Some Thoughts on Expert Elicitation

Prof. M. Granger Morgan
Department of Engineering and Public Policy
Carnegie Mellon University
412-268-2672
granger.morgan@andrew.cmu.edu

Portions of this talk are based on a paper titled "The Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy" that will appear in *PNAS*.

A definition

Expert elicitation involves the process of seeking carefully reasoned judgments from experts about an uncertain quantity or process in their domain of expertise, often in the form of subjective probability distributions.

Not a substitute for doing the science

Expert elicitation should *not* be viewed as a low cost alternative to doing the needed science.

However, when a decision must be made on a time-scale that is short compared with the time to complete additional needed science, well conducted expert elicitation can provide valuable insight and guidance to decision makers.

Interpretation of probability

A subjectivist or Bayesian interpretation of probability is employed 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.

While 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.

Three notes of caution

1. The topic must be one for which there are people who have predictive expertise.

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 upon 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.

Three notes of caution

2. Qualitative uncertainty words such as "likely" and "unlikely" are not enough. Such words can mean very different things to different people, or to the same people in different situations.

Three notes of caution

2. Qualitative uncertainty words are not enough.

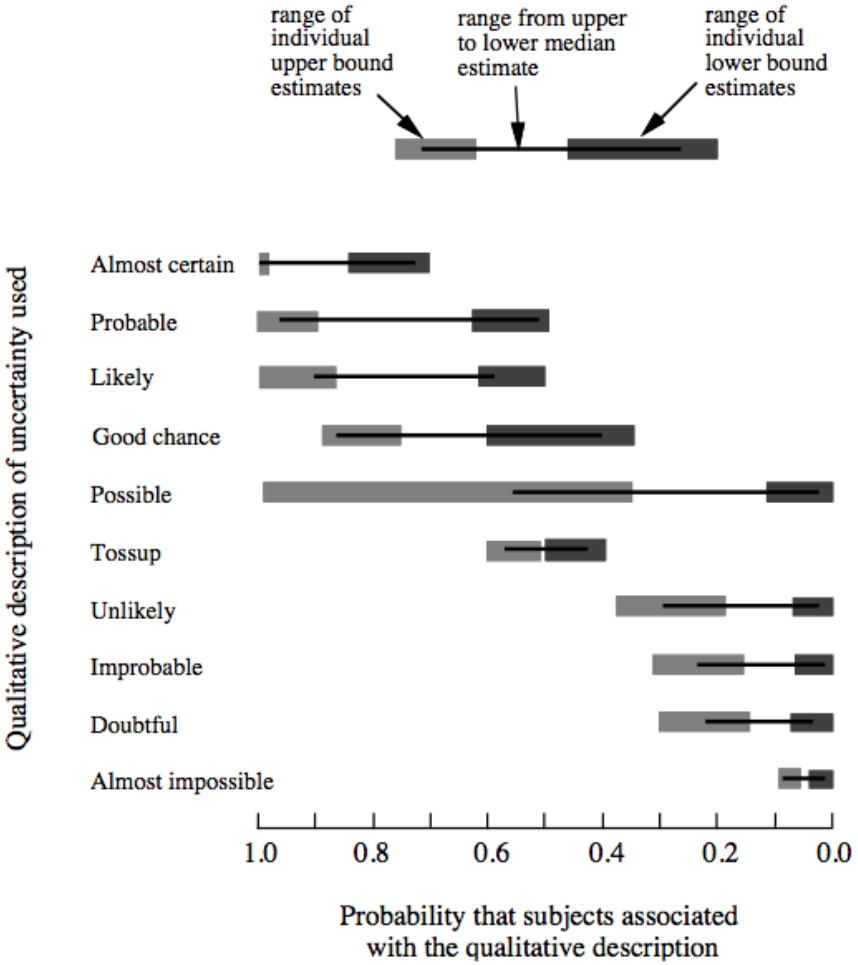


Figure redrawn from: T.S. Wallsten et al., "Measuring the vague meanings of probability terms," *Journal of Experimental Psychology: General* 155(4): 348-365, 1986.

Results obtained 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 the EPA cancer guidelines. Note that, even in this relatively small and expert group, the minimum probability associated with the word "likely" spans four orders of magnitude, the maximum probability associated with the word "not likely" spans more than five orders of magnitude, and there is an overlap of the probabilities the different experts associated with the two words.

Figure from M.G. Morgan, "Uncertainty analysis in risk assessment," *Human and Ecological Risk Assessment 4*(1): 25-39, 1998.

Three notes of caution

3. All such judgments are subject to the cognitive heuristics we all use when making judgments about uncertain events or quantities.

Three notes of caution

3. All such judgments are subject to the cognitive heuristics we all use when making judgments about uncertain events or quantities. Two that are especially relevant:

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 experts' 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.

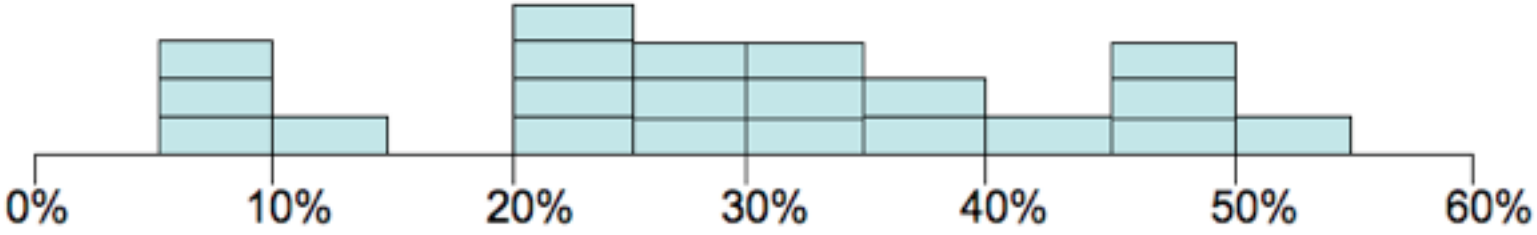
For details see: A. Tversky and D. Kahneman, "Judgments under uncertainty: Heuristics and biases," *Science, 185*(4157): 1124-1131, 1974; and D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and biases* (Cambridge University Press, UK and New York), 1982.

Three notes of caution

3. All such judgments are subject to the cognitive heuristics we all use when making judgments about uncertain events or quantities. Two that are especially relevant:

Anchoring and Adjustment: 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. In order 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.

One consequence of these heuristics is ubiquitous overconfidence



Percentage of estimates in which the true value lay outside of the respondent's assessed 98% confidence interval.

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 probabilities that are too narrow).

A more detailed summary can be found in M.G. Morgan, M. Henrion, with a chapter by M. Small, *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis* (Cambridge University Press), 1990.

Ubiquitous overconfidence…(Cont.)

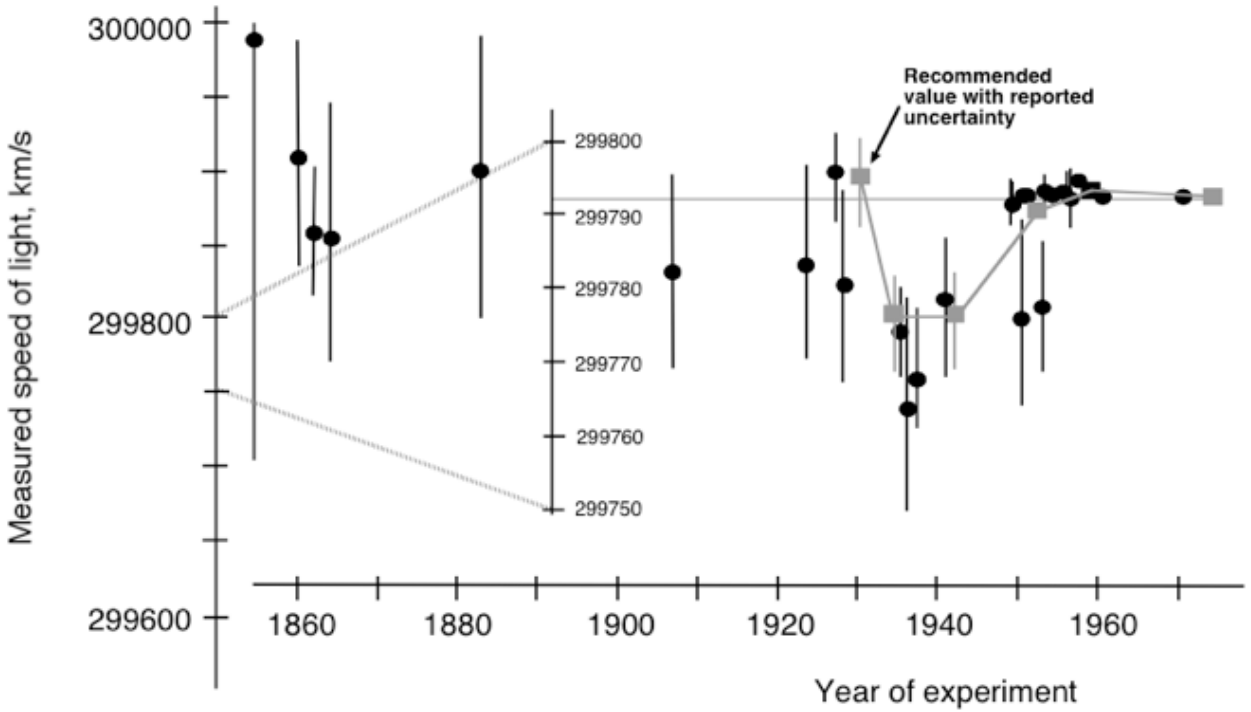


Figure redrawn from M. Henrion, B. Fischhoff, "Assessing uncertainty in physical constants," *American Journal of Physics* 54(9): 791-798, 1986.

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. 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, 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.

Developing an elicitation protocol

A primary output of many expert elicitations is a set of subjective probability distributions on the value of the quantities of interest.

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.

Steps in eliciting a CDF

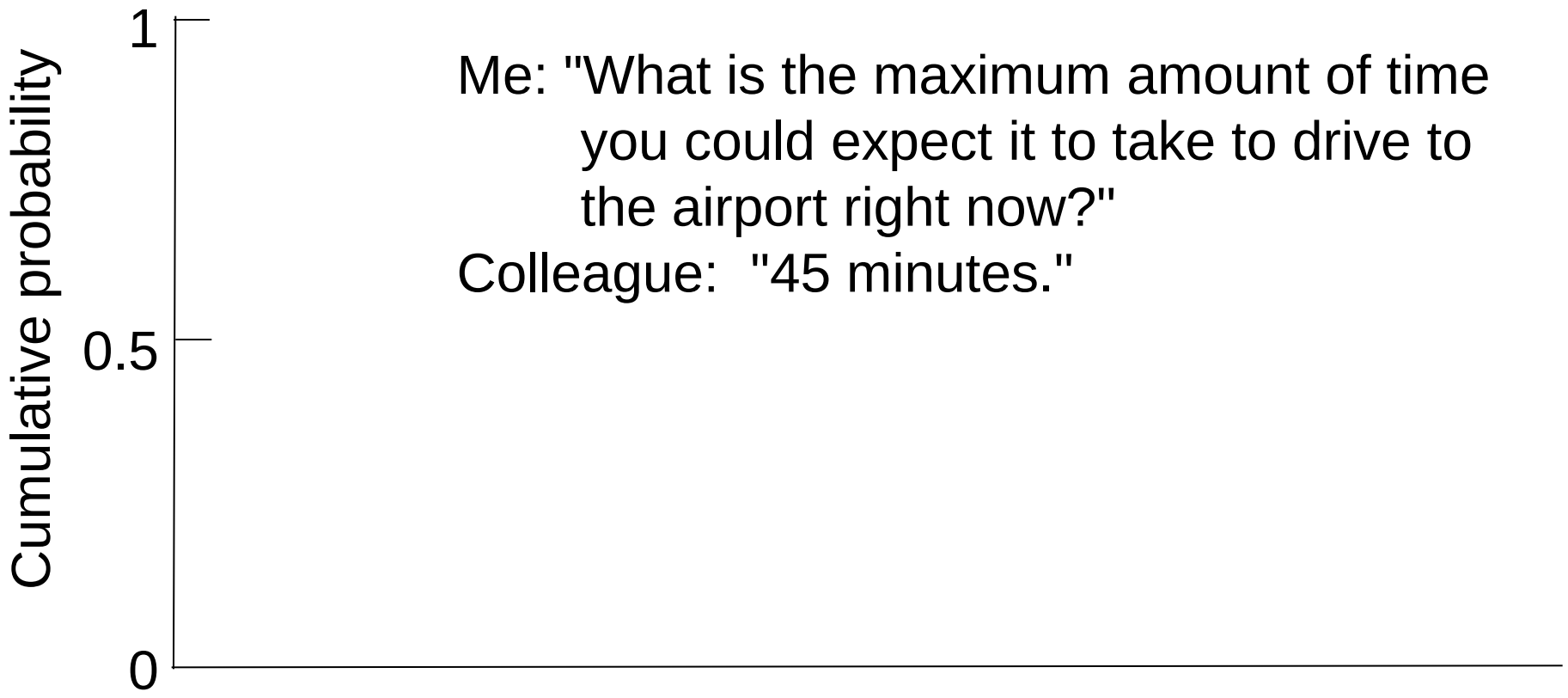
Suppose, for example, that I have a colleague who has driven to the airport mid-day, many times. I want to elicit a CDF of how long it will take him to drive to the airport if he leaves right now. I might break the question up into three parts:

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

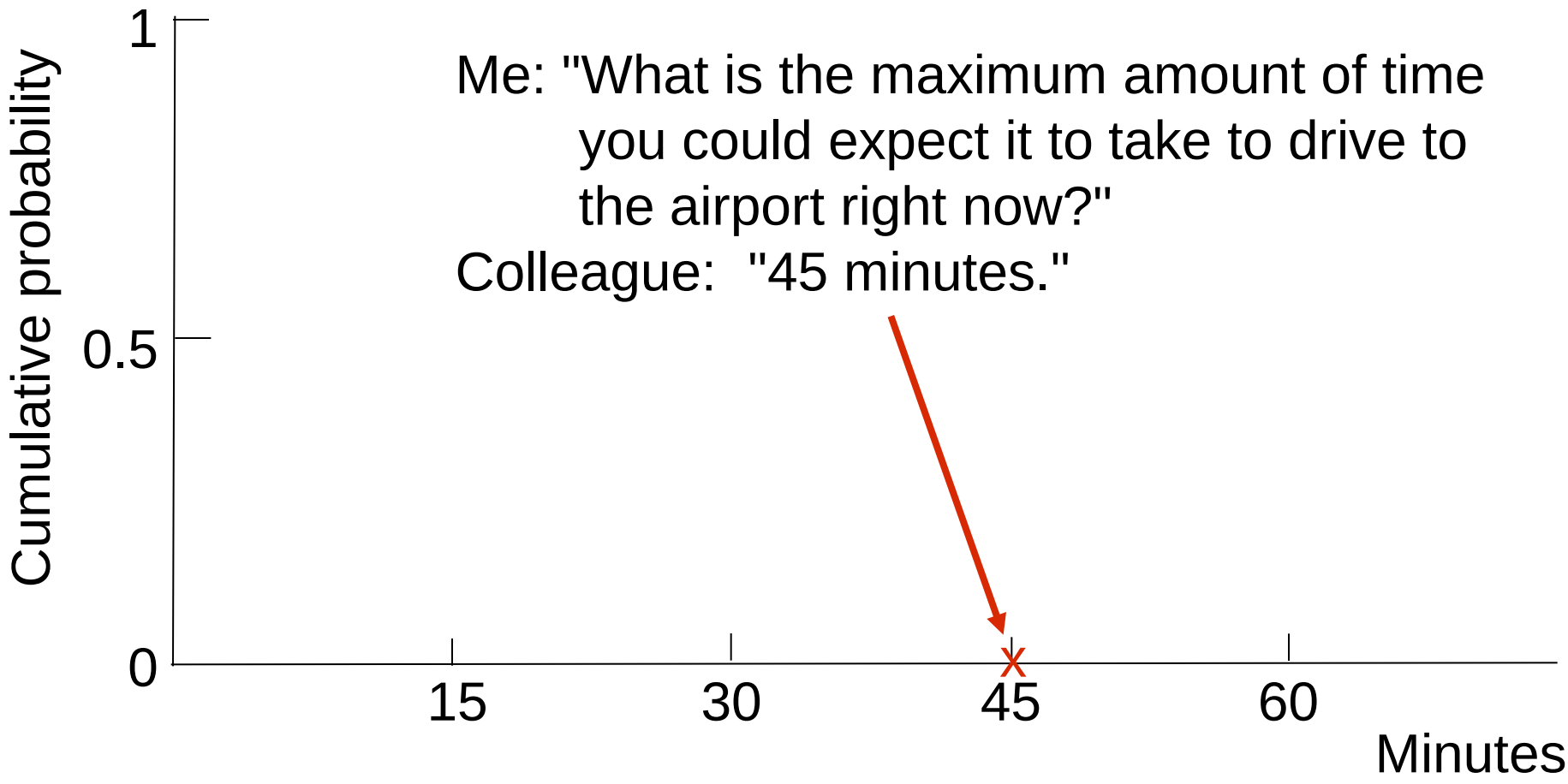
For simplicity I'll focus on just step 2.

We need to also specify just normal traffic, no major accidents, no Presidential motorcades, no ice storm, no terrorist attacks, etc.

Steps in eliciting a CDF…(Cont.)



Steps in eliciting a CDF…(Cont.)



Steps in eliciting a CDF…(Cont.)

Me: "Has it ever taken you any longer than that?"

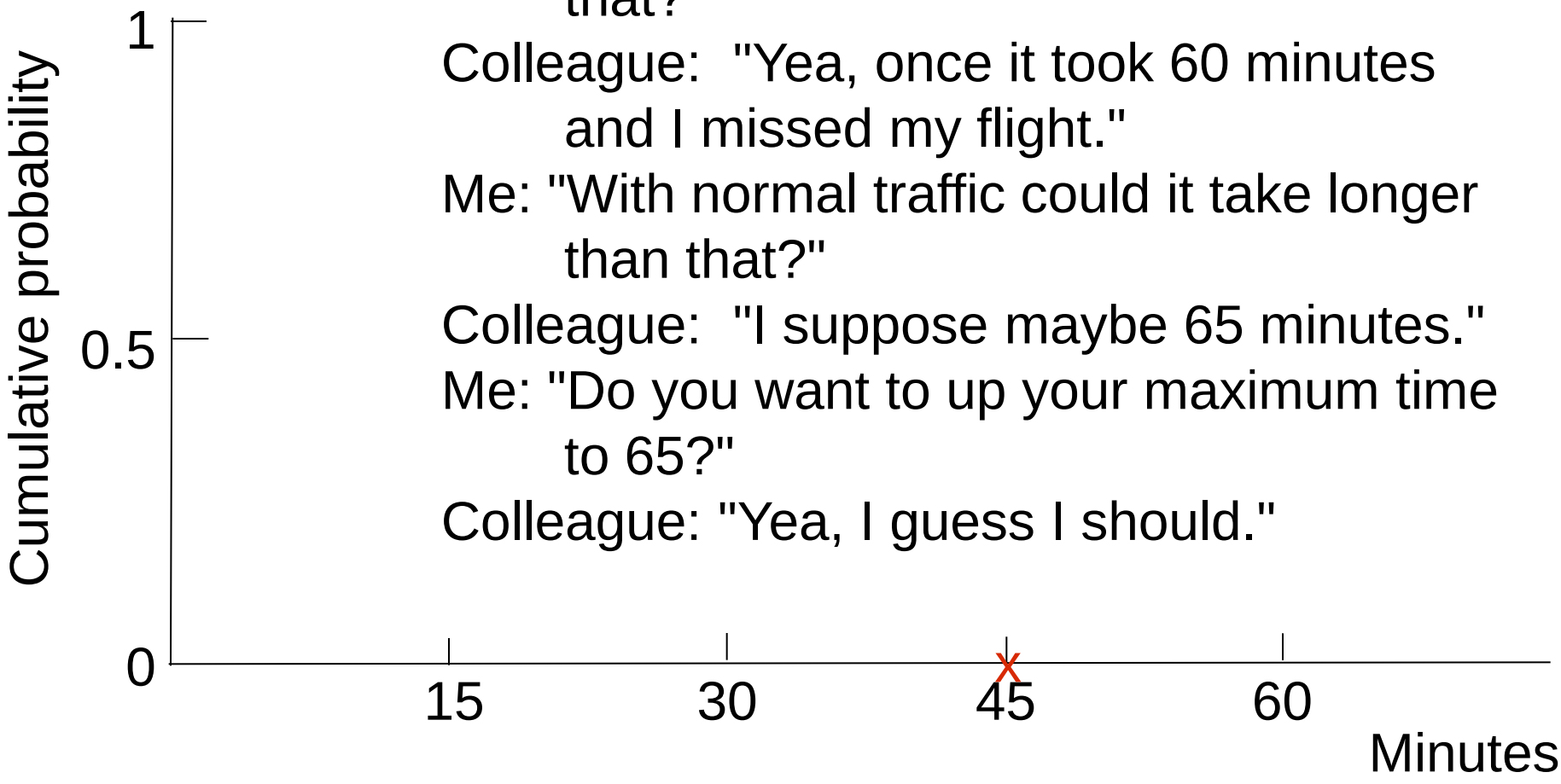
Colleague: "Yea, once it took 60 minutes and I missed my flight."

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 to 65?"

Colleague: "Yea, I guess I should."



Steps in eliciting a CDF…(Cont.)

Me: "Has it ever taken you any longer than that?"

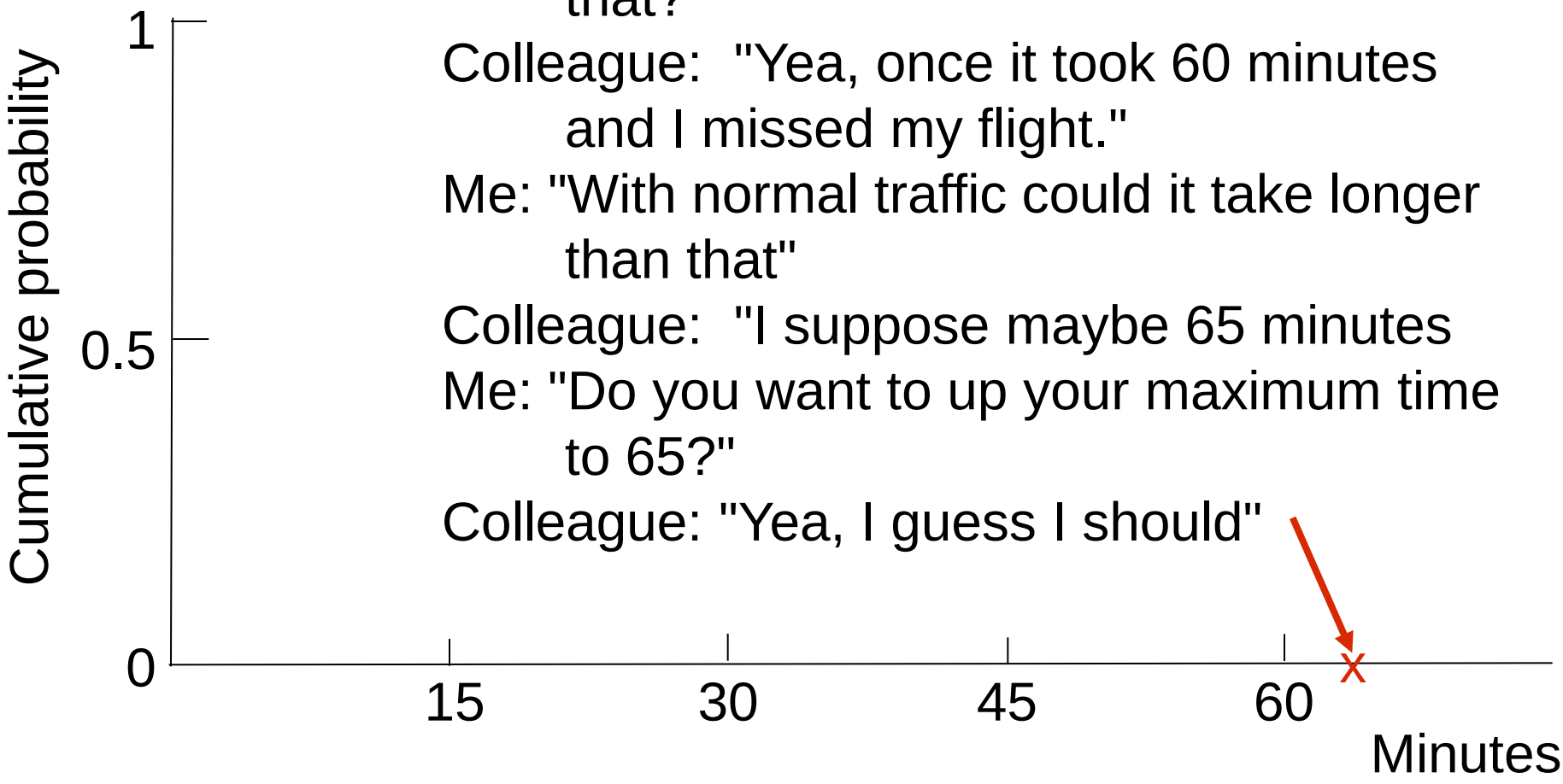
Colleague: "Yea, once it took 60 minutes and I missed my flight."

Me: "With normal traffic could it take longer than that"

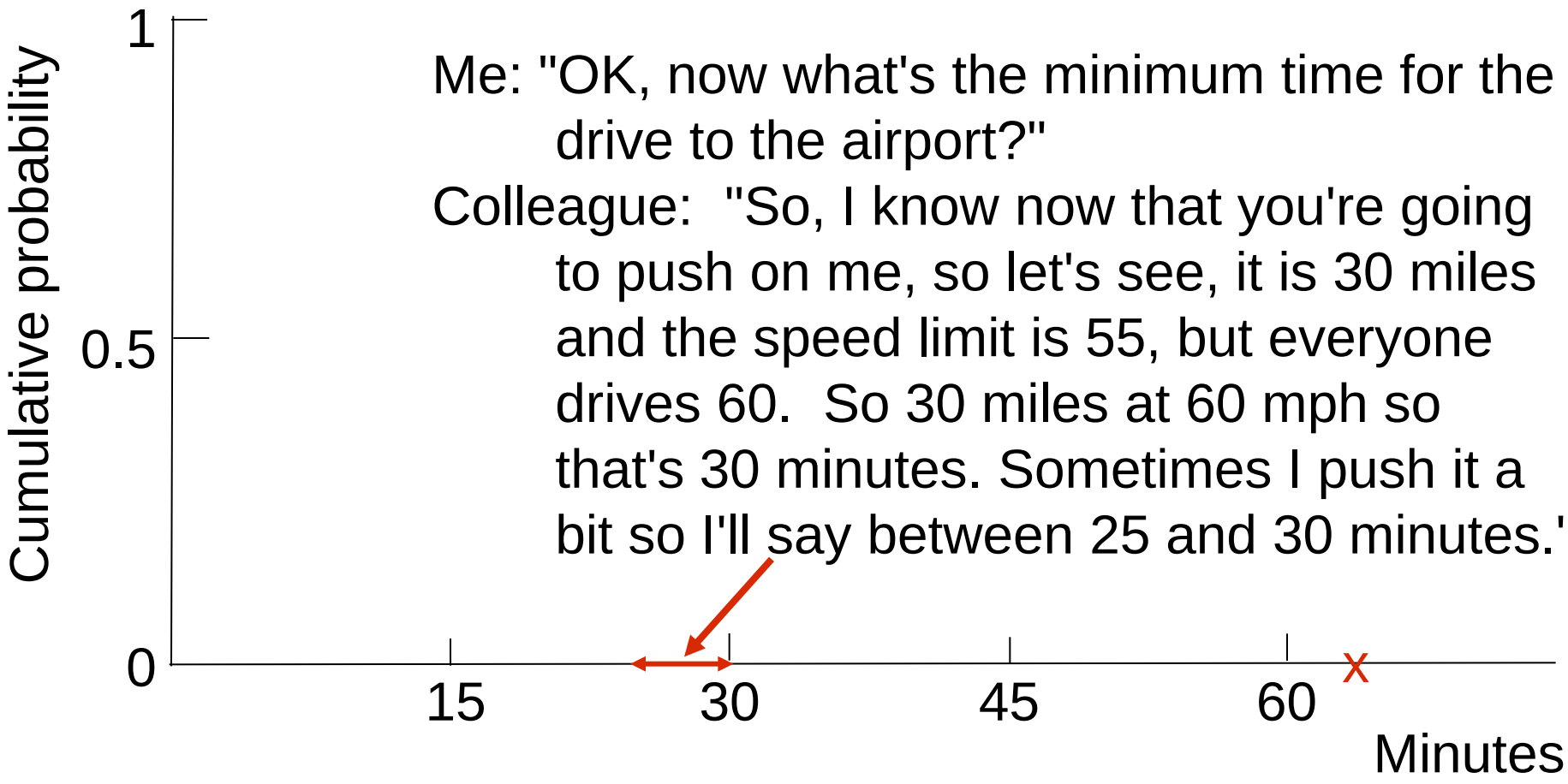
Colleague: "I suppose maybe 65 minutes

Me: "Do you want to up your maximum time to 65?"

Colleague: "Yea, I guess I should"



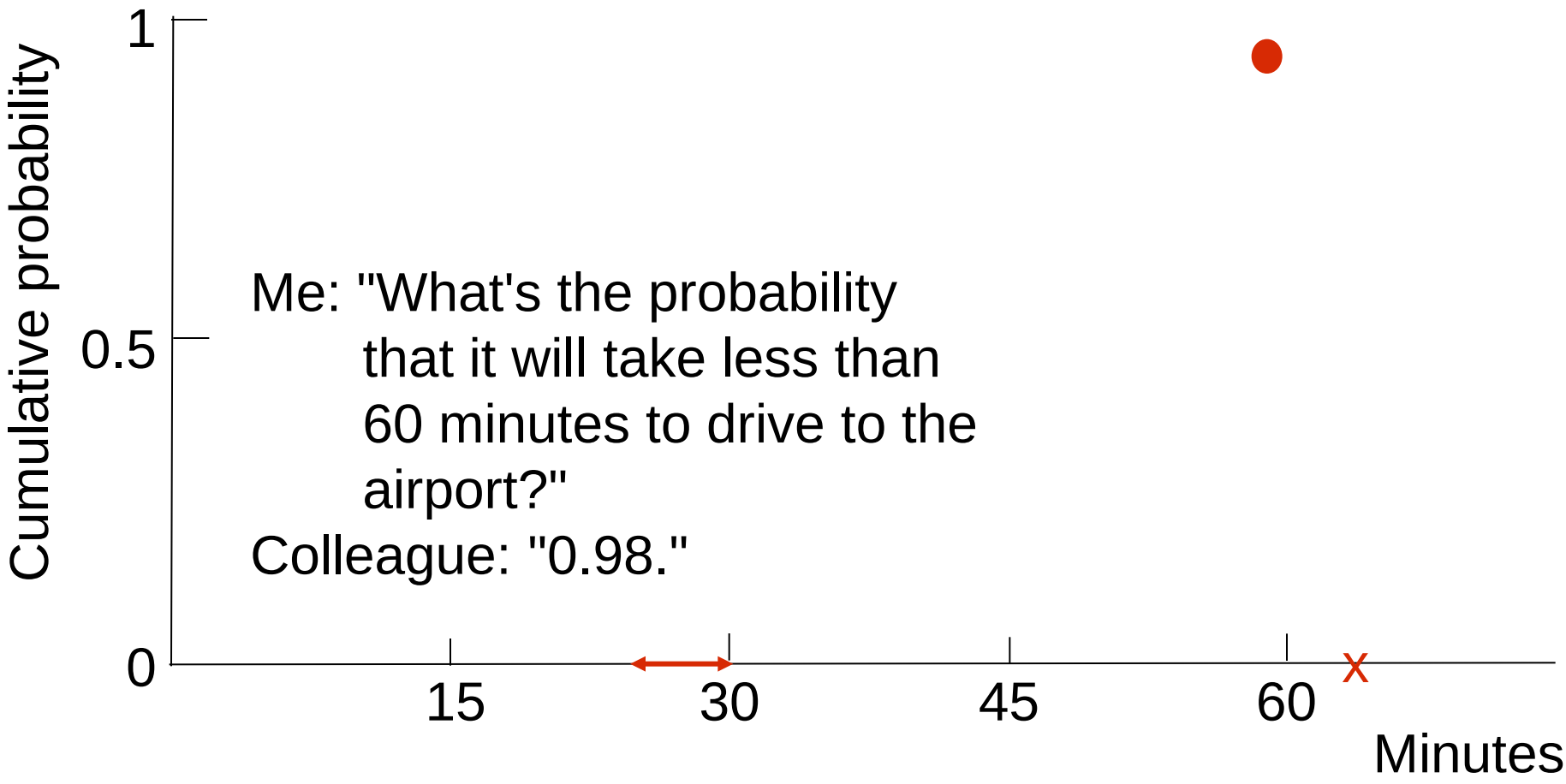
Steps in eliciting a CDF…(Cont.)



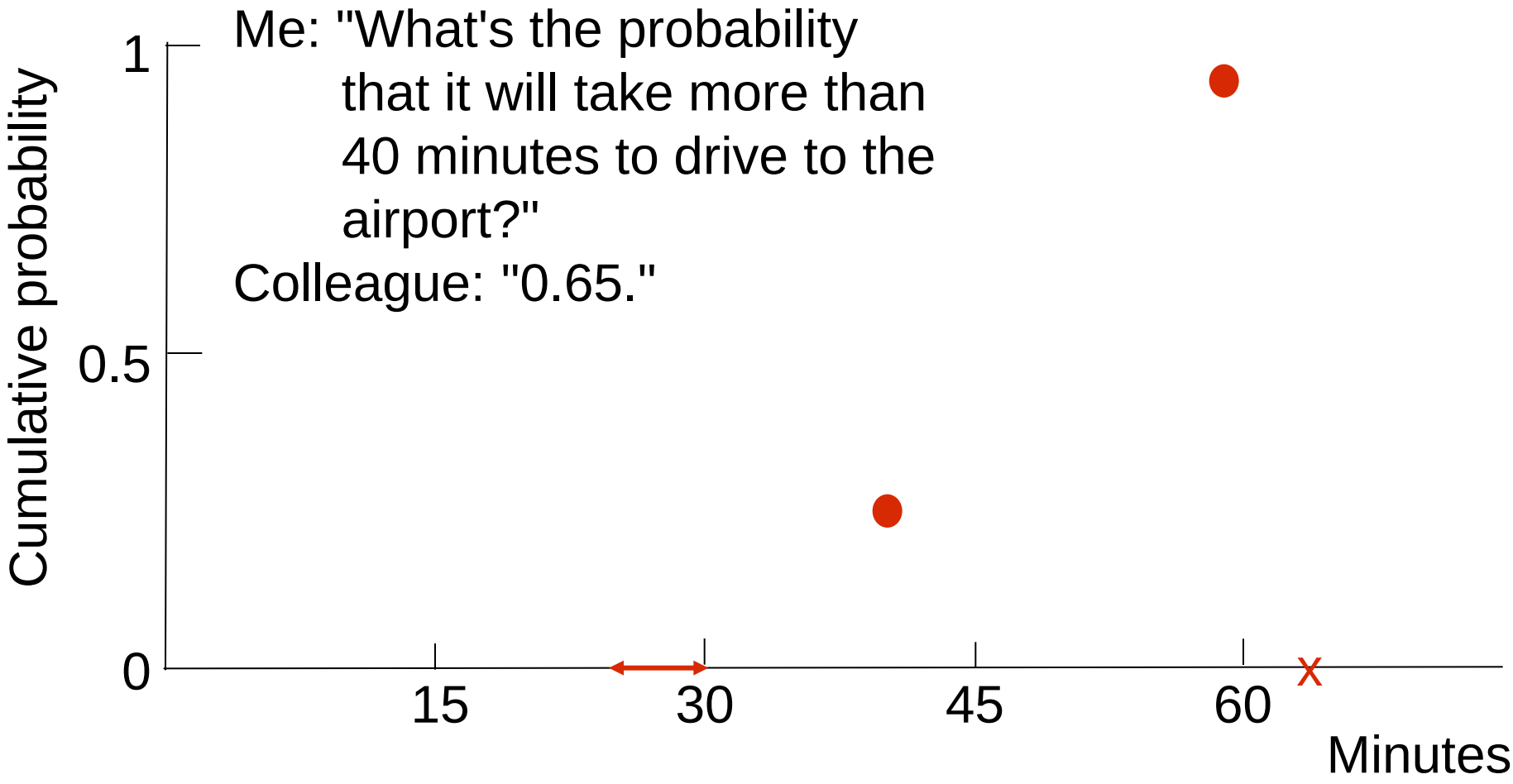
Me: "OK, now what's the minimum time for the drive to the airport?"

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

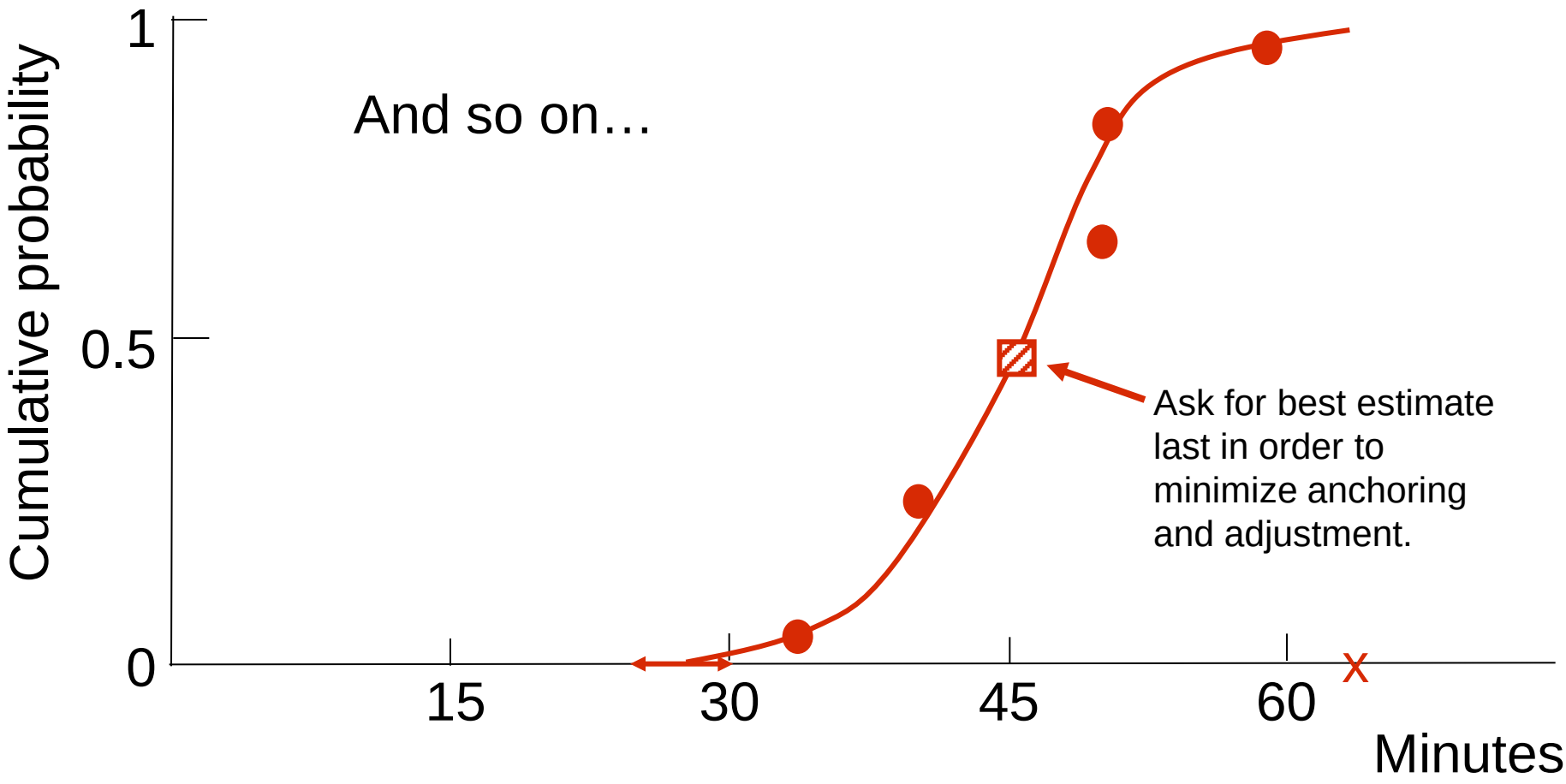
Steps in eliciting a CDF…(Cont.)



Steps in eliciting a CDF…(Cont.)



Steps in eliciting a CDF…(Cont.)



Some other issues

My colleagues and I have 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 strength of influences *versus* source of uncertainty, and to focus on the most important of the latter in formulating their probabilistic responses.

Other issues…(Cont.)

In most of the elicitations I have conducted, I have involved an excellent post-doc 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."

Selecting experts

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 the major perspectives and interpretations that exist within the community.

In studies we have conducted, we have relied on our own judgment and reading of the literature. There are more formal methods that can be used if that becomes important.

Should you combine experts?

There is a considerable literature on combining the judgments of experts into a single summary distribution.

In general, I think it is better not to try to combine them but to use the results to display the range of expert judgment.

Indeed, if experts are using very different underlying models of the science, the combined distribution may not adequately represent anyone's judgment.

One should definitely *not* combine expert PDFs if they are to be used as an input to a non-linear model.

Uncertainty about model functional form

We and a few others have worked on dealing with uncertainty about the underlying model - i.e., how the science works.

Perhaps most relevant to this audience is work by John Evans et al. They have 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 employed 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.

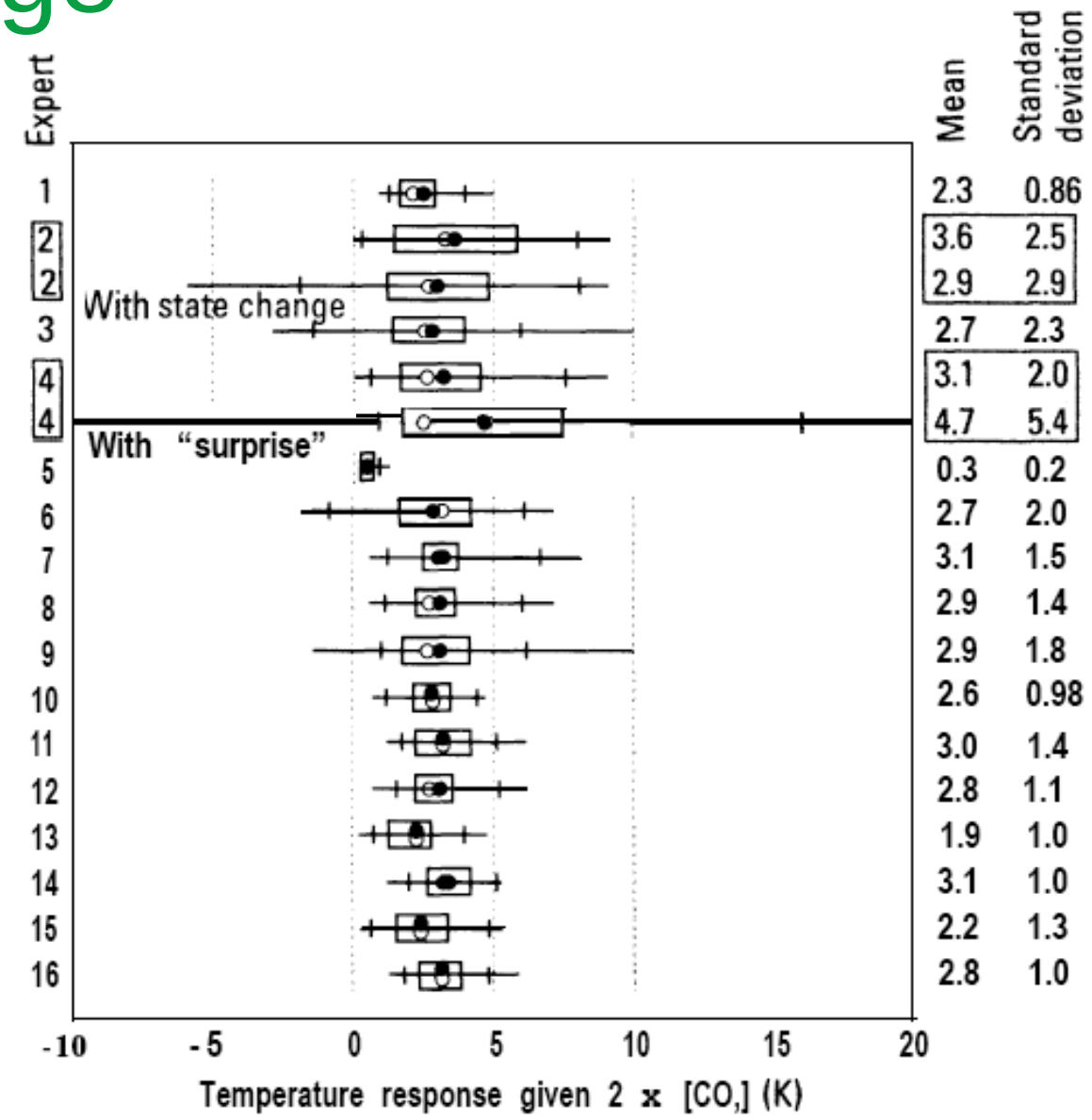
Evans JS, Graham JD, Gray GM, Sielken RL, Jr (1994) A distributional approach to characterizing low-dose cancer risk. *Risk Analysis* 14(1): 25-34.

Evans JS, et al. (1994) Using of probabilistic expert judgment in uncertainty analysis of carcinogenic potency. *Regulatory Toxicology and Pharmacology* 20(1 pt.1): 15-36.

A few examples drawn from some of the elicitations I have conducted…

Equilibrium change in global average temperature

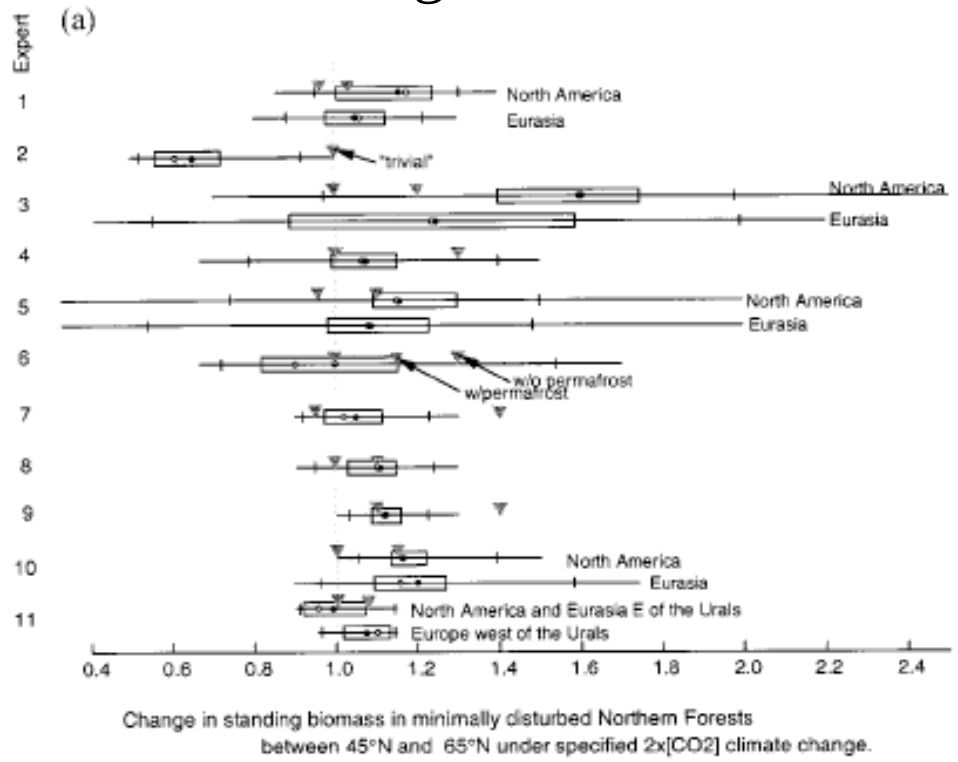
200 years after a $2xCO_2$ change



From M. Granger Morgan and David Keith, "Subjective Judgments by Climate Experts," *Environmental Science & Technology*, *29*(10), 468A-476A, October 1995.

Northern forests after 500 years

Standing biomass



Soil biomass

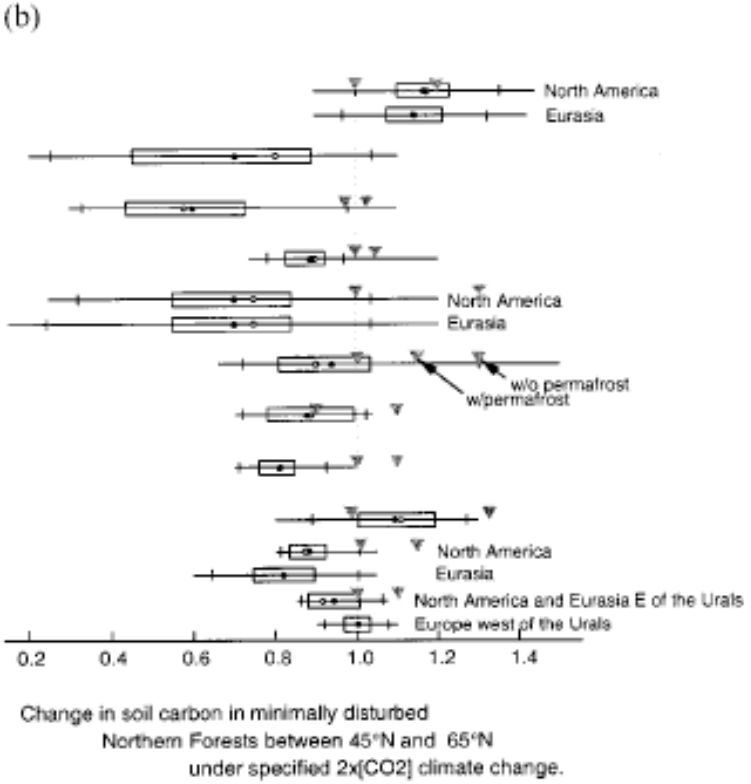
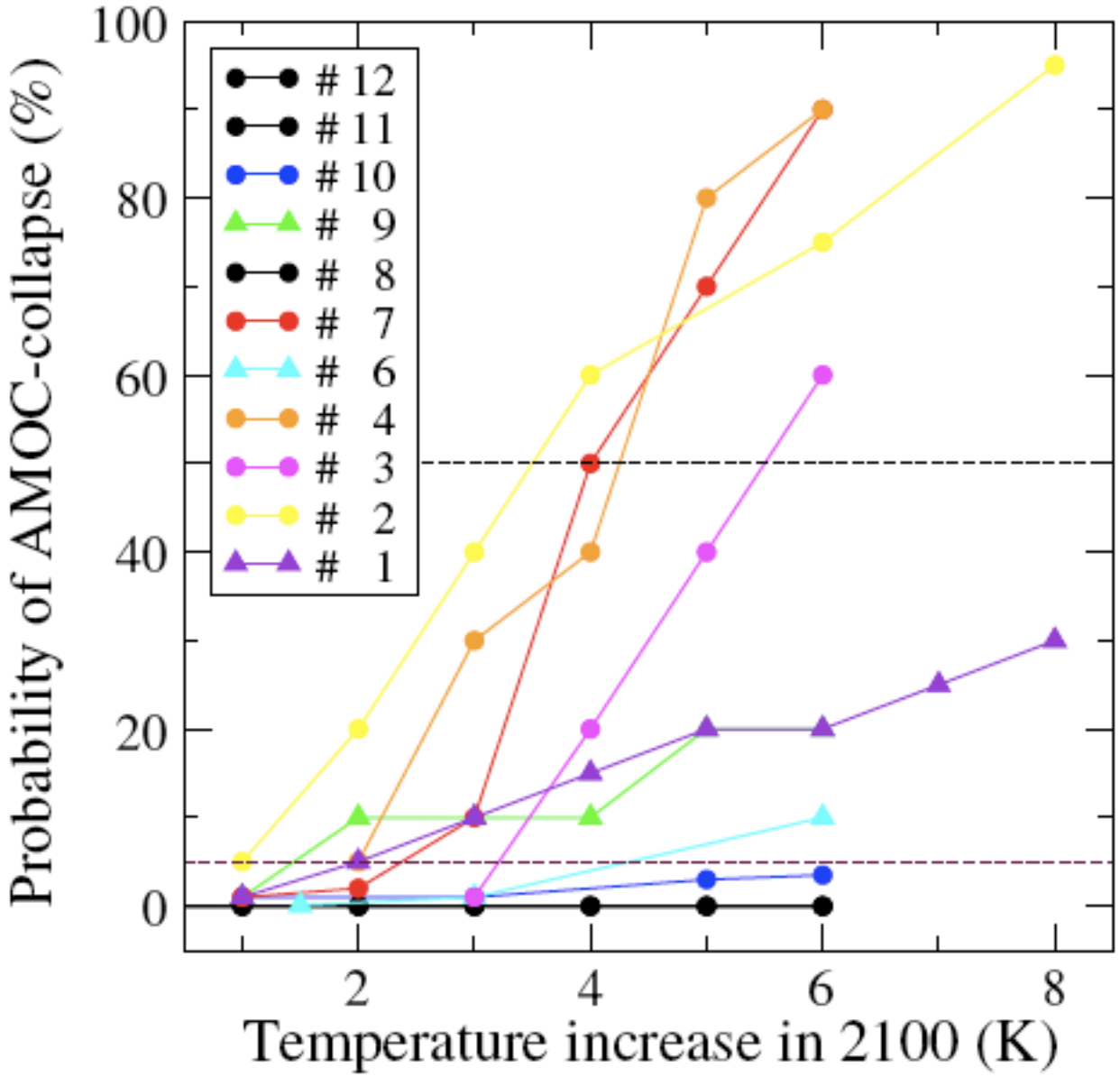


Figure 1. Box plots summarizing elicited expert subjective probability distributions of (a) standing biomass and (b) soil carbon in minimally disturbed northern forests at least 500 years after a specified $2 \times [CO_2]$ climate change. When not otherwise noted, results are for North America. Horizontal lines display the full range of the distributions. Vertical tick marks indicate the 90% confidence intervals. Boxes denote the 50% confidence intervals. Solid points indicate means, open circles indicate medians. The shaded triangles indicate the estimated range of response for a doubling of CO_2 alone, without any accompanying climate change.

From M. Granger Morgan, Louis F. Pitelka and Elena Shevliakova, "Elicitation of Expert Judgments of Climate Change Impacts on Forest Ecosystems," *Climatic Change*, *49*, 279-307, 2001.

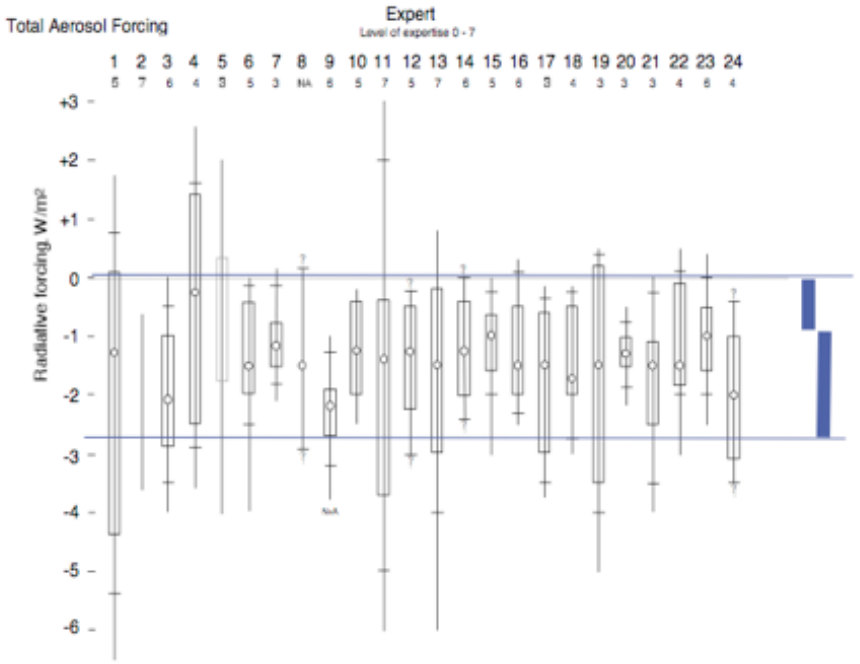
Probability of AMOC collapse



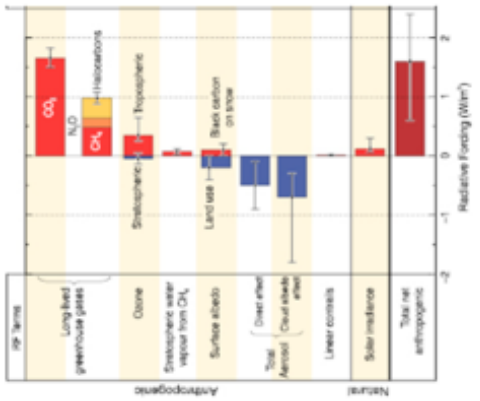
From Kirsten Zickfeld, Anders Levermann, Till Kuhlbrodt, Stefan Rahmstorf, M. Granger Morgan and David Keith, "Expert Judgements on the Response on the Atlantic Meridional Overturning Circulation to Climate Change," *Climatic Change*, *82*, 235-265, 2007.

Comparison with IPCC assessment consensus results

Individual expert assessments of total radiative forcing from aerosols from Morgan et al. (15)

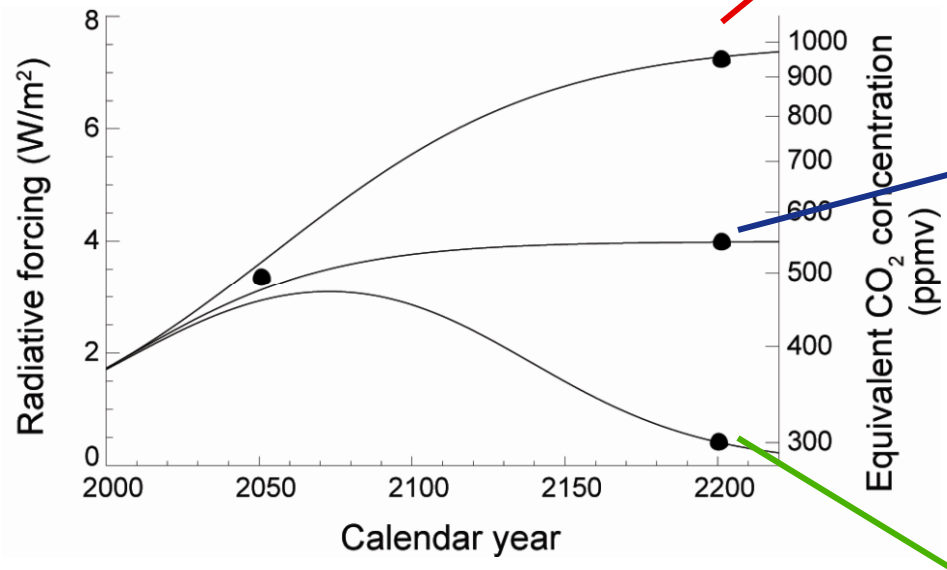


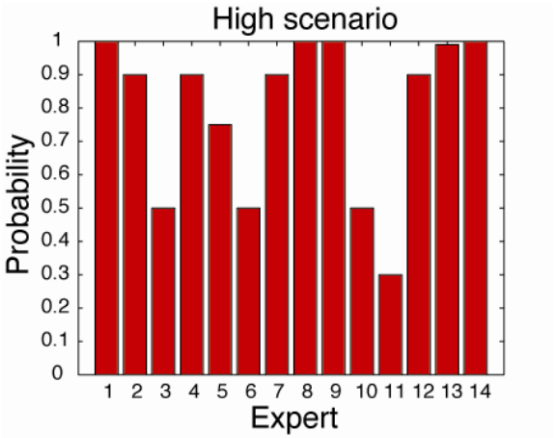
Consensus estimates of radiative forcing from the IPCC 4th assessment (61). The blue bars show the direct and indirect effects from aerosols:

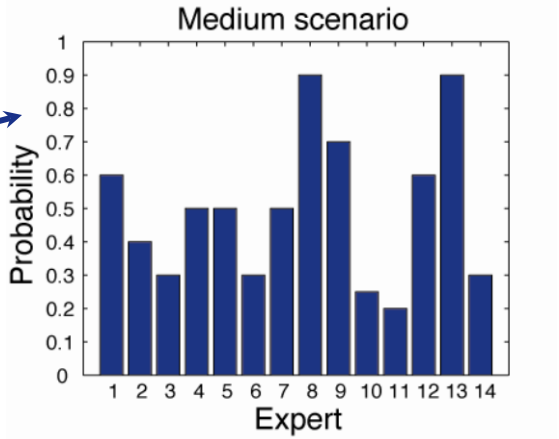


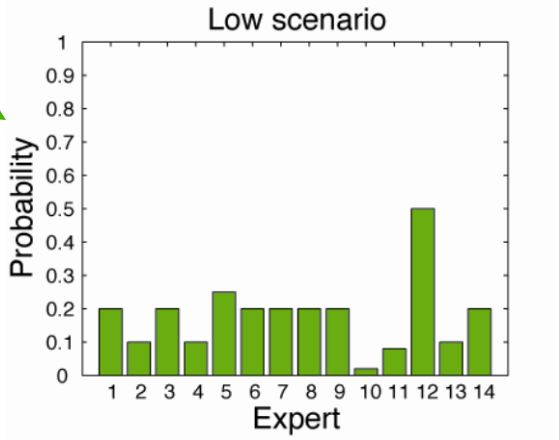
From M. Granger Morgan, Peter Adams, and David W. Keith, "Elicitation of Expert Judgments of Aerosol Forcing," *Climatic Change*, *75*, 195-214, 2006.

Probability of a basic state change



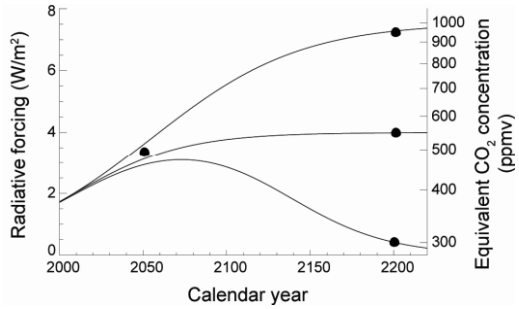


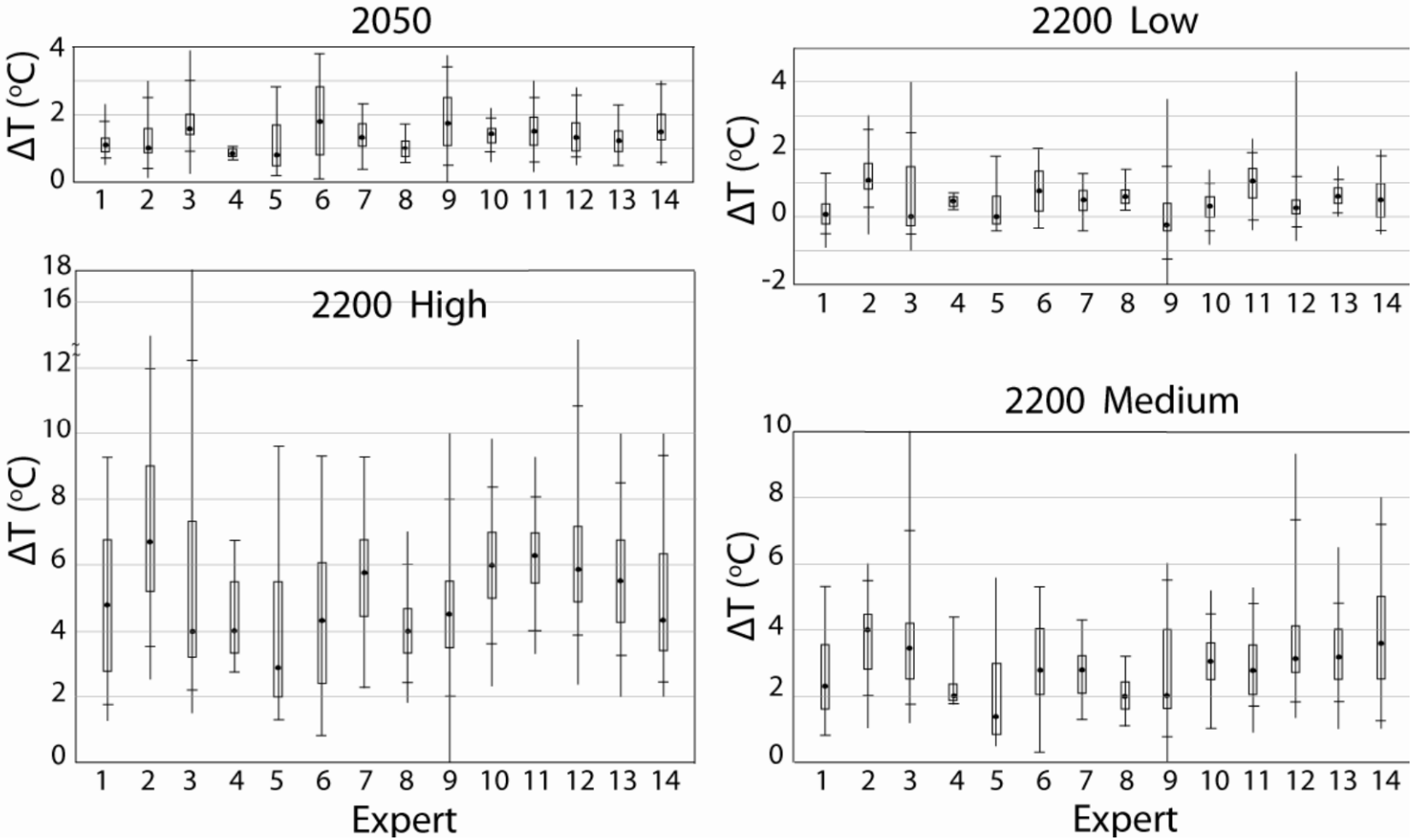




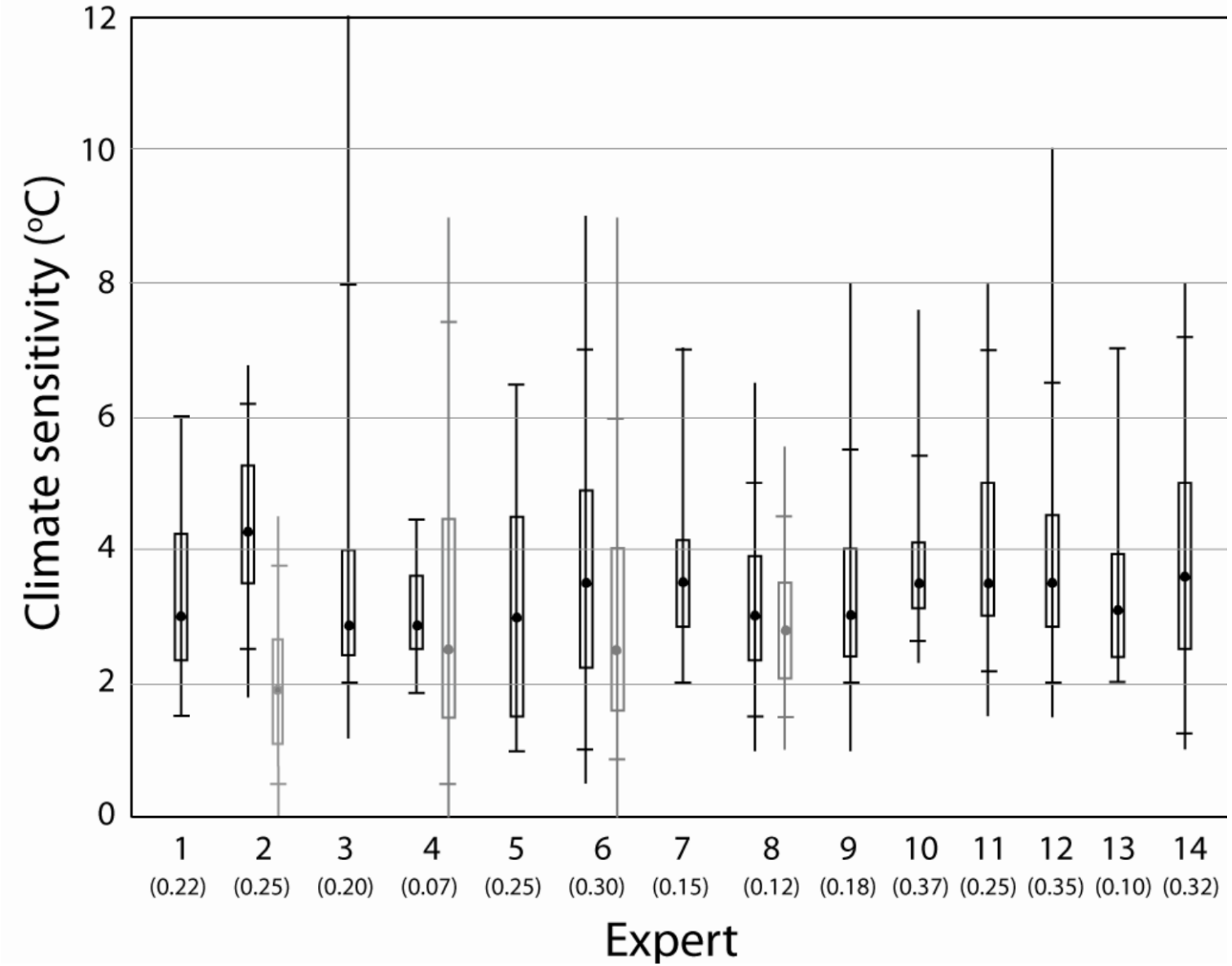
From Kirsten Zickfeld, M. Granger Morgan, David Frame and David W. Keith, "Expert Judgments About Transient Climate Response to Alternative Future Trajectories of Radiative Forcing," *PNAS*, *107*, 12451-12456, July 13, 2010.

Summary of PDFs in ΔT

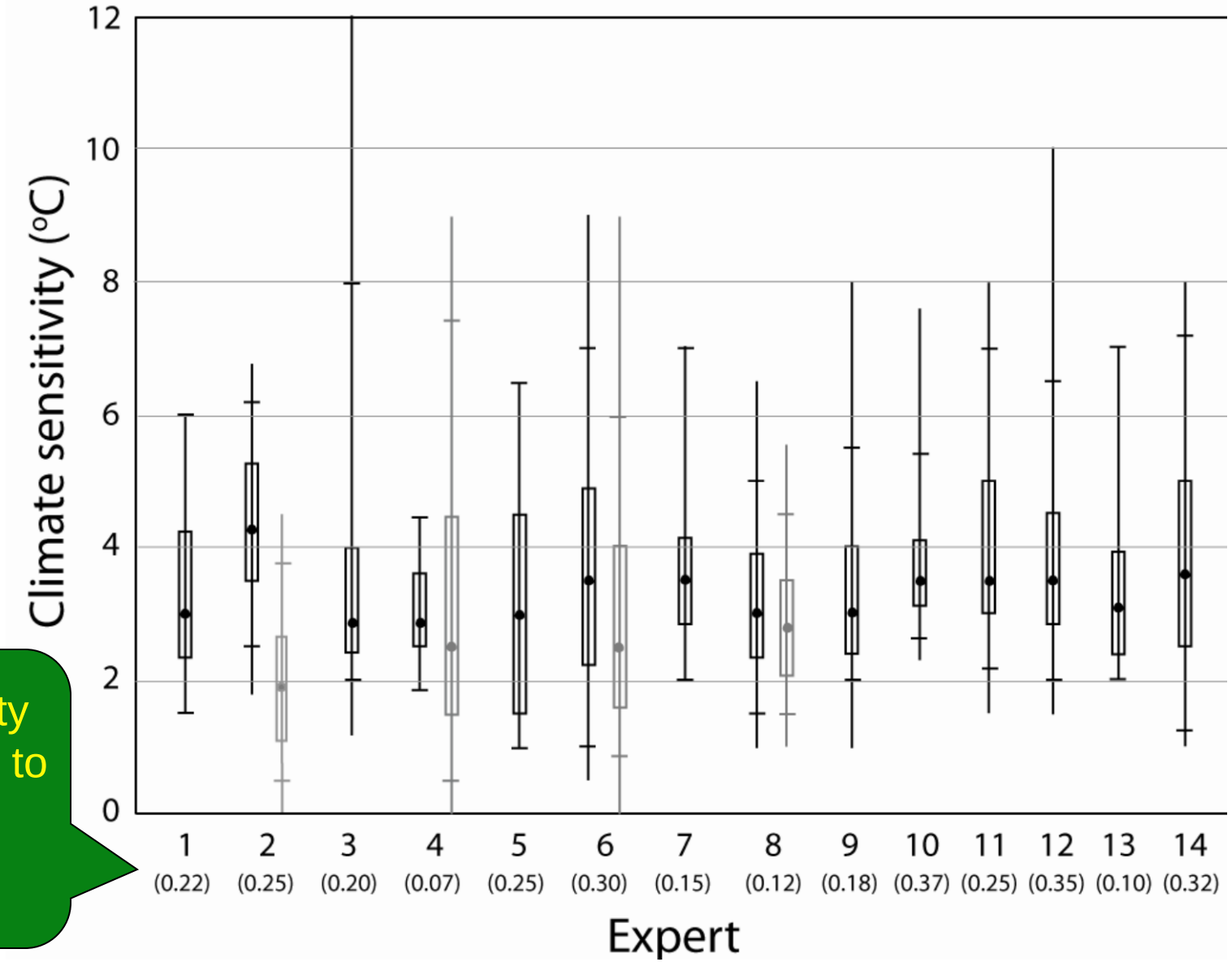




Climate sensitivity



Climate sensitivity



Probability allocated to values above 4.5° C

Likely cost of SMRs

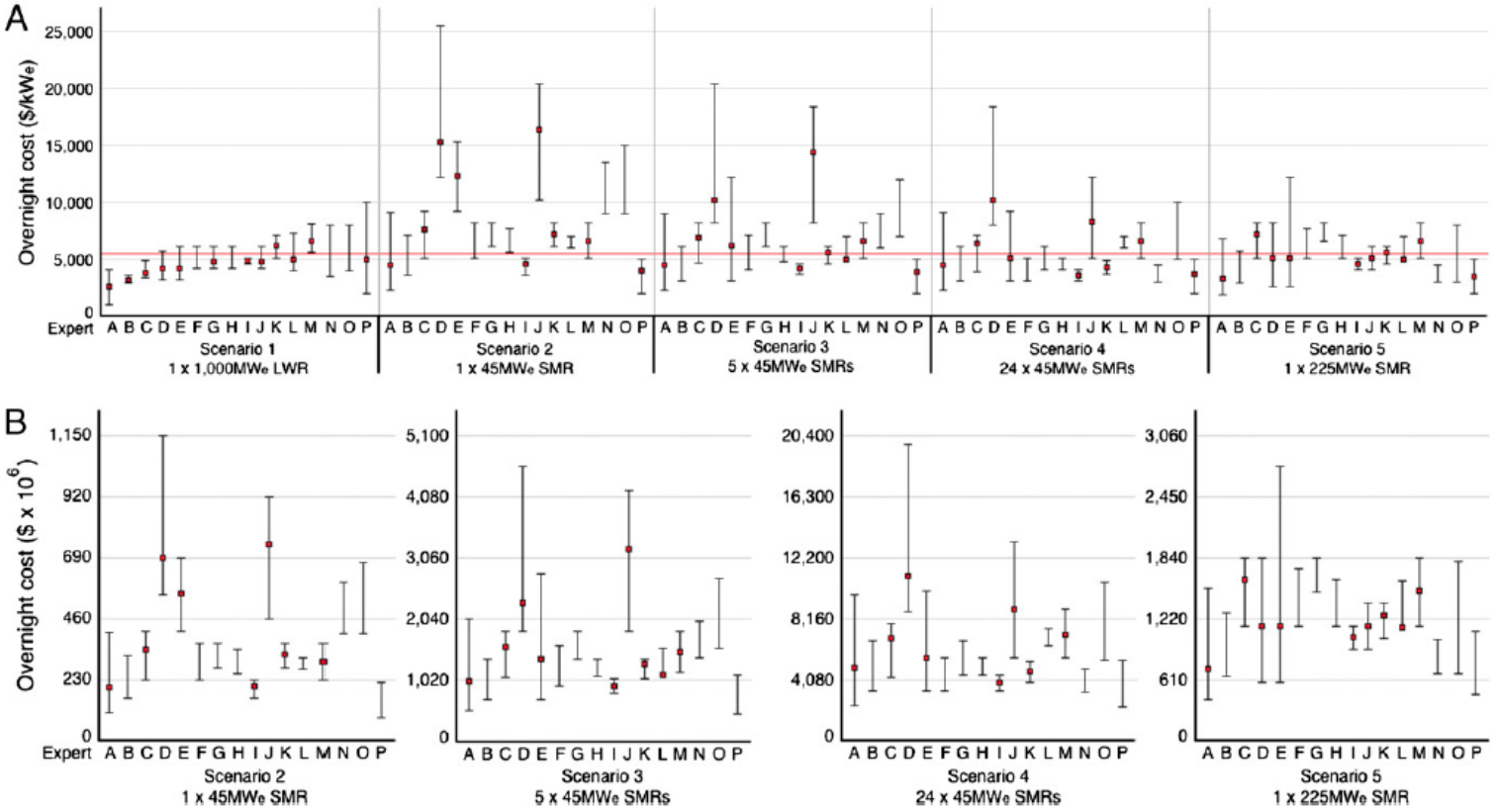


Fig. 1. Estimates of overnight cost elicited from sixteen nuclear power experts for each of five nuclear reactor deployment scenarios. (*A*) Each expert (A through P) provided estimates of the overnight cost per kilowatt of reactor capacity for each scenario. The details of the scenarios are noted on the horizontal axis. The solid line represents the Energy Information Administration's 2011 estimate of the overnight cost of a dual-unit large light water reactor (LWR) plant (30). (*B*) For the four SMR-plant configurations, each of the estimates in *A* is multiplied by plant capacity to arrive at project cost. Expert M's estimate included owner's cost (i.e., costs that fall out of the vendor's scope, such as site work, transmission upgrades, etc.).

From Ahmed Abdulla, Inês Azevedo and M. Granger Morgan, "Expert Assessments of the Cost of Light Water Small Modular Reactors," *PNAS*, *110*(24), 9686-9691, 2013.

Some sources

U.S. Environmental Protection Agency (2011). Expert Elicitation Task Force White Paper. Available on line at http://www.epa.gov/stpc/pdfs/ee-white-paper-final.pdf

Morgan MG, Henrion M, with a chapter by Small M (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis* (Cambridge University Press).

Morgan MG with Dowlatabadi H, Henrion M, Keith D, Lempert R, McBride S, Small M, and Wilbanks T (2009). *CCSP 5.2 Best Practice Approaches for Characterizing, Communicating, and Incorporating Scientific Uncertainty in Decisionmaking*. A Report by the Climate Change Science Program and the Subcommittee on Global Change Research. National Oceanic and Atmospheric Administration, Washington, DC, 96pp.

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, Stefan Rahmstorf, Anand Rao, William Rish, Edward Rubin, Stephen Schneider, Debra Shenk, Patti Steranchak, Kirsten Zickfeld and many others. Much of the work has been supported by grants from NSF and EPRI. Most recent support has been under NSF cooperative agreements SES-0345798 and SES-0949710 and from the John D. and Catherine T. MacArthur Foundation.