SH (2000) Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting. *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*, eds Pachauri R, Taniguchi T, Tanaka K (Intergovernmental Panel on Climate Change, Geneva), pp 33–51. Available at www.ipcc.ch/pdf/supporting-material/guidance-papers-3rd-assessment.pdf. Accessed April 25, 2014.

**35** National Assessment Synthesis Team (2001) *Climate Change Impacts on the United States: The Potential Consequences of Climate Variability and Change.* (Cambridge Univ Press, New York), 612 pp.
**36** Morgan MG, et al. (2009) CCSP 5.2 best practice approaches for characterizing, communicating, and incorporating scientific uncertainty in decision making. *A Report by the Climate Change Science Program and the Subcommittee on Global Change Research* (National Oceanic and Atmospheric Administration, Washington), 96 pp.
**37** Tversky A, Kahneman D (1974) Judgments under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.
**38** Kahneman D, Slovic P, Tversky A, eds (1982) *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge Univ Press, New York).
**39** Morgan MG, Henrion M (1990) *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis* (Cambridge Univ Press, New York).
**40** Henrion M, Fischhoff B (1986) Assessing uncertainty in physical constants. *Am J Phys* 54(9):791–798.
**41** Lichtenstein S, Fischhoff B, Phillips L (1982) *Judgment Under Uncertainty: Heuristics and Biases*, eds Kahneman D, Slovic P, Tversky A (Cambridge Univ Press, New York), pp 306–334.
**42** Murphy AH, Winkler RL (1977) Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest* 2:2–9.
**43** Charba JP, Klein WH (1980) Skill in precipitation forecasting in the National Weather Service. *Bull Am Meteorol Soc* 61(12):1546–1555.
**44** Christensen-Szalanski JJJ, Bushyhead JB (1981) Physician's use of probabilistic information in a real clinical setting. *J Exp Psychol Hum Percept Perform* 7(4):928–935.
**45** Kadane J, Fischhoff B (2013) A cautionary note of global recalibration. *Judgm Decis Mak* 8(1):25–28.
**46** Rao AB, Rubin ES, Keith DW, Morgan MG (2006) Evaluation of potential cost reductions from improved amine-based $CO_2$ capture systems. *Energy Policy* 34(18):3765–3772.
**47** Clark WC, et al. (August 15, 2011) Boundary work for sustainable development: Natural resource management at the Consultative Group on International Agricultural Research (CGIAR). *Proc Natl Acad Sci*, 10.1073/pnas.0900231108.
**48** Knol AB, Slottje P, van der Sluijs JP, Lebret E (2010) The use of expert elicitation in environmental health impact assessment: A seven step procedure. *Environ Health* 9:19.
**49** Aspinall W (2010) A route to more tractable expert advice. *Nature* 463(7279):294–295.
**50** Devilee JLA, Knol AB (2011) Software to support expert elicitation: An exploratory study of existing software packages (Dutch National Institute of Public Health and Environment, Bilthoven, The Netherlands), RIVM Letter Report 630003001/2011, 98 pp. Available at www.rivm.nl/bibliotheek/rapporten/630003001.pdf. Accessed April 25, 2014.
**51** O'Hagan A, Oakley JE (2010) SHELF: The Sheffield Elicitation Framework, Version 2.0 (School of Mathematics and Statistics, Univ of Sheffield, Sheffield, UK). Available at www.tonyohagan.co.uk/shelf/. Accessed April 25, 2014.
**52** Morris DE, Oakley JE, Crowe JA (2014) A web-based tool for eliciting probability distributions from experts. *Environ Model Softw* 52:1–4.
**53** Evans JS, Graham JD, Gray GM, Sielken RL, Jr. (1994) A distributional approach to characterizing low-dose cancer risk. *Risk Anal* 14(1):25–34.
**54** Evans JS, et al. (1994) Use of probabilistic expert judgment in uncertainty analysis of carcinogenic potency. *Regul Toxicol Pharmacol* 20(1 Pt 1):15–36.
**55** Budnitz RJ, et al. (1995) *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and the Use of Experts* (Lawrence Livermore National Laboratory, Livermore, CA), UCRL-ID 122160.
**56** Budnitz RJ, et al. (1998) Use of technical expert panels: Applications to probabilistic seismic hazard analysis. *Risk Anal* 18(4):463–469.
**57** Budnitz RJ, et al. (1997) *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts* (US Nuclear Regulatory Commission, Washington), NUREG/CR-6372, Vol 2.
**58** Dowlatabadi H, Morgan MG (1993) A model framework for integrated studies of the climate problem. *Energy Policy* 21(3):209–221.
**59** Morgan MG, Dowlatabadi H (1996) Learning from integrated assessment of climate change. *Clim Change* 34(3-4):337–368.
**60** Refsgaard JC, et al. (2006) A framework for dealing with uncertainty due to model structure error. *Water Resour* 29(11):1586–1597.
**61** Mastrandrea MD, et al. (2010) *Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties, Intergovernmental Panel on Climate Change (IPCC).* Available at www.ipcc.ch/pdf/supporting-material/uncertainty-guidance-note.pdf. Accessed April 25, 2014.
**62** Intergovernmental Panel on Climate Change (2013) Summary for policymakers. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Stocker TF, et al. (Cambridge Univ Press, New York).
**63** Funtowicz SO, Ravetz JR (1990) *Uncertainty and Quality in Science for Policy* (Kluwer, Norwell, MA), 229 pp.
**64** van der Sluijs JP, Risbey JS, Ravetz J (2005) Uncertainty assessment of VOC emissions from paint in The Netherlands using the NUSAP system. *Environ Monit Assess* 105(1-3):229–259.
**65** van der Sluijs JP, et al. (2005) Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: The NUSAP system. *Risk Anal* 25(2):481–492.
**66** Intergovernmental Panel on Climate Change (2007) *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Solomon S, et al. (Cambridge Univ Press, New York).
**67** Intergovernmental Panel on Climate Change (2001) *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, eds Houghton JT, et al. (Cambridge Univ Press, New York).
**68** Schneider S, et al. (2007) *Climate Change 2007: Impacts, Adaptation, and Vulnerabilities. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Parry ML, et al. (Cambridge Univ Press, New York), pp 779–810.
**69** Oppenheimer M, O'Neill BC, Webster M, Agrawala S (2007) Climate change. The limits of consensus. *Science* 317(5844):1505–1506.
**70** Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. *Risk Anal* 19(2):187–203.
**71** Clemen RT, Winkler RL (2007) *Advances in Decision Analysis: From Foundations to Applications*, Edwards W, Miles RF, Jr., von Winterfeldt D (Cambridge Univ Press, New York), pp 154–176.
**72** Cooke RM, Goossens LLHJ (2008) TU Delft expert judgment data base. *Reliab Eng Syst Saf* 93(5):657–674.
**73** Clemen RT (2008) Comment on Cooke's classical method. *Reliab Eng Syst Saf* 93(5):760–765.
**74** Lin SW, Cheng CH (2009) The reliability of aggregated probability judgments obtained through Cooke's classical method. *Journal of Modeling in Management* 4(2):149–161.
**75** Karvetski CW, et al. (2013) Probabilistic coherence weighting for optimizing expert forecasts. *Decis Anal* 10(4):305–326.
**76** Dalkey N, Helmer O (1972) An Experimental Application of the Delphi Method to the Use of Experts (RAND Corporation, Santa Monica, CA) Report RM-727/1 Abridged.
**77** Sackman H (1975) *Delphi Critique* (Lexington Books, Lexington, MA).
**78** Woudenberg F (1991) An evaluation of Delphi. *Technol Forecast Soc Change* 40(2):131–150.

SI for M. Granger Morgan "The Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy"

# A Simple Illustration of the Process of Eliciting a Subjective Probability Distribution

As the main text explains, a well-developed protocol for expert elicitation may entail a variety of activities, only some of which involve asking an expert to assess the likely value of an uncertain coefficient as a subjective probability distribution.

An example of the protocol used in Zickfeld et al. (16) can be found at:
http://www.pnas.org/content/suppl/2010/06/28/0908906107.DCSupplemental/Appendix.pdf

The protocol used in Abdulla et al. (24) can be found in Appendix S2 at:
http://www.pnas.org/content/suppl/2013/05/22/1300195110.DCSupplemental/sapp.pdf

This box provides a very simple illustration of how that actual process of eliciting a probability distribution might proceed.

Suppose that I have a colleague who has driven to the airport midday from our offices, many times. It is midday now and the colleague is sitting next to me in my office. I want to elicit a probability distribution that provides his judgment of how long he believes it will take him to drive to the airport if he leaves for the parking lot to get his car right now.

First, we should probably break the question up into at least three parts:

1. Time to get to his car
2. Time to drive to the airport
3. Time to get from his car to the gate

For simplicity in this illustration I'll focus on just part 2.

Before I ask my colleagues any questions we need to agree on some general assumptions. I am interested in his judgment assuming normal traffic at this hour, no major accidents, no Presidential motorcades, no ice storms, no terrorist attacks, etc. We also assume that his car starts, has adequate gas, and has no mechanical problems.

Having agreed on these general assumptions, the interview dialogue might run something like this:

Me: Once you are in your car what is the maximum amount of time you could expect it to take to drive to the airport right now?
Colleague: 50 minutes.
Me: Has it ever taken you any longer than that?
Colleague: Yea, once it took 60 minutes and I missed my flight.
Me: With normal traffic could it take longer than that?
Colleague: I suppose maybe 65 minutes.
Me: Do you want to up your maximum time from 45 to 65?
Colleague: Yea, I guess I should.

Me: OK, now what's the minimum time for the drive to the airport?

Colleague: So, now I know that you're going to push me on this, so let's see, it is 30 miles and the speed limit is 55, but everyone drives 60. So 30 miles at 60 mph, that's 30 minutes. Sometimes I push it a bit more so I'll say between 25 and 30 minutes.

This dialogue results in my marking the range illustrated in Fig. B1. The objective in these initial exchanges is to get all the evidence brought to mind for my colleague as to minimize the impact of the heuristic of "availability" (see main text). In doing this, it is common to use strategies such as counter examples, so as to establish as wide a range as possible and minimize overconfidence. In more technical examples, a common strategy is to say something like "you said the minimum [maximum] value is X. Suppose that when the actual value becomes known it turns out to be 0.95[1.05] X. Can you think of any way in which that might occur?" If the expert can offer an explanation, then he or she might decide to increase the bounds.

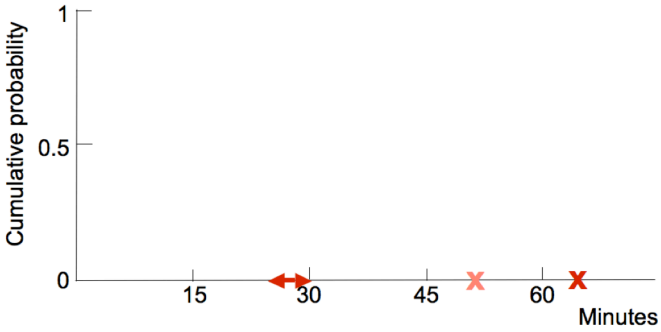


**Fig. B1.** Establishing the upper and lower bounds on the time it will take my colleagues to drive to the airport.

Continuing with the airport drive-time example, having established the range, I would then start to ask questions such as:

Me: What's the probability that your drive to the airport will take less than 60 minutes?

Colleague: 0.98.

Me: What's the probability that the drive will take more than 40 minutes?

Colleague: 0.65.

Me: What's the probability…etc.

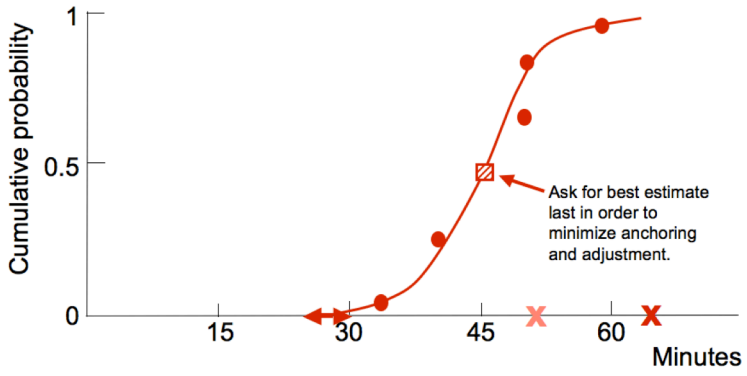


**Fig. B2.** Elicited distribution of the time it will take for my colleagues to drive to the airport.

Through a series of such questions we would build up a distribution of the sort shown in Fig. B2. If my colleague's estimates appear to be scattered, I might also phrase questions in the form "Give me a time such that you think there is at least a 30% chance you can drive to the airport in less time than that."

Finally, while I will ask my colleague for a best or median estimate, I will not pose that question until I have completed all my other questions so as to minimize the influence of the heuristic of "anchoring and adjustment" (see main text).

In virtually all the elicitations I have run, the experts have been very numerate and have chosen to answer questions directly in terms of probabilities. When respondents are not very numerate, folks in the decision analysis community sometimes ask the expert to respond by adjusting the colored section of a probability wheel of the sort shown in Fig. B3.

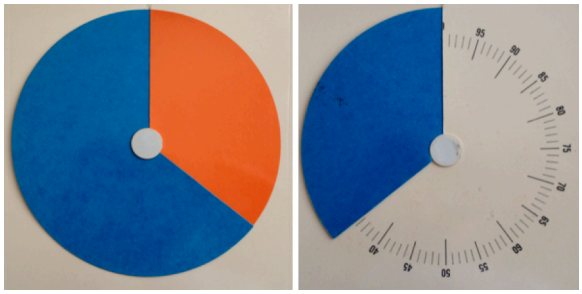

**Fig. B3.** Example of the sort of probability wheel that is sometimes used by the decision analysis community when eliciting experts who are not particularly numerate. Respondents are asked to adjust the size of the orange pie section (left) to match their probability. The value can then be read off the scale on the back (right). The specific wheel shown was made by Decision Focus, Inc.