RTI INTERNATIONAL | Memorandum

TO: Ron Evans, EPA

FROM: Carol Mansfield, RTI International

DATE: July 30, 2004

SUBJECT: Peer Review of Expert Elicitation

1. Introduction

The purpose of this memo is to synthesize responses given by reviewers selected to review and comment on the following document:

An Expert Judgment Assessment of the Concentration-Response Relationship Between $PM_{2.5}$ Exposure and Mortality, Prepared by Industrial Economics, Incorporated under subcontract to Abt Associates, Inc., for U.S. Environmental Protection Agency, Research Triangle Park, NC, April 23, 2004.

This section provides an overall summary of the responses received from the reviewers. This summary is followed by a section containing the specific questions sent to reviewers and a section containing a synopsis of the answers to these questions. Following the summary, a full list of the references cited in the reviews is presented. Finally, copies of the full responses supplied by the reviewers, including a detailed discussion covering the "best practices" for expert elicitation, are attached as Appendix A.

1.1 Review Summary

The overall response of the reviewers is that the expert elicitation study was a defensible body of work that should at least be considered a pilot study or best first step, if not something more substantial. The following are the main strengths the reviewers found:

- The selection criteria used are sound. The experts chosen for elicitation are well known and respected in their respective fields.
- The number of experts selected was small, but there is no magic number for such an exercise. The five experts chosen were considered a good starting point. This number could possibly be expanded in future elicitations.
- The overall study and processes involved were well documented and followed the "standard elicitation protocol."

- Despite the lack of group interactions (discussed below), the elicitation study team was able to find some consensus among the experts.

The major criticisms of the report and elicitation related to analytical topics:

- The encoding process of elicitation could be improved. In the elicitation process, the reviewers interpreted that some of the experts provided judgments based on a central tendency before providing judgments on extreme values (upper and lower ranges). This type of sequencing may introduce anchoring or adjustment heuristics, which are associated with biased estimates of uncertainty. Although the authors of the elicitation report introduce the topic of heuristics, reviewers felt that a more substantive discussion on how the study addressed known sources and any other potential forms of bias was necessary.

- The experts should have communicated before and/or after the individual interviews. Group communication prior to the individual interviews would have aided in the motivation and conditioning steps of the elicitation, while communication, either in person or through a summary document, would have allowed an expert to adjust his response based on the responses of the other experts.

- Some reviewers had problems with the manner in which the expert judgments were combined. Dr. Chris Frey included a detailed discussion on the potential problems of averaging expert judgments as done in the EPA report (see pp. A-18 to A-20). To summarize his critique, Dr. Frey believes that the combined distributions do not adequately capture the opinions of individual experts, but rather average them out. It is possible in such cases that the combined judgments may generate results that none of the experts could agree on.

2. Questions Sent to Reviewers

Figure 1 displays the cover letter, and Figure 2 displays the accompanying list of questions that was sent to reviewers to guide their evaluations of the expert elicitation study.

3. Responses to Administered Questions

Reviewers were given copies of the report and the above set of questions seeking input on their responses to several aspects of the document. Additionally, the reviewers were given the opportunity to supply specific, detailed comments on the overall expert elicitation document and an opportunity to indicate the strengths, weaknesses, and suggested areas of improvement for future elicitations. The original questions posed to the reviewers are presented below along with responses from the reviewers. In cases where the responses were consistent across all reviewers, responses are summarized to represent all reviewers. Responses found to be detailed, significant, insightful, or contrary to that of the overall group are highlighted along with the name of the reviewer providing the response.

Figure 1. Cover Letter

July 27, 2004

Name
Address

Dear Dr. [name]:

Thank you for agreeing to serve as a peer reviewer of EPA's Expert Judgment Assessment of the Concentration-Response Relationship between $PM_{2.5}$ Exposure and Mortality.

At the suggestion of the National Research Council (NRC), EPA is taking steps to improve its characterization of uncertainty in its benefit estimates. Mortality effects associated with air pollution comprise the majority of the benefits estimated in EPA's retrospective and prospective Section 812A benefit-cost analyses of the Clean Air Act (EPA, 1997, 1999) and in regulatory impact analyses (RIAs) for rules such as the Heavy Duty Diesel Engine/Fuel Rule (EPA, 2000). However, calculating uncertainty bounds is often hampered by the absence of conclusive scientific data. In the absence of such data, NRC recommended that probabilistic distributions can be estimated using techniques such as formal elicitation of expert judgments.

EPA recently conducted an elicitation of expert judgment of the concentration-response relationship between $PM_{2.5}$ exposure and mortality. The charge for this peer review is to provide technical feedback on the methods employed for this expert elicitation, with particular emphasis on the strengths and weaknesses of the methodology. We are also interested in methods for combining the results from the individual experts, especially given the differences in the forms of the functions. Finally, we are interested in your suggestions for improving similar elicitations (e.g., if there is another iteration of this particular elicitation or an application to a similar problem). To the extent that the results of the elicitation are influenced by the method, please also comment on the utility of the technical conclusions.

Below you will find a list of both general and specific questions that we would like you to consider in conducting your review. We do not expect you to answer each question individually, but we would like you to use them as a guide in preparing your review. Please address as many of these issues as possible but feel free to focus on areas that correspond best with your technical expertise and interests. As you read the report, you will see that the authors of the report point out several potential areas for improvement. Feel free to comment on their suggestions as well.

We request that you submit a written review no later than [insert date]. You can e-mail the review to me at carolm@rti.org. Please organize the review in the form of a memorandum or a short report (preferably in WordPerfect but otherwise in MSWord), beginning with your general impressions of the elicitation and then moving to your more specific comments.

Thanks again for your participation. If you have any questions, please feel free to contact me via e-mail or at (919) 541-8053 or Wanda Throneburg at (919) 541-6261.

Sincerely,

Carol Mansfield
Senior Economist
Environmental and Natural Resource Economics Program
RTI International

Enclosure

Figure 2. Questions to Guide Review

Questions for Reviewers

Please feel free to address other topics you consider important.

General topics:

- What are "best practices" for conducting expert elicitations? What "best practices" are essential for a defensible elicitation?
- How does the EPA's elicitation compare to "best practices"? What are the strengths and weaknesses of this elicitation?
- How should the individual expert opinions be combined?
- How could the elicitation be improved?
- Are the elicitation methodology and process appropriate as a benefit analysis technique to characterize uncertainty?

Specific topics:

1. Participants
    a. Did the set of participants chosen reflect the views of other scientists in the field?
    b. Was the number of participants appropriate?
    c. Was the method for choosing participants acceptable?
    d. For these participants, does the potential for motivational bias exist?
    e. Are the relevant fields represented?
2. Topic: Long- and short-term mortality
    a. Were the topics adequately described to the participants (eliminated ambiguity)?
    b. Were the correct questions asked (i.e., did they effectively address the topics without introducing bias)? Do you think that word choice, structure, or the order of the questions affected the quality of the results?
3. Preparation
    a. Were any materials missing that should have been included in the Briefing Book? Should any materials have been excluded?
    b. Was the time allowed for review of the materials adequate?
    c. Some participants were more prepared than others-is this a problem? If so, how can this problem be addressed?

(continued)

Figure 2. Questions to Guide Review (continued)

d. Were expectations effectively communicated to the participants prior to the meeting?

e. Were the questions sufficiently tested prior to the elicitation?

f. Was adequate training provided for the participants prior to the elicitation? Would a pre-elicitation workshop have been helpful?

4. Meeting format

a. How do individual interviews compare to group interviews?

b Is it important to achieve consensus or not?

c. Did the interviewers perform their roles properly?

d. Was the length of the interview appropriate?

5. Aggregation of opinions and final report

a. How should the quantitative and qualitative results from the individual interviews be combined?

b. Was it appropriate to weight all the opinions equally? Are there alternative weighting systems that could have been used? EPA considered ways to calibrate the experts, which can be used to weight the experts' responses, but could not construct an applicable calibration question for this topic. Do you agree that you cannot calibrate experts on this topic? If not, do you have ideas for calibration questions.

c. Can the results of the individual interviews reasonably be combined given the differences in forms of the functions? Of the four methods presented for combining the results, are any of them sufficient given the statistical approaches taken, and would you recommend other methods for combining the results?

d. Are all of the essential elements included in the final report? Are any unnecessary elements included?

6. Recommendations for future elicitations

a. Absolute necessities for an expert elicitation (based on best practices)

b. Major strengths and weaknesses of this expert elicitation

3.1 General Topics

1. What are "best practices" for conducting expert elicitations? What "best practices" are essential for a defensible elicitation?

   Drs. Stieb and Frey provide several reference documents that provide in-depth discussions of "best practices" for expert elicitations (Merkhofer, 1987; Morgan and Henrion 1990; Morgan, Henrion, and Morris, 1980). Dr. Frey includes a detailed discussion of five steps for eliciting expert judgment and techniques associated with these steps, all of which are based on the material referenced above (see pp. A-10 to A-16).

2. How does the EPA's elicitation compare to "best practices"? What are the strengths and weaknesses of this elicitation?

   Each of the reviewers stated that they believe EPA's expert elicitation methodology generally compares favorably with "best practices." Dr. Frey again provides the most thorough analysis by identifying what he thinks were the report's strengths and weaknesses (see pp. A-16 to A-18). Dr. Stieb indicated that the experts Samet and Zeger, who worked closely on an influential study (NMMAPS) and are currently at the same institution, could have been excluded because they may have narrowed the range of views represented.

3. How should the individual expert opinions be combined?

   Several of the reviewers preferred that the expert opinions not be combined or stated that they knew of no agreed-upon method for combining results from expert elicitations. This allows for the differences in the individual distributions to be recognized. Two of the reviewers indicated that they were reasonably comfortable with the method used in this study to combine the results. Particular issues with the EPA elicitation's method of combining responses are discussed below.

4. How could the elicitation be improved?

   The reviewers were asked to identify ways in which the expert elicitation could have been improved. All reviewers provided substantial input on this topic, and their responses are summarized and grouped into three sections: Methodological/Analytical Improvements, Expert Interaction, and General Improvements.

Methodological/Analytical Improvements

A major area of concern raised by Dr. Frey was in regard to the encoding step, or the quantification of the expert's judgment. This step involves the quantitative description of the subjective probability distribution that best reflects the expert's beliefs. The reviewer suggests that a disaggregated approach could be more appropriate. It is recognized that the study team acknowledges the benefits of such an approach and did take steps to quantify the experts' judgments; however, the exact steps or process involved in the encoding is not detailed in the report. The reviewer continues by pointing out that all sources of potential bias and ways in which the elicitation protocol intends to address them need to be discussed in the report.

As identified by a number of the reviewers, a key element of the expert elicitation that could be improved is the method used for combining the expert judgments. In the report, the values provided in the elicitation were combined by taking the arithmetic mean. However, as Dr. Frey pointed out in detail, this value may result in an answer that none of the experts would agree with or capture the full range of the values suggested by the experts. He recommends a mixture distribution approach for this purpose (see pp. A-18 to A-19). Dr. Crawford-Brown suggested combining the responses after the individual distributions have been followed through to the benefits analysis stage.

A final methodological/analytical concern was due to the use of Monte Carlo analyses. Dr. Crawford-Brown felt that this was the result of the way in which the problem or question was introduced: "a single estimate of change in mortality per unit exposure." He was concerned that this may have forced some of the experts to condense their more complex understanding of the subject into a simplified context. A possible solution to this problem may be to ask the experts for confidence intervals surrounding the point estimates of any incremental change in effect along the C-R Curve. He also suggested exploring the benefits of using lognormal distributions. Using lognormal distributions would mean the distributions have the property that they are phrased in terms of "accurate to within a factor of" (see pp. A-2).

Expert Interaction

Most of the reviewers were concerned about the lack of group interaction. All reviewers said that interviews need to be conducted on an individual basis to avoid motivational or other biases. However, the benefits of group interaction before and/or after the interviews are conducted could be substantial. The interaction prior to the interview will aid in the conditioning step of the "best practices," referred to earlier. A pre-elicitation

workshop or similar meeting will allow experts to consider other experts' opinions and perspectives when formulating their responses. The post-interview interaction will aid in reconsidering the conditioning step and serve as an impetus for experts to refine their judgments. This interaction will allow experts to challenge the lines of reasoning by others and as a result give them the option to revise their responses.

General Improvements

Dr. Stieb pointed out that the range of views represented in the expert elicitation may need to be widened. This concern relates to the comment in Question 2 regarding the two experts who may have similar views based on past research efforts.

Dr. Morgan suggested that including questions about future research needs in the elicitation process would be useful. He provides several expert elicitations that have been completed recently as examples (Morgan and Keith, 1995; Granger-Morgan et al., 2001). He indicated that such questions may yield useful insights. Although individual expert elicitation is a fine strategy, one might also consider the collective expert–workshop strategy that the seismic risk community has adopted (Budnitz et al., 1995; Budnitz et al., 1998).

5. Are the elicitation methodology and process appropriate as a benefit analysis technique to characterize uncertainty?

Overall, the reviewers concluded that the expert elicitation presented and conducted is an appropriate technique for characterizing uncertainty even though they had some concerns about the post-elicitation combination of the results.

3.2 Specific Topics

1. Participants

The reviewers thought five experts was an acceptable number of participants, although most commented that a larger number might have provided some benefit. Dr. Crawford-Brown doubted including more experts would have significantly changed the analysis in any way other than in a statistical sense. A number of reviewers commended the process by which experts were selected and noted that it served as a basis for deciding that five was an appropriate number of experts. Dr. Crawford-Brown stated that the EPA elicitation should have followed its original established procedure in reference to the decision to allow the expert who chose not to participate to recommend individuals who were added to the group. Dr. Frey suggested consulting literature on the acceptable number of experts in the field of nuclear power plant probablistic risk assessment,

keeping in mind that a health-related field that is more multidisciplinary may require more experts.

A more pressing issue than the number of participants was the representativeness of the range of views. Overall, the reviewers agreed that the experts chosen reflected the views of other scientists in the field. As mentioned previously, Dr. Stieb questioned the inclusion of both Samet and Zeger as experts. He also noted that Krewski is based in Canada even though one of the stated criteria for inclusion was being based in the United States. Dr. Frey suggested that the authors state more clearly whether they believe that the experts provide an appropriate range of views. Dr. Stieb commented that a toxicologist could have been included but noted that he or she may not have been as familiar with the quantitative aspect of the epidemiological evidence.

2. Topic: Long- and short-term mortality

The reviewers thought the topics were adequately described to the participants and the correct questions were asked. Dr. Frey mentioned that the follow-up questions, especially those asking the expert to consider values outside of the range chosen, are important. He did suggest that the length of time meant by "long-term" could have been more clearly stated. Dr. Morgan suggested that, in the next steps, including some questions about future research needs may yield useful insights. Another new topic for questions, suggested by Dr. Stieb, was to ask experts how they generated their answers (e.g., consulted primary sources, computed values from specific tables).

In response to the issue of word choice, Dr. Crawford-Brown found it interesting that experts tended to focus on increases in mortality following increases in exposure rather than on decreases, as the original question prescribed. He could not think of a reason why this change would affect results but did point out that at times subtle changes in wording can change answers. Dr. Stieb noted that questions about specific percentiles should be randomized. Likewise, Dr. Frey remarked that the order in which the percentiles were elicited should be described, and the specific order should be listed in the report if order varied among experts. Drs. Stieb and Frey agreed that the extreme values should be elicited first and the central tendency elicited last to avoid biased estimates of uncertainty associated with anchoring and adjusting heuristics.

3. Preparation

The reviewers determined that the briefing book, as Dr. Crawford-Brown stated, included a "representative sample of the most influential studies." Dr. Frey suggested also including a literature review. Adequate time was allowed for reviewing the materials,

although some experts did not use the time, so that some participants were more prepared than others. The reviewers considered this an issue. One reviewer indicated that the questions had been sufficiently tested. Two of the reviewers did not comment, and one did not feel adequate to address the issue but thought that the process seemed reasonable.

All of the reviewers commented on the importance of a pre-elicitation workshop, many recommending it in future expert elicitations. Dr. Morgan thought that a pre-elicitation workshop could be used "to get the experts to place themselves and each other in some taxonomy of the field so that you could show coverage." According to Dr. Frey, a pre-elicitation workshop could serve the purpose of knowledge sharing among experts, taking care to not elicit expert judgment at this time because it could introduce motivational and other biases. He noted that the workshop could aid in the conditioning step as well. Dr. Stieb also recognized the importance of this element but understood the difficulty in scheduling the workshop around five busy schedules.

4. Meeting format

As mentioned previously, the reviewers agreed that individual interviews are preferable, but that aspects of group communication before and/or after the individual interviews have many benefits. The reviewers did not believe that achieving consensus in the experts responses was necessary. In fact, some reviewers said consensus may be an unrealistic expectation or, if a consensus is found, it might produce an artificial result with no merit. Dr. Frey thought it was a strength that the authors found some aspects of consensus among the experts, such as agreement on the relevant literature studies and this allows for establishing that the differences between the experts' responses are attributable to other factors.

The reviewers indicated that the interviewers performed their roles properly. The length of the interview was appropriate given the material. Dr. Stieb admitted to being surprised that the often busy experts agreed to such a lengthy elicitation interview.

5. Aggregation of opinions and final report

All of the reviewers had major concerns about combining results and the way they were combined. Dr. Stieb stated that he was unaware of an agreed-upon way to combine the results. Dr. Crawford-Brown questioned whether the method used was the best way to combine the results. However, these two experts were reasonably comfortable with the method used to combine results. Drs. Morgan and Frey were more critical of the combined results.

Dr. Crawford-Brown thought that the procedure currently used to combine results should only be used if it was impossible to send each individual distribution through the benefits analysis. Because no examples were provided, he was not certain what procedure was used. He stated that he would have used "the weighted average of the percentiles associated with a given parameter value, rather than the weighted average of parameter values associated with a given percentile" (see pp. A-7 to A-8).

Dr. Morgan preferred not combining the results and instead using each expert's judgments individually in analysis. He was concerned that the algorithm used to combine results might not capture the tails if only a few of the experts have wide distributions (see pp. A-25). Dr. Frey was also concerned that the extremes of the distributions are not captured in the current approach.

Dr. Frey's uncomfortableness with the method of averaging results has already been stated. In his report, he describes in detail the downfalls of the averaging approach, in particular the assumption of a 100 percent correlation between all experts' distributions (see pp. A-18 to A-20). Alternatively, he suggests using a "mixture" approach.

All reviewers preferred equal weighting and were skeptical of other weighting schemes. Dr. Frey recommended reexamining the weighting approach used.

Reviewers had some specific comments on the elements of the final report. Those not described here can be found in Appendix A. Dr. Stieb felt that the exact order of the quantitative elicitation questions could be more specific. Dr. Crawford-Brown suggested moving the Monte Carlo analysis material after the individual results to prevent confusion (see pp. A-2). Dr. Frey commented that the systematic discussion of known heuristics and other sources of bias and how the biases were addressed should be organized in one section rather than scattered throughout the report.

6. Recommendations for future elicitations

The reviewers' recommendations for future elicitations are noted throughout this memo. Their primary recommendations are

- using pre-elicitation workshops,
- using alternate methods for combining results or not combining results at all, and
- improving the encoding step.

The strengths of this elicitation are

- the procedure used to choose respondents,
- the coverage of views,
- the use of individual interviews, and
- the in-depth briefing book; these elements should be implemented in future elicitations.

References Contained in Peer Reviews

Budnitz, Robert J., George Apostolakis, David M. Boore, Lloyd S. Cluff, Kevin J. Coppersmith, C. Allin Cornell, and Peter A. Morris. 1995. "Recommendations for Probabilistic Seismic Analysis: Guidance on Uncertainty and Use of Experts." NUREG/.CR-6372 and UCRL-ID-122160, Lawrence Livermore National Laboratory, 170pp.

Budnitz, Robert J., George Apostolakis, David M. Boore, Lloyd S. Cluff, Kevin J. Coppersmith, C. Allin Cornell, and Peter A. Morris. August 1998. "The Use of Technical Expert Panels: Applications to Probabilistic Seismic Hazard Analysis." *Risk Analysis* pp 463-470.

Morgan, M. Granger and Max Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, New York, 1990 (reprinted 1998).

Kaplan, S. 1992. "'Expert Information' versus 'Expert Opinions': Another Approach to the Problem of Eliciting/Combining/Using Expert Knowledge in PRA." *Reliability Engineering and System Safety* 35:61-72.

Merkhofer, M.W. 1987. "Quantifying Judgmental Uncertainty: Methodology, Experiences, and Insights." *IEEE Transactions on Systems, Man, and Cybernetics* 17(5):741-752.

Morgan, M.G., and M. Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. New York: Cambridge University Press.

Morgan, M.G., M. Henrion, and S.C. Morris. 1980. "Expert Judgment for Policy Analysis." Brookhaven National Laboratory. BNL 51358.

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

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

Spetzler, C.S., and Stael von Holstein. 1975. "Probability Encoding in Decision Analysis." *Management Science* 22(3).

Appendix A:

Comments from Expert Reviewers

Review of
"An Expert Judgment Assessment of the Concentration-Response Relationship Between $PM_{2.5}$ Exposure and Mortality"

Review prepared by
Dr. Douglas Crawford-Brown
Professor and Director, Carolina Environmental Program
May 10, 2004

I have completed my review of the document "An Expert Judgment Assessment of the Concentration-Response Relationship Between $PM_{2.5}$ Exposure and Mortality" dated April 23, 2004. My comments are below.

1. General Comments

My overall impression is quite positive. I think the authors are selling the study short to some degree by referring to it solely as a pilot study. While I might prefer a slightly larger sample of experts, I believe the study was well conducted and that a larger sample would cause only minor adjustments in the final confidence intervals. I would not have been surprised to find this study being used as the basis for a regulatory benefits analysis. I would position it intermediate between simply a pilot study and a full-blown study for use in assessments, and probably closer to the latter than the formal.

The methodologies of elicitation appear to me quite sound. I know there are problems associated with wording of questions, and the ordering of questions, that can lead to the classical issues of grounding, bias, etc, explored in the literature. I am not, however, convinced that it is possible to do anything more than what the authors already have done to prevent these distortions (most of which lead, as the authors suggest, to underestimates of the confidence intervals). I reviewed the material provided to the subjects, and went through both the questions and the order in which they occurred, and can see no reason to believe that this had any significant effect on the answers given. And I say this despite the fact that the experts clearly decided to approach the final answers in a somewhat different direction than that originally envisioned by the authors. That is to be expected at times, and I prefer the decision by the authors to let the experts go at the issue as they wished, rather than forcing them into an approach (I refer here mainly to the issue of whether minimum and maximum values should be assigned first, followed by progressive inclusion of specific percentiles).

I found it interesting that the experts uniformly preferred to focus on increases in mortality following increases in exposure, rather than decreases following decreases in exposure. Some thought should go into addressing the issue of whether this somewhat subtle change might hide a difficulty that will complicate translating confidence intervals on increases into confidence intervals on decreases. I can't think of any reason why the same confidence intervals

wouldn't apply to both cases, but then I am always surprised at how seemingly insignificant changes in a task can affect the answers in elicitations. So this issue deserves some thought.

The main reason the authors had to resort to the Monte Carlo analyses was the way in which the problem or question was stated: as a single estimate of change in mortality per unit exposure. I don't think experts tend to think of a single point on a C-R curve, but rather of the entire curve. Unless the curve is linear with no threshold, there is no single parameter that will capture all of the considerations going into the expert's understanding of the issue. By looking for a single parameter initially, I think the study forced some of the experts (especially B and C) to try to squeeze their more nuanced understanding into a simpler scheme. My preference would have been to ask for confidence intervals on point estimates of the incremental change in effect along the entire C-R curve (e.g., a 1 unit change in concentration at baseline levels of 8, 10, 20, etc). I know this increases the work of the interview, but it also would decrease the amount of post-processing of the answers.

On this same note, I found the discussion of the treatment of answers for Experts B and C difficult to follow. Part of the reason is that the Analytic Methods material appears before the actual results from the individuals. Not having yet seen the results, I couldn't understand what was being said in Chapter 2. This was in part because I didn't understand at that point why these Monte Carlo manipulations of raw answers were needed. I finally came to understand the issue on the second reading through Chapter 3, when it dawned on me that the benefits analysis would be being asked to assess the effect of a uniform 1 or 10 unit increase (well, decrease in the original questions) across the board in the US, and that this is why the distribution of baseline concentrations needed to be sampled from. But this was not at all clear to me in either Chapter 2 or 3, and I wonder whether it was clear to the experts. At some point, there is a need to state clearly the precise decision formulation, or benefits analysis, these answers are intended to support, and why the Monte Carlo analyses are needed in order to place all 5 responses onto a common footing in regards to this decision/analysis. I also think that the Monte Carlo analysis material could come after the results, rather than in the Analytic Methods chapter, as this would allow the reader to see the results first and then confront the difficulty of placing them onto a common footing. Chapter 2 could focus only on the analysis that produces the figures showing results of individual experts (un-pooled), with the methods to produce the pooled results described only after the individual results are presented. That would also be the point at which the REASONS for needing to place the results on a common footing would be discussed.

While the analysis, or questioning, didn't impose any specific distributional form on the judgments of the experts (i.e., they were not required to use normal distributions to describe their uncertainty), the authors might consider the utility of lognormal distributions in the future. These distributions have the property that they are phrased in terms of "accurate to within a factor of Y." For example, the geometric standard deviation of a lognormal distribution lends itself quite naturally to such an interpretation. Perhaps the experts all preferred to think in terms

of an X percent increase or decrease around some central tendency value, but they might also find it useful to think in terms of multiplicative factors around the central tendency values. Just a thought. I am not convinced it would have made much of a difference.

I very much liked the use of background questions concerning causality, mechanisms, etc, to help the experts frame the problem in their head. I agree with the authors that some form of disaggregated assessment might be useful, although I wouldn't insist that the experts reduce their entire final judgment of confidence intervals to some rigorous combination of these disaggregated judgments. My own experience is that overall judgments should stand in some relation to the disaggregated judgments, but that one can never think of all the factors that might, and should, affect the final judgment, and so the final judgment of the confidence interval can never be reduced to a calculus of the disaggregated judgments. But the disaggregated judgments would still provide useful insights for the experts in forming their final judgments (and I think did to a large extent in the current study). It is important that this process of disaggregation follow closely the lines of reasoning used by scientists, and recognize that different causal theories can lead to differences in the way evidence is used. So, the disaggregated judgments should not reflect only one causal theory.

One of the large issues in such elicitations is whether the expert is anchoring on evidence that supports his or her preconceptions, or is systematically sorting through supporting and counter evidence for a given belief and reaching a judgment balancing this conflicting information. A strong elicitation process causes the expert to not only go through his or her line of reasoning, but to explore the "paths not taken" through the evidence and explain why these were not taken. It is a good feature of the current study, therefore, that experts were asked to consider evidence they had not selected, or had perhaps rejected.

Finally, I am torn on the issue of having the experts give their judgments in isolation. I understand the issues of logistics, and also the problems of dominant personalities swaying the group inappropriately. But I still think it is useful for individuals to confront the lines of reasoning offered by others and to consider whether they want to adjust their beliefs accordingly. This can be done through circulation of at least one round of summaries of each expert's results and reasoning to the entire group. This adds a bit of time, but not much (I would guess another 8 hours per expert to read the other summaries and then decide how he or she wants to adjust the percentiles of their distributions). I agree that looking for a consensus distribution is unwarranted, and prefer the idea (stated in the report) of carrying each distribution forward into the benefits analysis and then post-processing the results from the individual benefits analyses. Of course, this could significantly increase the work of the benefits analysis. But it avoids creating a fictitious individual whose distribution does not correspond to that of any one individual expert.

2. General Topics

Actual questions (in italics) were inserted to aid the reading of this review.

1. *What are "best practices" for conducting expert elicitations? What "best practices" are essential for a defensible elicitation?*

I think this report used reasonable best practices. The selection of experts was quite good (even if sample size could be somewhat larger); the review materials developed were good; the questions were appropriate; and the ways of analyzing these results were appropriate.

2. *How does the EPA's elicitation compare to "best practices"? What are the strengths and weaknesses of this elicitation?*

The methods used here compare well with best practice. But then, I must confess that I am more concerned about the nature of the scientific discourse that goes on during the elicitation process than I am about formal procedures. So, since I found the scientific discourse good here, I am comfortable with the methods of elicitation.

3. *How should the individual expert opinions be combined?*

I prefer the method of following the distributions of individuals through the benefits analysis and only then combining them. This would be done by combining their cumulative distribution functions as was done in the report.

4. *How could the elicitation be improved?*

I discussed this issue in the first section.

5. *Are the elicitation methodology and process appropriate as a benefit analysis technique to characterize uncertainty?*

Yes, I think they are appropriate and suitably combine expert judgment and data analysis (since the experts are asked to relate their confidence intervals to the confidence intervals in the underlying data). I think confronting the experts with results such as Figure 8 (although restricting the left-most bars to the results for an individual rather than the pooled results), and seeing if they want to adjust anything, would be good.

3. Specific Topics

1A. *Did the set of participants chosen reflect the views of other scientists in the field?*

Yes, their views reflect the range of views.

1B. *Was the number of participants appropriate?*

I think it was. A larger number might be slightly better, but I am unconvinced it would change the analysis appreciably in anything other than a statistical sense. A larger sample always gives a better estimate of population parameters, but what one wants here is not the view of the entire scientific community, but the view of suitably trained individuals who have been through the structured process of elicitation. I personally would be comfortable with a sample of 5, given the good process of selecting them.

1C. *Was the method for choosing participants acceptable?*

Yes, it was a very good method. The one thing that bothers me, however, was the decision at one point to allow an expert who chose not to participate to in turn suggest some new individuals, and then re-sample from this larger pool. I think the procedure needs to be established at the beginning and followed rigorously.

1D. *For these participants, does the potential for motivational bias exist?*

This potential is always there, but these individuals have done the best they could.

1E. *Are the relevant fields represented?*

Yes, the relevant fields are represented.

2A. *Were the topics adequately described to the participants (eliminated ambiguity)?*

Yes.

2B. *Were the correct questions asked (i.e., did they effectively address the topics without introducing bias)? Do you think that word choice, structure, or the order of the questions affected the quality of the results?*

I like the questions asked. They were not overly prescriptive, but they caused the experts to consider the most important issues before forming their judgments. I do not believe any different ordering would have affected the results. But then I am always surprised as to how small changes in wording can affect the answers to such questions. The areas of study in which these effects appear, however, are different from this area, and generally are restricted to questions having to do with risk trade-offs (which did not occur in this study).

3A. *Were any materials missing that should have been included in the Briefing Book? Should any materials have been excluded?*

I liked the briefing book. It is always possible to add more studies, but the book contained what I consider to be a representative sample of the most influential studies.

3B. *Was the time allowed for review of the materials adequate?*

Yes, although it was interesting to see that some individuals didn't quite make use of that time!

3C. *Some participants were more prepared than othersCis this a problem? If so, how can this problem be addressed?*

Ideally, everyone should be equally prepared. I do consider this to be at least an issue, and perhaps a problem. If an individual is not prepared, there is too much of a chance that he or she will be unduly influenced by a single result they can call readily to hand and scan during the interview, or by clues from the interviewer. I think this aspect needs to be tightened up a bit.

3D. *Were expectations effectively communicated to the participants prior to the meeting?*

Yes.

3E. *Were the questions sufficiently tested prior to the elicitation?*

I can't really address this. The process described seems reasonable, but I didn't sit in on the actual tests and so don't know what issues arose. Let's just say the final questionnaire contained the questions I would have wanted to see. I can't comment on how the resulting discourse proceeded, although the nature of this discourse (i.e., how well it is guided by someone who knows the field well and can help the expert consider alternatives lines of reasoning) is often as important as the original question in getting a reliable elicitation. In that regard, I like the use of an additional expert to participate at all times in the discussions.

3F. *Was adequate training provided for the participants prior to the elicitation? Would a pre-elicitation workshop have been helpful?*

I think a pre-elicitation workshop is important. This places all participants on a common footing in regards to the evidence; allows them to see the competing lines of reasoning in the literature; and would allow them to normalize their judgments of uncertainty by sending them through some common, and well-established, examples of uncertainty analysis. I recommend conducting one in the future.

4A. *How do individual interviews compare to group interviews?*

I think the use of individual interviews was appropriate. It is extremely hard to document group interviews and to understand the dynamics of the discourse.

4B. *Is it important to achieve consensus or not?*

It is not important to achieve consensus, and I would even avoid looking for it. I think you will get an artificial result.

4C. *Did the interviewers perform their roles properly?*

Yes.

4D. *Was the length of the interview appropriate?*

Yes.

5A. *How should the quantitative and qualitative results from the individual interviews be combined?*

I discussed this in the first section. I think it is appropriate to combine the results IF one assumes it is not possible to send each individual distribution through the benefits analysis. But as far as presenting pooled results from this particular elicitation, I believe there may be a slight problem. I say "may" because there is no example of the procedure applied. I think it was as follows: (1) create the cumulative distribution function, using specific percentiles, for each individual, (2) for each individual, determine the parameter value associated with a specific percentile, such as the 75th, and repeat for each individual, (3) take the weighted average of these 5 estimates of the parameter value, with the weighting being 0.2, (4) repeat over all percentiles of interest.

This is not the method I would have used. The correct method is to take the weighted average of percentiles associated with a given parameter value, rather than (as in the method described above) the weighted average of parameter values associated with a given percentile. For example, imagine there are two individuals, each supplying a cumulative distribution function for parameter X that must lie somewhere between 0 and 1. For individual A, the percentiles of the CDF (shown in parentheses) for specific estimates of X are:

0 (0); 0.2 (0.3); 0.4 (0.5); 0.6 (0.9); 0.8 (1.0); 1 (1.0)

For individual B:

0 (0); 0.2 (0.1); 0.4 (0.3); 0.6 (0.4); 0.8 (0.9); 1 (1.0)

If the two individuals are weighted equally, the composite CDF is:

0 ([0+0]/2); 0.2 ([0.3+0.1]/2); 0.4 ([0.5+0.3]/2); 0.6 ([0.9+0.4]/2); 0.8 ([1.0+0.9]/2); 1 ([1.0+1.0]/2)

or

0 (0); 0.2 (0.2); 0.4 (0.4); 0.6 (0.65); 0.8 (0.95); 1 (1.0)

But the procedure (I think) used to construct the CDF in the report would start from the specific percentiles and calculate the average of the parameter values at those percentiles. The resulting CDF would be different from the one above. Of course, the advantage of the method used in the report is that it can be based solely on the numerical values elicited from the respondents, who were asked to give the numerical values of X corresponding to a given percentile of the CDF, rather than (as required by the more exact method) the percentiles of the CDF corresponding to specific numerical values of X. But the more exact procedure could be done by plotting the results of the elicitation for an individual as a CDF, and using the smoothed curve of the CDF to estimate the percentiles of the CDF corresponding to each specific value of X. The method actually used is only an approximation, which can introduce significant errors at times.

5B. *Was it appropriate to weight all the opinions equally? Are there alternative weighting systems that could have been used? EPA considered ways to calibrate the experts, which can be used to weight the experts' responses, but could not construct an applicable calibration question for this topic. Do you agree that you cannot calibrate experts on this topic? If not, do you have ideas for calibration questions.*

I do not like calibration of experts, or differential weighting. I think the appropriate approach was used: establish a procedure to get the proper sample of people in the study, and then treat them all equally. In other words, do your weighting up front in choosing participants, not in any post-processing of their individual results.

5C. *Can the results of the individual interviews reasonably be combined given the differences in forms of the functions? Of the four methods presented for combining the results, are any of them sufficient given the statistical approaches taken, and would you recommend other methods for combining the results?*

Yes, I think they can be combined, and am reasonably comfortable with the methods used. Part of the problem lies in looking for a single value (e.g., a 1 unit increase above a baseline) rather than asking the participants to construct values at different baselines. I would prefer the latter approach. I don't know what is meant by the "four methods." I can't see where 4 alternatives are presented systematically at any point in the text. So, sorry, but I can't answer this part of the question.

5D. *Are all of the essential elements included in the final report? Are any unnecessary elements included?*

I think the report is very good, and the only recommendations I have for re-structuring were described in the first section of my review.

6. *Recommendations for future elicitations: Absolute necessities for an expert elicitation (based on best practices) and Major strengths and weaknesses of this expert elicitation*

   These issues were addressed in my general comments at the beginning.

4. Some Additional Specific Comments

1. I found the issue of separating the long-term from short-term effects somewhat confusing. This doesn't mean the experts also found it confusing, but at least the material I reviewed didn't adequately explain this issue, why it was important, how it relates to interpretation of specific studies, etc. And I don't see any discussion of long-term exposures increasing susceptibility to short-term exposures, or how experts might consider this in their reasoning. I am confident the experts did so (given their backgrounds, and the fact that they are all very thoughtful scientists), but it would have been better to make this more explicit.

2. Bullet 3 on Page 4 calls for a balanced pool covering the range of views. This is a good general principle, but it is important not to let the issue of a full range unduly influence the sample, providing over-representation of people at the extremes.

3. One weakness of the use of individual elicitations without sharing of answers (as in a Delphi method) is that the procedure fails in some sense to mimic the procedures of science, where discourse and comparison of views is such a vital activity that informs the rationality of judgments.

4. Figure 1 was a mystery to me. It needs a much better explanation and figure heading.

5. One additional question should focus on fraction of time spent indoors and outdoors. If high levels of ambient pollution force people indoors, and if the indoor air is a major contributor to effects (as many scientists believe), this issue will be significant.

6. Figure 3 is not well described. It needs a more complete heading.

7. The conclusion in the last sentence of the second full paragraph on page 70 does not seem warranted to me. It is not clear the impact on benefits analyses will be so significant.

Review of
"An Expert Judgment Assessment of the Concentration-Response Relationship Between $PM_{2.5}$ Exposure and Mortality"

Review prepared by
Dr. H. Christopher Frey
Associate Professor, Department of Engineering, North Carolina State University
July 6, 2004

1. Introduction

This is a review of the following document:

An Expert Judgment Assessment of the Concentration-Response Relationship Between $PM_{2.5}$ Exposure and Mortality, Prepared by Industrial Economics, Incorporated under subcontract to Abt Associates, Inc., for U.S. Environmental Protection Agency, Research Triangle Park, NC, April 23, 2004.

This review is organized in response to several key questions given in the charge. This review addresses those topics that are within the expertise of the reviewer. Therefore, not all charge questions are addressed here. For example, since this reviewer is not an expert in the domain subject matter, it was not deemed to productive to offer comments as to whether there might have been other topics that should have been discussed during the conditioning step.

2. General Topics

This section covers the following major issues:

- What are "best practices" for conducting expert elicitation? What "best practices" are essential for a defensible elicitation?
- How Does the EPA's Elicitation Compare to Best Practices? What are the strengths and weaknesses of the elicitation?
- How should the individual expert opinions be combined?
- How could the elicitation be improved?
- Are the elicitation methodology and process appropriate as a benefit analysis technique to characterize uncertainty

2.1 Best Practices for Conducting Expert Elicitations

The discussion of best practices given here focuses on the expert elicitation protocol and related issues. This is based upon a working paper written by the reviewer that is also the basis for materials used in a doctoral level course, CE/NE 772 Environmental Exposure and Risk Analysis.

There are several protocols that have been developed for eliciting expert judgments about uncertainty. One of the most widely reported protocols is one developed in the 1960s and 1970s at Stanford and the Stanford Research Institute (Spetzler and von Holstein, 1975). The Stanford/SRI protocol involves five steps. Similar protocols have been developed by others.

We will describe the basic steps in an expert elicitation protocol. We assume that there is an elicitor who implements the protocol, and an expert whose judgment is sought. The discussion here is based on the five steps of the Stanford/SRI protocol. Each of the five steps is discussed in the following sections. The material presented here is based on Morgan and Henrion (1990), Morgan, Henrion, and Morris (1980), and Merkhofer (1987).

2.1.1 Step 1: Motivating the Subject: Establishing Rapport

The purpose of this step is for the elicitor to establish rapport with the expert. In preparation for this discussion, it is useful for the elicitor to have some knowledge of the subject matter and for the interview to be located in a place where the expert has full access to relevant materials (e.g., in the expert's office).

At first, the elicitor generally explains to the expert the nature of the problem at hand and the analysis being conducted. The purpose of this discussion is to give the expert some context regarding how their judgments will be used.

A key part of this step is a discussion of the methodology for probabilistic assessment and expert judgments. Such a discussion has several benefits: it helps the expert understand why the elicitor is approaching the elicitation in the format employed; it indicates to the expert that the elicitor is trying to do a careful job of the elicitation; and it satisfies professional obligations to inform the expert of the potential problems and limitations of expert elicitation. The discussion may include an explanation of the types of heuristics that experts may use when making judgments, and how the elicitation protocol is structured to help the expert make the best use of information.

Another component of the motivation phase is to identify any motivational biases. For example, the subject's personal involvement with corporate or institutional positions regarding the uncertain quantity should be discussed. If such biases are found, it may be possible to overcome them by disaggregating the problem or restructuring the problem in such a way that bias is less likely to be of concern. One approach is simply to take notes, which may encourage the expert to think about other points of view besides their own in an attempt to appear balanced.

2.1.2 Step 2: Structuring: Defining the Uncertain Quantity

Structuring typically consists of four steps: defining the variable; identifying the possible range of outcomes; disaggregating if necessary; and selecting an appropriate measurement scale.

At the outside, the elicitor seeks is to arrive at an unambiguous definition of the quantity to be assessed. The definition should pass the "clairvoyance test"—the definition should be specific enough that a clairvoyant could look into the future (or the past, as appropriate) to discover the exact value of the quantity. The definition should also be stated in form in which the expert will most likely be able to provide reliable judgments. For example, the elicitor should work with whatever units with which the expert is most comfortable. The expert should also begin to think about the full range of possible outcomes for the variable.

A critical aspect of structuring is to determine whether there are conditioning factors that may influence the value of the quantity. If so, then these conditioning factors need to be clearly stated. In the case of an air quality model, it is necessary for the expert to have a clear idea of the scenario (e.g., time frame, geographic domain) upon which the estimate is predicated. It is also important for the elicitor to obtain from the expert a list of all key assumptions that the expert is making. If the expert is assuming, for example, that no significant changes in emissions sources occurred during the episode or that there were no major shifts in human activity patterns, these assumptions must be known to the elicitor.

It is also important to identify, to the extent possible, quantities whose uncertainties are statistically independent from uncertainties in other variables. Sometimes disaggregating the variable into more elementary variables can help. Judgments would then be elicited for these more elementary variables.

2.1.3 Step 3: Conditioning: Get the Expert to Think About All Evidence

The purpose of this step is get the expert to think about all relevant knowledge related to the uncertain variable. This includes thinking about available data, theoretical models for how the system of interest behaves, and how the expert plans to use the information. The expert typically will consider "case-specific" information, which relates directly to the quantity being assessed, and "surrogate" information, which relates to quantities similar to that being assessed. In the later case, the expert may draw on analogies with similar systems to make inferences about possible outcomes for the quantity of interest.

The elicitor needs to help the expert think about the problem from different perspectives, to help draw in as much relevant information as possible. One approach is to ask the expert to react to various scenarios constructed by the elicitor ("suppose the concentration at this specific location and time turned out to be X. Why would this occur?") or to ask the expert to invent such scenarios for extreme outcomes. Another approach is to restate the question in different forms.

As part of conditioning, some investigators suggest discussing with the expert heuristics used in making judgments and their potential for creating biases (e.g., Merkhofer, 1987). Others recommend doing this as part of the motivating stage of the elicitation (e.g., Morgan and

Henrion, 1990). In either approach, part of the purpose of the conditioning phase is to attempt to overcome biases such as anchoring and adjustment, and availability, by getting the expert to think about a range of outcomes without fixating on any particular one.

2.1.4 Step 4: Encoding: Quantifying the Expert's Judgment

The purpose of the encoding phase is to arrive at a quantitative description of the subjective probability distribution which best reflects the expert's beliefs about the possible range of outcomes, and their likelihood, for the uncertain quantity. The expert's judgment is for the quantity which emerged in the structuring phase, using the knowledge base discussed in the conditioning phase.

As a deliberate attempt to counteract the availability and anchoring and adjustment biases (and the resulting tendency toward overconfidence observed in many studies),many elicitors recommend that the encoding start with extreme upper and lower values of the distribution. Thus, one might first ask the expert for what they consider to be extreme values for the quantity. Then the elicitor would ask for scenarios that might lead to outcomes outside of these extremes. In addition, the expert might ask the expert to pretend that we had just learned that the value of the quantity was actually just outside the extremes previously discussed. Can the expert invent any plausible explanation for such an outcome? If so, then the expert may revise their estimate of the extreme values.

There are several techniques for eliciting probability distributions for continuous random variables. They fall under several categories. A few of the most important ones are briefly described:

- **Fixed Value Methods.** In this approach, the expert is asked to estimate the probability that the actual value of a quantity is higher (or lower) than some arbitrary number. For example, what is the probability that the VOC emissions from dry cleaners in an urban grid cell on a summer weekday could be 20 percent higher than the point estimate used in the emissions inventory? This type of elicitation may be done with the aid of a probability wheel. The probability wheel is a graphical tool for communicating to an expert the meaning of probability.

- **Fixed Probability Methods.** Here, the expert is asked to estimate the value of a quantity such that the probability of higher or lower values is some specified amount. For example, what is the drycleaner VOC emission rate such that there is only a 10 percent change of a higher rate?

- **Interval Method.** These methods involve partitioning the probability distribution into ranges of equal probability. For example, to assess the median of a distribution, the elicitor may ask the expert to react to an arbitrary value. The value is adjusted until the expert is indifferent as to whether the actual value of the quantity is higher or lower. Then, to assess the quartiles (25 and 75 percentiles), the expert is asked

whether it is equally likely that the value is within the interval bounded by the extreme or the median.

- **Reference Lottery.** The expert is asked to choose between two bets. One is a reference lottery, in which the probability of winning can be adjusted. The other is whether the actual value of a quantity will be above or below some specified value. The probability of the reference lottery is adjusted until the expert is indifferent between the two best.

The fixed value and fixed probability methods are examples of *direct* methods, in which the expert must respond to questions with numbers. The interval and reference lottery methods are *indirect* approaches, in which the expert has to make choices between alternatives but does not have to specify numbers. The fixed value methods appear to have found favor among many elicitors. The indirect methods often appear to be frustrating for experts with quantitative backgrounds, who may be more comfortable responding with numbers. Thus, the selection of an appropriate approach depends in part on the preferences of the expert.

It is generally preferred that the quantities for which judgments are sought be statistically independent of each other. However, in cases where it is unavoidable, approaches may be used to elicit judgments for two dependent uncertain quantities. One approach is to assess a probability distribution for one variable, and then to assess several probability distributions for the second variable conditioned on selected values of the first variable. An approach often used instead is to arbitrarily categorize a correlation as being high, medium, or low, and then to use a nominal correlation coefficient for each category (e.g., 0.9, 0.6, 0.3). However, it is often difficult to make judgments regarding correlations in this manner.

2.1.5 Step 5: Verifying: Checking the Answer

The purpose of this phase is to test the probability distribution constructed in the encoding phases against the expert's beliefs to make sure that the distribution correctly represents those beliefs. The elicitor can form what appear to be equally likely outcomes based on the elicited distribution and ask the expert if they would be willing to make bets that one or the other outcomes are more likely to occur. If the distribution correctly reflects the expert's beliefs, then the expert should be indifferent between such bets. If not, then it may be necessary to iterate on previous steps, including conditioning and encoding, to obtain a better estimate of the distribution.

The results can also be plotted as a cdf or a pdf and shown to the expert. Any key features of the distribution, such as the location of the mode, the presence of more than one mode, or extreme skewness, should be reviewed and discussed with the expert.

Expert Groups vs. Individuals

Our discussions so far have focused on the case in which an elicitor is working with a single expert. However, there may be cases, for practical reasons, in which an elicitor may have to work with groups of experts. In such situations, group interactions may interfere with the purpose of obtaining a subjective probability distribution that represents beliefs about uncertainty. For example, face-to-face interactions among group members can be counterproductive if the group is dominated by particular individuals because of personality traits. Similarly, the group can be counter productive if it is susceptible to managerial biases. If the boss is part of the group, then the group members may tend to offer statements about goals and desirable outcomes in lieu of statements regarding their beliefs about the possible outcomes for the variable.

Multiple Experts. One method to avoid the downsides of group interactions is simply to avoid using groups. Instead, elicitations are performed with multiple experts on an individual basis. The resulting judgments can be considered separately to determine whether any significant difference in modeling results occur. Alternatively, the judgments can be aggregated. Some investigators have suggested that simply combining the judgments with equal weights often performs as well as more complex schemes in which different weights are assigned to different experts. The assignment of such weights can be based on approaches in which the experts are asked, in an anonymous forum, to rate each other and themselves. The weights may also be based on the judgment of a decision-maker. Any assumptions regarding weights are often viewed as tenuous or controversial. Therefore, sensitivity analysis on the weights is important to determine whether the choice of weights significantly affects the answer. If the degree of sensitivity is very high, then differences of opinion should be considered more explicitly.

Group Interactions. A key benefit of group interactions is the sharing of knowledge. Thus, one approach to dealing with groups is to use them for the motivating, structuring and conditioning phases of an elicitation, and then to separate the group and ask individuals to make judgments in the encoding and verification phases. In this approach, the group is used to bring to discussion a potentially larger body of evidence than any single individual possesses. The elicitation process may be tied back to the group through the use of Delphi techniques, in which anonymous encoded subjective probability distributions, together with supporting statements regarding reasons and rationales for the judgments, are circulated back to the group for reassessment. Interchange of these assessments and explanations may lead to some convergence of opinion. The Delphi approach is used to avoid personal interactions that might be deleterious to a frank exchange of opinions.

Expert Information from Groups. Alternatively, an analyst may construct a distribution based on the information input of the group, without eliciting the judgments of

individuals. In this approach, which is suggested by Kaplan (1992), the group as viewed as a source of expert information, as opposed to a source of expert opinion. This approach aims at obtaining a consensus regarding the body of evidence relevant to estimating uncertainty about a quantity. The elicitor may choose to iterate with the group regarding a consensual distribution, but this step is optional. Ideally, the range of uncertainty in the distribution emerging from the group deliberations would be sufficiently broad to capture all outcomes that any of the groups members believe could occur. In practice, it is not clear how this can be implemented without introducing significant motivational biases.

Who Should Do the Elicitation and How Long Does it Take?

The expert and the elicitor should be two different people. It is desirable for the elicitor to have some background or training in the theory and practice of elicitation. The elicitor also should develop a working knowledge of the subject matter for which judgments are sought, to enable communication with the expert and the development of questions to help condition the expert. The process of expert elicitation is not amenable to cookbook solutions. Each elicitation problem is a special case, because experts have different styles and preferences. The expert often helps shape the definition of the quantity which is the subject of the elicitation.

An elicitation may take anywhere from a half hour to a day, in most cases. The elicitations should be conducted in a setting comfortable for the expert and which allows maximum access to relevant files, books, data, etc. The expert should be provided with sufficient information in advance, in the form of groundrules or a briefing packet, regarding the technical assumptions and design basis upon which judgments should be based, and the methodology for the elicitation.

2.2 How Does the EPA's Elicitation Compare to Best Practices? What are the strengths and weaknesses of the elicitation?

The approach described in the report for EPA generally conforms to good practice. The strengths of the elicitation are that:

- multiple experts who are recognized in their fields were subjects,
- the process generally followed the steps of established elicitation protocols,
- elicitations were conducted individually with each expert,
- the conditioning step was conducted with each expert and seemed effective at getting experts to think about relevant information prior to the encoding step. The briefing material that contained a list of questions that encouraged experts to think about relevant evidence is a useful approach.
- The authors were able to identify some areas of consensus among the experts, such as agreement that the original Six-Cities and ACS studies were well-conducted.

Agreement regarding sources of information can be very useful even if the expert inferences and opinions based upon them differ. In such cases, the differences can be attributed to other factors, which helps focus debate/discussion in the future

- the verification approach was reasonable in terms of general checks on the procedures and evaluation of answers; however, it is not clear if the experts had a chance to review the final distributions and models developed by the study team based upon each expert's judgment. If so, then the procedure is generally good. If not, then the lack of the final verification step could be a substantial weakness.
- The follow-up questions, and particularly those that challenge the expert to consider whether higher or lower values than what they had provided are possible, are important.
- There was good discussion on some methodological points, such as aggregated versus disaggregated judgments, and there was acknowledgement of some potential limitations.
- the process is well-documented. This is especially a strength in terms of detail regarding the basis for the experts' judgments. The study team did a good job of recording information from the experts and explaining the basis for the experts' judgments.

The weaknesses are:

- the manner by which multiple judgments were combined,
- there was no knowledge sharing between experts;
- there should be a systematic discussion of known heuristics and other sources of bias and how each one was addressed in the study. There is some of this scattered throughout the report, but having it in one place would make it easier to evaluate the method used.
- how the encoding process was implement - it appears that some experts provided judgments regarding the central tendency before providing judgments regarding upper and lower ranges. This type of sequence of judgments can involve the well-known "anchoring and adjustment" heuristic, which is associated with biased (overconfident) estimates of uncertainty.
- The use of a fixed percentile encoding method could be a weakness if this method is not well-suited to a particular expert. The effectiveness and accuracy of the fixed percentile method can depend on the specific manner in which it is implemented. For example, one extreme is just to ask the expert for the values associated with the percentiles. Another is to use a graphical tool (e.g., probability wheel) or other technique to assist in conveying what each percentile represents. The order in which the questions are asked is also important.

- There was only one execution of the process. It would be useful to share the knowledge among the experts (but not necessarily their judgments) to see if any experts would wish to refine their judgment in light of evidence or theoretical inferences that they might not originally have given much consideration.

The selection process for the experts was very good. It was reasonable to start with a pool of experts that had in a sense been pre-screened by the National Research Council and to invite a few other experts to participate in the process of selecting experts. It was also appropriate to select experts so as to cover a range of relevant disciplines. Although the study might have included more experts, the use of five experts is a good start and may be adequate. The project team should review the practice in probabilistic risk assessment in the nuclear industry with regard to how many experts to include as a perhaps useful precedent. Of course, it is possible that in a health-related field that may be more multidisciplinary that there could be a need for a larger number of experts than for a narrowly defined subcomponent of an engineered process.

The use of mathematical approaches for combining judgments is not a major shortcoming, but the specific approach used seems to be. The text (e.g., on page 14) describes the averaging as follows: "the equal weight combination method we used involves averaging responses across experts for each percentile and for the minimum and maximum values elicited." The text goes on to the secondary topic of whether there is dependence in the judgments and argues that expert responses were treated as if they were completely independent. Of course, this is an incorrect statement if the averaging was done for specific percentiles (e.g., 1st percentile, 5th percentile, 10th percentile, and so on to the 90th, 95th, and 99th percentiles) because this type of averaging assumes 100 percent correlation between the experts' distributions. That is, the assumption in this type of averaging is that if Expert A is predicting a high (on a relative basis) value of the quantity, then Expert B is simultaneously predicting a high (on a relative basis) value of the quantity. Clearly, this is unlikely to be the case unless both experts agree on the mechanisms that are causing the high end of the outcomes, even if they might disagree on the absolute value of the high outcomes.

A key implication of the type of averaging that is described in the report is that the combined distribution used as an input to the analysis may be one to which none of the experts could agree, even in part. For example, Figure 1 illustrates what happens if two distributions from hypothetical Experts A and B (shown individually in the panel on the left) are "averaged" using the technique described on page 14 of the report. The result is the center panel. Note that the domain of the "average" distribution lies in an area for which neither expert provided judgments. Thus, neither expert would agree that the averaged distribution is a fair representation, even in part, of their judgment. Instead, the two distributions could be weighted as part of a mixture of distributions, as shown in the panel on the right side of the figure. A

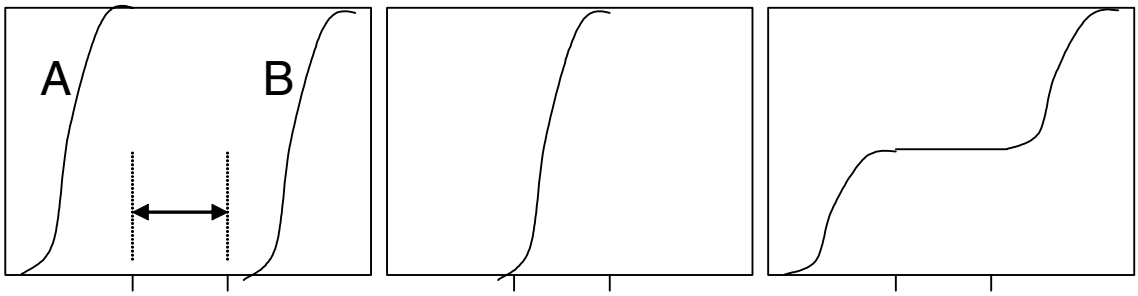


Figure 1. Comparison of Methods for Combining Judgments: (a) Individual Judgments of two Experts, A and B; (b) "average" of the percentiles of the two judgments; (c) equally weighted mixture of the two judgments. The vertical axis is cumulative probability and the horizontal axis is the value of quantity for which the elicitation is being done.

mixture of distributions is comprised of component distributions. Each component is recognizable as a distribution obtained from an individual expert. Furthermore, the mixture approach has the advantage of including the entire range of opinions, whereas the "average" approach has the unfortunate consequence of dispensing with very low and very high outcomes given by the experts.

In practice, the use of averaging may or may not cause as severe of a problem as illustrated in Figure 1, but the potential for this type of problem is always there in some form if averaging as described on page 14 is used. For example, Figure 4 on page 43 of the report compares the "combined" distribution to the individual distributions. Since all experts included outcomes that were very low, the combined distributions seems to adequately capture the opinion across all experts that there could be very small increases in mortality. However, the averaging seems to negate that some experts implied that only small percentage increases were possible (e.g., Expert C) whereas another expert (Expert E) considered the possibility of much greater percentage increases than any other expert. It is somewhat difficult to evaluate the combined and individual distributions based only on modified Tukey plots such as shown in Figure 4. It would have been more helpful to see the CDF of each distribution and of the combined distribution.

The sensitivity analysis shown on Figure 7 is useful and reinforces the point that the average process seems to average out the more extreme experts, since the results are more sensitive to the removal of the more extreme experts (C and E) than to any other expert. By "extreme" in this case the intent is only to distinguish the range and central tendency versus that of other experts, and is not to suggest anything negative about the opinion of any of the experts.

The shortcomings of the averaging approach are more pronounced in Figure 10 on page 61. In this case, the averaging process produces one overall distribution that does not distinguish (as a mixture distribution would) the differences in opinions among the experts nor does it seem

representative of any of the experts. Clearly, Experts B, C, and D produced smaller averages and narrower ranges than the combined distribution, whereas Experts A and E produced higher averages and wider ranges than the combined distribution. Thus, one has to question whether any of the experts could agree with the combined distribution or feel that their opinion was adequately included.

The fact that a combined distribution is more symmetrical than the individual distributions upon which it is based (as noted in the text) can arise simply as a function of the process of averaging (e.g., Central Limit Theorem). When adding many distributions, the sum will tend to approach normality. Of course, the manner in which distributions are added does not typically assume 100 percent correlation, as seems to be the case in the report.

The "combined" distributions developed in this work were compared to published studies and were evaluated for reasonableness. The comments above are mainly aimed at the point that the combined distribution does not adequately capture the opinion of individual experts, but averages it out.

2.3 How Should the Individual Expert Opinions Be Combined?

This point is addressed in the previous section.

2.4 How Could the Elicitation Be Improved?

The most important improvements are with respect to the encoding step. This improvement should be in context. In general, the study team did a good job of picking experts and of doing the motivating, conditioning, and structuring steps. The study team acknowledges that a disaggregated approach could be better, and of course this could be an improvement in the future. The study team did accommodate some different structures among the experts. However, the specifics of how the encoding was done could be more clear in the main report. As mentioned in Section 2.2, there should be a critical discussion of all potential sources of bias and of how the elicitation protocol attempts to deal with each one. Although heuristics are briefly mentioned, they are not adequately discussed. In particular, the anchoring and adjustment heuristic possibly could have been a problem in some of the elicitations. This does not seem to get attention until near the end of the report.

The term "elicitation" may be used differently by EPA versus my interpretation. I interpret "elicitation" to refer to the entire process of obtaining expert judgments, but not to the process of combining the judgments into a weighted distribution of some type. A key area in need of improvement is the method for combining expert judgments, which is a post-elicitation exercise. As noted in Section 2.2, the use of an arithmetic average can produce a result to which no expert would agree and does not capture the full range of values produced by the expert. Therefore, the use of multiple expert judgments should be in such a way so that the entire range of opinion is captured. A mixture distribution approach is recommended for this purpose. The

mixture approach is easy to implement in a Monte Carlo simulation framework. If there are e experts, and if there are n samples generated in the Monte Carlo simulation, then on average n/ e of the samples should be drawn from each of the experts' distributions. For example, if there are 5 experts and 100 samples, then on average 20 samples should be drawn from each of the five distributions corresponding to each expert. This will create a mixture of the five distributions.

Other improvements that could be made in the future would be to include an information-sharing workshop in order to help with the conditioning step. However, the expert judgments should be elicited on an individual basis to avoid motivational or other biases (see discussions in Section 2.1). A similar improvement would be to have a post-elicitation meeting or workshop in order to reconsider the condition step and as a precursor to providing an opportunity for experts, on an individual and private basis, to refine their judgments based upon information that might be new to them. If a face-to-face meeting was not practical, then information sharing could be achieved by circulating draft materials such as a summary of qualitative expert responses as given in Appendix C of the report.

2.5 Are the Elicitation Methodology and Process Appropriate as a Benefit Analysis Technique to Characterize Uncertainty

Yes. Suggestions for improvement are given in the previous sections.

3. Specific Comments

These are some specific comments based upon a detailed review of the document.

On page 2 there is mention that OMB staff scientists were involved in the "Project Team." Since a role of OMB is to provide management and oversight of regulations proposed by other federal agencies, a question may be whether it is appropriate for OMB to be part of a study team of an agency for which it provides oversight. From a scientific perspective, there is no problem with scientists from multiple agencies collaborating, and in general this should be encouraged. However, OMB is not typically thought of as a science agency but rather as an agency that provides review and oversight. This is more of a policy than a technical point. Perhaps there was some special purpose to OMB's role in this case, such as to establish new inter-agency procedures. If so, this could be explained.

Page 3. There could be some discussion of how the number of experts is chosen in the field of nuclear power plant probabilistic risk assessment. There is a track record in the nuclear industry of using expert elicitation based upon multiple experts. Some literature could be cited here.

Page 4. While the caveat that the pool of experts may not be "fully representative of scientific opinions" may be true, the way this is written does not convey whether the project team really thinks this is a problem or not. Is there some known lack of coverage as far as

expertise or opinion that would have been desirable to include? In the conclusions (page 71) it is stated that a strength of the analysis was that the expert group "represented a reasonable range of expertise and opinion." Thus, some context and judgment regarding whether this is really a problem could be included in the discussion on Page 4.

Related to this point, the concerns of the SAB as summarized on pages 71-72 with regard to the number of experts could be more specifically articulated, if possible. Are there some important views or minority views that are known to be missing? Or is the concern expressed by the SAB just a generic one? As a practical matter, it becomes more expensive to conduct expert elicitations rigorously with each expert as the number of experts increases, since some of the cost is a linear function of how many experts are involved. Of course, it is important to select from a pool of true "experts" and to cover a range of opinions. The peer review process is also a way of getting feedback on whether some opinion is missing from the pool of judgments. The study team might build in some time and resources to include one or two additional experts after a round of peer review of a draft report (such as this) in case any gaps in important or minority views are identified by reviewers.

On page 5 it would help to have some context as to which criteria were the most binding as far as the number of experts included in the pool (e.g., qualifications or composition of the panel).

Page 9. The time period implied by "long-term" could be more clearly stated.

Page 9. The order in which percentiles were elicited is important and should be described. If the order varied among the experts, then the specific order used with each expert should be documented in the Appendices and should be summarized somewhere in the main report.

Page 10-11. The pilot testing was a good component of the study.

A minor comment is that some of the writing on these pages is too conversational, but that is a stylistic and not a technical issue.

Page 12. As noted, the use of workshops to aid in the condition step is desirable. The reason given ("project schedule precluded it") seems weak, given that the pilot occurred over a four month period, prior to the actual elicitations. It would have seemed possible to have a one day workshop sometime during the project.

Page 13. How was anchoring and adjustment avoided? This should be discussed.

Page 13. Were the experts allowed to review and comment on the combined distributions?

Page 14. The choice of equal weights seems like a practical approach, but the method for doing the weighting is not recommended as discussed in previous sections. As a sensitivity analysis, the weights could be varied or 100 percent weight could be assigned to each of the most extreme cases to get some idea of how sensitive results are to differences in weights.

Page 14. The averaging of percentiles actually assumes 100 percent correlation between experts. Thus, the discussion of dependence here needs to be changed. Of course, the use of averaging is not a preferred approach, as discussed in previous sections.

Some specific comments on the recommendations are as follows:

- "Experts need to be convinced that the elicitation protocol is asking the right questions." This could be stated that the elicitor and expert should participated in a structuring step so that the expert is comfortable that the right questions are being addressed. In turn, this implies that the expert should be allowed to change the structure of the model and to disaggregate as appropriate so that the expert can provide judgments for quantities with which they have better direct knowledge.
- The briefing material should include a literature review, and not just a compilation of literature.
- As noted, a workshop could be used to address many of the motivating and conditioning aspects, and perhaps some of the structuring aspects, of the elicitation. Overall, this might save time or allow for more hours of encoding when meeting on an individual basis with the experts. However, care should be taken not to let the experts state any positions regarding ranges of uncertainty, so as to avoid the perception or reality of motivational biases creeping into the encoding more than they might otherwise.
- The comments offered in previous sections of this review should be considered and incorporated as appropriate.

Review of
"An Expert Judgment Assessment of the Concentration-Response Relationship Between $PM_{2.5}$ Exposure and Mortality"

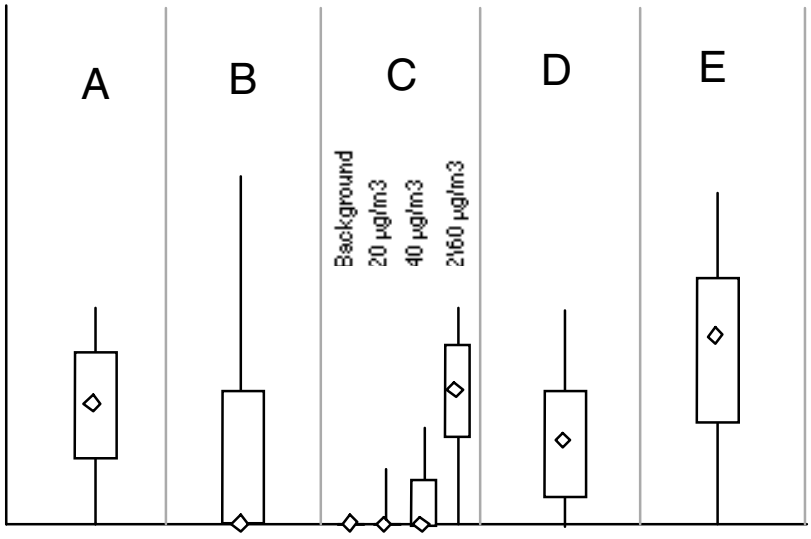
Review prepared by
Dr. M. Granger Morgan
University and Lord Chair Professor in Engineering; Professor and Department Head, Engineering and Public Policy, Professor, Electrical and Computer Engineering and the H. John Heinz III School of Public Policy and Management.
May 22, 2004

My overall evaluation is that this is a high quality piece of work which reflects thoughtful planning and execution. The investigators have invested considerable thought and care in developing the protocol and have displayed admirable flexibility in its application. The effort lays an excellent foundation for future work by EPA/OAQPS in this area.

The staged process which gets experts thinking seriously about the issues before performing the actual quantitative elicitation is good. The discussion of the need in future studies for a more systematic review of evidence, in easy to use summary form, is sensible.

The authors are to be commended for being flexible and allowing the experts to adopt a variety of different models so as to reflect how each frames and thinks about the problem.

In displaying the results, since expert C gave a more complex answer which results in multiple results, a slightly different display might be in order so as to help casual readers to recognize that there are only five responses, not eight. Perhaps something like:



The conclusion (p69) that a more detailed focus on toxicity of specific constituents is probably premature, given the current level of knowledge, seems reasonable.

The concerns that apparently others have raised about the selection of experts are not serious. Clearly it would be nice to have more experts, but the basic strategy of trying to develop a sample across the principle views that exist in the field is the right approach. This is *not* a matter of sampling from a distribution. One of these folks may be right and the rest all wrong. The key point is to get all the main streams of thought represented. If a pre-elicitation workshop is held in the future it might be possible to get the experts to place themselves and each other in some taxonomy of the field so that you could show coverage.

The idea of running a pre-elicitation workshop is a good one. The elicitation questions should still be open for modification and refinement as part of that process.

While I like the caution against using the combined results, and the stress on using each expert's judgments separately in subsequent analysis, that advice may not always get followed. In view of that I am a bit concerned that the algorithm being used to combine experts may not adequately capture the tails when only one or two expert's have wider distributions. Consider Figure 4 (pg 43). Fully 1/3 of expert E's probability appears to lie above the upper 95th percentile on the combined result - yet if I only had the combined result I'd never even know this. I urge the authors to think a bit more about this.

Two thoughts with respect to next steps. First the authors might consider including some questions about future research needs. For example of elicitations which have done this see:

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

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

Questions about research may yield useful insights, and also have the practical benefit of reminding EPA that expert elicitation is not a cheap substitute for doing the needed science.

The second thought involves the basic procedure. While individual expert elicitation is a fine strategy, the authors might also consider the collective expert-workshop strategy that has been adopted by the seismic risk community. For details on this see;

Robert J. Budnitz, George Apostolakis, David M. Boore, Lloyd S. Cluff, Kevin J. Coppersmith. C. Allin Cornell, and Peter A. Morris. 1995. "Recommendations for Probabilistic Seismic Analysis: Guidance on Uncertainty and Use of Experts." NUREG/.CR-6372 and UCRL-ID-122160, Lawrence Livermore National Laboratory, 170pp.

Robert J. Budnitz, George Apostolakis, David M. Boore, Lloyd S. Cluff, Kevin J. Coppersmith. C. Allin Cornell, and Peter A. Morris. August 1998. "The Use of Technical Expert Panels: Applications to Probabilisitic Seismic Hazard Analysis." *Risk Analysis*, pp 463-470.

Review of
"An Expert Judgment Assessment of the Concentration-Response Relationship Between $PM_{2.5}$ Exposure and Mortality"

Review prepared by
Dr. David Stieb
Adjunct Professor, Department of Community Medicine and Epidemiology, University of Ottawa
Medical Epidemiologist, Air Quality Health Effects Research Section, Environmental Health Directorate, Health Canada
May 10, 2004

Overall, this was a very impressive effort. It was thoughtfully planned, meticulously executed and clearly and thoroughly reported.

1. General Topics

1. *What are "best practices" for conducting expert elicitations? What "best practices" are essential for a defensible elicitation?*

I don't claim to be an expert on the methodology for expert elicitation, but it appears that the key issues (as laid out in e.g., Henrion and Grainger's book on uncertainty [Morgan and Henrion, 1990]) are appropriate preparation of experts, avoiding anchoring and other heuristics which can lead to biased and poorly calibrated results, and employing a defensible approach to combining results across experts, with appropriate use of sensitivity analysis.

2. *How does the EPA's elicitation compare to "best practices"? What are the strengths and weaknesses of this elicitation?*

The EPA elicitation compares favourably to best practices. I have already mentioned strengths in my overall comments above. The only feature which might qualify as a weakness, is the inclusion of two experts (Samet and Zeger) who have worked extremely closely on one particularly influential study (NMMAPS), thus possibly narrowing the range of views represented.

3. *How should the individual expert opinions be combined?*

I am not aware that there is any evidence that there are particular advantages to weighting schemes other than equal weighting.

4. *How could the elicitation be improved?*

As above, perhaps widening the range of views represented by the experts.

5. *Are the elicitation methodology and process appropriate as a benefit analysis technique to characterize uncertainty?*

Yes, but it would be interesting to evaluate the impact on benefits estimates of employing this approach versus the status quo, or alternative approaches (e.g., meta-analysis).

2. Specific Topics

1. Participants

   a. Did the set of participants chosen reflect the views of other scientists in the field?

      Yes, a reasonable range of views, although one might question the inclusion of both Samet and Zeger, who have worked very closely together, particularly as investigators in the highly influential NMMAPS study, and are from the same institution. Also, Krewski was apparently invited to participated, even though one of the criteria was being based in the U.S., although he is based in Canada.

   b. Was the number of participants appropriate?

      Appeared to be appropriate for a pilot study. There is of coursr no magic number, but rather it is more important to ensure that a range of reasonable views is represented.

   c. Was the method for choosing participants acceptable?

      I think the approach exhibited due diligence and would stack up well against alternative approaches.

   d. For these participants, does the potential for motivational bias exist?

      There's always potential. I suppose one would need to analyze whether each expert would stand to benefit from the views he expressed.

   e. Are the relevant fields represented?

      Might have considered including a toxicologist, but they might have less quantitative familiarity with the epidemiological evidence. Have included an expert on clinical studies (Utell).

2. Topic: Long- and short-term mortality

   a. Were the topics adequately described to the participants (eliminated ambiguity)?

      Yes.

   b. Were the correct questions asked (i.e., did they effectively address the topics without introducing bias)? Do you think that word choice, structure, or the order of the questions affected the quality of the results?

      Generally, yes (1st question) and no (2nd question). However, my understanding is that the quantitative elicitation is best done starting at the extreme values/ probabilities, and leaving the mid-values to the end. Also, that the order of questioning about specific percentiles be random. It wasn't clear exactly how this was carried out. It was also not clear exactly how the experts came up with their quantitative assessments, since it's hard to believe that they just pulled the numbers out of the air. Did they go back to the primary sources and actually compute values from specific tables or paragraphs in each source? This type of detail would be informative.

3. Preparation

a. Were any materials missing that should have been included in the Briefing Book? Should any materials have been excluded?

   No on both counts.

b. Was the time allowed for review of the materials adequate?

   Yes.

c. Some participants were more prepared than others—is this a problem? If so, how can this problem be addressed?

   I think this will always be an issue. One might consider underweighting the opinions of those who are obviously less prepared, but one would need some criteria for evaluating the level of preparedness.

d. Were expectations effectively communicated to the participants prior to the meeting?

   Yes.

e. Were the questions sufficiently tested prior to the elicitation?

   Yes.

f. Was adequate training provided for the participants prior to the elicitation? Would a pre-elicitation workshop have been helpful?

   It may have been helpful, but seems unrealistic both in relation to the additional time commitment and the logistics of coordinating schedules for 5 or more busy experts.

4. Meeting format

a. How do individual interviews compare to group interviews?

   My understanding is that group sessions must be handled very carefully in order to avoid the occurrence of group dynamics issues which colour the outcome of the deliberations.

b. Is it important to achieve consensus or not?

   This may be an unrealistic expectation.

c. Did the interviewers perform their roles properly?

   Yes.

d. Was the length of the interview appropriate?

Frankly, I was surprised that the experts went along with an 8 hour elicitation interview. It seems like an appropriate length given the amount of material to cover, but I would have thought many of these individuals would be reluctant to commit this much time, given other responsibilities.

5. Aggregation of opinions and final report

a. How should the quantitative and qualitative results from the individual interviews be combined?

I'm not aware that there is agreement on the best way to do this.

b. Was it appropriate to weight all the opinions equally? Are there alternative weighting systems that could have been used? EPA considered ways to calibrate the experts, which can be used to weight the experts' responses, but could not construct an applicable calibration question for this topic. Do you agree that you cannot calibrate experts on this topic? If not, do you have ideas for calibration questions.

   As I indicated earlier, I am not aware of evidence which supports the use of alternative weighting schemes. I believe it was appropriate to examine the sensitivity of results to the exclusion of individual experts, and to compare combined results to the various key pieces of evidence.

c. Can the results of the individual interviews reasonably be combined given the differences in forms of the functions? Of the four methods presented for combining the results, are any of them sufficient given the statistical approaches taken, and would you recommend other methods for combining the results?

   I think what was done was reasonable.

d. Are all of the essential elements included in the final report? Are any unnecessary elements included?

   Could have been more explicit about the exact order of by the quantitative elicitation questions.

6. Recommendations for future elicitations

   a. Absolute necessities for an expert elicitation (based on best practices)

      See above.

   b. Major strengths and weaknesses of this expert elicitation

      See above.