RTI INTERNATIONAL
3040 Cornwallis Road ▪ PO Box 12194 ▪ Research Triangle Park, NC 27709-2194 ▪ USA
Telephone 919 541-6000 ▪ Fax 919 541-5985 ▪ www.rti.org

TO: Ron Evans, EPA

FROM: Carol Mansfield and Sumeet Patil, RTI International

DATE: September 29, 2006

SUBJECT: Peer Review of Expert Elicitation

---

1. INTRODUCTION

This memo synthesizes responses provided by experts selected to review and comment on *Expanded Expert Judgment Assessment of the Concentration-Response Relationship between $PM_{2.5}$ Exposure and Mortality*. (Prepared by Industrial Economics, Incorporated [IEc] for the U.S. Environmental Protection Agency [EPA], Research Triangle Park, NC, August 25, 2006). The six reviewers were Drs. John Evans, Douglas Crawford-Brown, Granger Morgan, Thomas Wallsten, David Stieb, and Warner North.

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. Finally, we attach copies of the full responses as Appendix A.

1.1 Review Summary

Overall, the reviewers unanimously agreed that EPA conducted a high quality expert elicitation. The elicitation follows best practices and can serve as a model of good practice for expert elicitations in a variety of agency-wide settings. The reviewers agree that the elicitation protocol provides a reliable basis for eliciting the probabilistic distributions of uncertainty in the $PM_{2.5}$ C-R relationship. We reproduce some of the kudos by the peer reviewers that can better summarize the review than our summary above.

> Dr. Granger Morgan – "*Both EPA-OAQPS and IEc deserve warm congratulations for a job well done, which others across the Agency would do well to emulate. In addition to finalizing this report, EPA should encourage the authors, in the strongest possible terms, to publish the results as a refereed paper.*"
>
> Dr. Douglas Crawford-Brown – "*My summary judgment is that this study represents best practice and will hold up to scrutiny at*

> *least as well as other expert elicitations I have seen and participated in.*"
>
> Dr. Thomas Wallsten – "*Overall, the report is well written, well organized and highly readable. The described work is of very high quality and is carefully and thoughtfully done.*"
>
> Dr. David Stieb – "*The elicitation compares favorably with accepted practices…. The elicitation protocol was very thoroughly thought through and benefited from extensive consultation and pre-testing.*"
>
> Dr. Warner North – "*This exercise has been done commendably well, and the documentation in the report is excellent in explaining what was done and why*." *[he follows this sentence with his concern that expert elicitations, in general, can be taken out of context that these are after all only judgments]*
>
> Dr. John Evans – "… *as a member of the NAS committee which urged the EPA to improve its treatment of uncertainty in dose-response, as a consultant to the EPA SAB panel which reviewed the EPA Pilot Study, and as a scientist who has conducted several expert elicitations I believe that the EPA should be quite proud of the excellent work reflected in the report …. It is absolutely first rate and sets a fine example of what can be done with this approach. There is no question that the characterization of knowledge and uncertainty about PM mortality effects provided by this work is far superior to any previous analysis of this topic.*"

The reviewers specifically listed the following main strengths:

- The quality of experts and adherence to best practices were impressive
- The elicitation protocol benefited from extensive consultation, pretesting, real-time feedback, pre- and post-workshops, and effective use of technology.
- Selection of experts via a two-step nomination process with consideration to representation across fields clearly conforms with the best practices in the field.
- The experts were asked to provide their distributional information and a narrative justification for that distribution, which results in a better theoretical foundation for the distribution and more information for the reader and the other experts.
- The pre- and post-elicitation workshops represent the best practice. The post-elicitation workshop allowed experts to adjust their judgments if they had misinterpreted the intent of a question relative to the other experts.
- The report is well written, well organized, and informative.

Although the reviewers agreed that the elicitation was high quality, reviewers pointed out some of the issues with expert elicitation in general, and provided suggestion to further improve the exercise. Most of these comments are reviewed in Section 3. Below we note the key comments:

- Dr. North underlined that it was important to also recognize that expert elicitation is only one step in addressing uncertainty. He said that more research and dialogue are needed for improved understanding of uncertainty in C-R functions before they are widely applied in the decision making related to an important topic such as PM-induced mortality. Dr. Crawford-Brown also did not believe that subjective judgment is a proper sole basis for characterizing uncertainty. These comments are true for expert elicitations in general and not specific for this expert elicitation.
- Although the expert selection process was lauded, reviewers pointed out a few possible improvements. For example, one expert indicated that the panel could have been bolstered had it included government scientific assessors or evaluators who consider empirical evidence in developing standards. Another reviewer felt that inclusion of more toxicologists, researchers who understand the biomechanics of PM-induced mortality, and a well-known reviewer with a skeptical or critical viewpoint would have benefited the representation of the panel.
- Additional discussion and comparison of the summary results would be useful for the reader. For example, a comparative summary figure and explanations for some of the larger differences that emerged among the experts would be helpful. Currently the information is available only in the detailed discussions and specific sections provided in the report.

One important aspect of the elicitation that all reviewers commented on was the decision to provide experts with two options for specifying uncertainty distributions for the C-R function. One option was to develop a distribution that jointly considers (1) the probability of causality and (2) the conditional uncertainty (given causality) in the C-R function. The second option was to separately characterize above two quantities. Although the reviewers had mixed reaction to these options, in general, they agree that the resulting distributions were based on a careful elicitation, and thus defensible. The views presented on the issue are summarized here:

- Dr. Crawford Brown commended the elicitation for providing a choice that allows each expert to use an option that best reflects his/her views and understanding. However, he also suggested that EPA could have asked some experts to produced PDFs using both options, and then compare these as a sensitivity analysis. This sensitivity analysis could have informed us whether or not the experts were truly separating the causal claims from the conditional uncertainty.
- Dr Crawford Brown argued that experts found it confusing to understand and address causality claims as evident during their interviews. He believed that such confusion arose because experts were not confronted with the claims of causality at different levels of exposure. Therefore, experts could have assumed causality at a high level of exposure and thus produce high level of confidence on causality at a high exposure levels. He also recommended that EPA combines the causality judgment and conditional PDF in to an aggregate distribution for more proper use of these distributions in benefits assessment as opposed to suggestion by some of the experts in the elicitation.

Dr. Stieb also questioned whether causality and conditional PDF are independent of each other or not. May be the precision in measuring causality – conditional PDF – is limited but the evidence can suggest high probability of causality. Other the other hand precision in measuring causality is high, but the evidence suggests a lower probability of causality. He believes that some experts expected these two quantities to be dependent on each other. However, he recommends resolution of above question because it pertains to application of the findings from the elicitation to subsequent benefits assessments.

Dr. Morgan was also unsure about whether the experts understood the concept of incorporating the likelihood of a causal relationship (between PM and mortality) directly in their "composite" distribution.

- Dr. Stieb also argued that the concept of causality and conditional PDFs were interpreted inconsistently by the experts. He claimed that combining the probability of a causal relationship with the quantitative conditional PDF of uncertainty problematic can cause ambiguity in interpreting the distributions elicited from the two groups of experts – one that combined the two parts and the other that kept them separate. He argued in favor of quantifying the probability of causality and conditional PDF separately.

  On the other hand, Dr. Evans indicated his preference for not disaggregating the probability of causal interpretation and the existence and location of a threshold from experts' answers. He questioned the availability of sufficient evidence, and thus expertise, on the issue of a population threshold for PM mortality. Therefore, he sided with not disaggregating the two components of the uncertainty distribution. However, he acknowledged that the EPA/IEc team presented especially sound and acceptable rationale for above decisions considering that such choices are often arbitrary in any expert elicitation.

  Dr. Morgan was also unsure about whether the experts understood the concept of incorporating the likelihood of a causal relationship (between PM and mortality) directly in their "composite" distribution.

---

2. QUESTIONS SENT TO REVIEWERS

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

---

3. REVIEWERS' RESPONSES

Reviewers were given copies of the report and the set of questions seeking input on several aspects of the document. The reviewers were also provided with additional background information, including

- Chapter 5 of the National Academy of Science (NAS) review of EPA's methods for conducting benefit analyses,
- the Pilot Expert Elicitation report,
- the peer review of the Pilot Expert Elicitation report,
- EPA=s response to the peer review,

- a summary of each expert's interview, and
- a summary of the pre- and post-elicitation workshops.

Figure 1: Charge and Questions

Dear Dr. [name]:

Thank you for agreeing to serve as a peer reviewer of EPA's Expanded Expert Judgment Assessment of the Concentration-Response Relationship between PM2.5 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 consensus on how to interpret the 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 PM2.5 exposure and mortality. In 2004, EPA conducted a pilot elicitation on this topic. The peer review of that elicitation informed the design of the current, final elicitation we are asking you to review.

The charge for this peer review is to provide technical feedback on the methods employed for this expert elicitation, with particular emphasis on whether EPA used best practices in the design of the elicitation. Does the report accurately characterize the results of the individual elicitations? EPA wants to ensure the procedures and tools used to conduct the elicitation were adequate and in accordance with general guidelines for conducting expert elicitations. 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.

In addition to the final report, hard copy and included on the CD, we have attached a CD with a few additional documents that you may find helpful.

- Chapter 5 of the NAS review of the EPA=s methods for conducting benefit analyses
- the Pilot PM Expert Elicitation report
- the peer review of the Pilot elicitation report

- § EPA=s response to the peer review
- § Summary of each expert's interview
- § Summary of pre- and post-elicitation workshops

We request that you submit a written review no later than September 13. 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 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 (carolm@rti.org) or at (919) 541-8053 or Sumeet Patil at e-mail (spatil@rti.org) or (919) 316-3931.

Sincerely,

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

Enclosure

Questions for Reviewers

Please feel free to address other topics you consider important.

General topic: Is the EPA's expert elicitation defensible in terms of assumptions, methodology, and prevalent best practices in the expert elicitation field? What are the strengths and weaknesses of this elicitation?

Specific topics:

1. Selection of Experts
2. Design of the Elicitation Protocol
3. Background Materials/Briefing Book
4. Communication with experts pre- and post-elicitation
5. Elicitation

6. Summary of Findings and Final Study Report (IEc, 2006)
7. Responsiveness to reviewers
8. Overall Comments

The table below lists the topics for the review on the left with more detail on the issues you might address as part of the topic on the right.

Topics for Review	Detail on Topics
Selection of Experts	Was the method for choosing the experts consistent with standard practices? Are the relevant fields represented? Did the set of experts selected reflect the views of other scientists in the field? Was the number of experts appropriate given the topic covered by this elicitation and the number of studies and experts on the topic?
Design of the Elicitation Protocol	Did the elicitation cover all the topics relevant to PM mortality? Were the topics adequately described to the participants (eliminated ambiguity)? Do you think that word choice, structure, or the order of the questions affected the quality of the results? Did the protocol design adequately control for heuristics and biases in the process?
Background Materials/Briefing Book	Were any materials missing that should have been included in the Briefing Book? Should any materials have been excluded? Were any biases introduced given the set of materials provided to the experts?
Elicitation	Were expectations of the elicitation process effectively communicated to the participants prior to the interview process? Was adequate training provided for the participants prior to the elicitation? Was the pre-elicitation workshop properly conducted (based on the description provided in the report)? Was the length and format of the interview appropriate? Were the tools used during the interview process acceptable (i.e., use of a domain expert and expert in elicitation methods, web link to two additional observers/recorders, transcription and summary of the

Topics for Review	Detail on Topics
	interview, cards for key questions, electronic visual of expert's distribution provided as immediate feedback to allow for adjustments, etc)? Was the interaction and feedback after the elicitation appropriate? Were the summaries of each interview adequate and appropriate? Was the post-elicitation workshop properly conducted (based on the description provided in the report)?
Summary of Findings and Final Study Report (IEc, 2006)	Are all of the essential elements included in the report? Is there adequate information in the report to understand how the interview went and issues that were addressed during the interview? Can you suggest other analyses that could have been done with the data? Can you suggest other ways to present results (e.g., other than box and whiskers)?
Responsiveness to reviewers	Did the elicitation adequately address the concerns and comments from the peer reviewers of the pilot elicitation? Were any biases introduced given the changes made to the protocol as a result of the pilot?
Overall Comments	Overall, how does the EPA=s elicitation compare to Abest practices@ or acceptable practices for a defensible expert elicitation? What are the strengths and weaknesses of this elicitation?

Below, we present the questions posed to the reviewers with a summary of the responses. We typically try to identify common themes in the reviewers' response to specific questions and summarize them. We also highlight more detailed of contrary comments by specific reviewers. We recommend reading a reviewer's report in totality (Appendix A) to gain a holistic understanding of the reviewer's comments. Please note that we avoid repeating the praise reviewers bestowed on this expert elicitation and sometimes even tone it down to avoid redundancies, and instead, focus more on comments targeted at further improving the analysis.

3.1 Selection of Experts

In general, the reviewers agreed that the methods used in selection of experts were in line with best practices in the field. They find the process commendable in terms of selection process, representation of the relevant fields and diversity of opinion, and the number of experts. Most reviewers recognized that the panel was representative enough, but they wished for inclusion of some other fields to further improve it.

a. Was the method for choosing the experts consistent with standard practices?

All reviewers agreed that the selection methods were in line with best practices and EPA went well-beyond what was necessary. All researchers agreed that the selection process resulted in identifying twelve highly respected and suitable experts. Dr. Evans found the process to select 3 toxicologists randomly from a list of 10 a bit unusual but accepted the argument in favor of including toxicologists in the panel. Dr. Crawford-Brown, as discussed in his review in more detail, thought that the report could have done a better job describing the nomination process. Dr. North and he wished for experts with alternative or skeptical opinions regarding the link between PM and mortality in the panel. Dr. Crawford-Brown also indicated that "publication bias" can affect the selection of the panel. The selection process focused on published literature to find experts and may have excluded people without readily available publications or standard peer-reviewed journal articles. Therefore, although EPA has a good panel of experts, 'perception' about representativeness of the panel might be an issue to consider.

b. Are the relevant fields represented?

The reviewers found that the final panel is adequately representative of the relevant fields. The panel mainly consists of epidemiologists, which the reviewers felt was reasonable given that they have good insight into the available data. However, some reviewers had suggestions for other expertise in the panel such as government staff who review and evaluate evidence in the development of standards, and more toxicologists and general medicine researchers with knowledge of the biological mechanism of PM-induced mortality. Dr. North argued in favor of including at least one well-known expert with critical or skeptical viewpoints.

c. Did the set of experts selected reflect the views of other scientists in the field?

The reviewers agreed that a range of views in the scientific community are reflected in this exercise, although it is difficult to evaluate whether 'all' viewpoints in the field are well reflected. For example, Dr. Stieb found that the range of uncertainty judgments expressed in the pilot expert elicitation was wider than those expressed in the current elicitation and wondered if the selection of experts or maybe the emergence of new evidence explained the current narrower range of uncertainty judgments. Dr. North recommended further discussion and review to evaluate how representative the viewpoints of the 12 experts are compared with other scientists in the field. Dr.

Crawford-Brown argued that despite his desire for experts from additional fields, the expert panel participated in a rigorous exercise and produced estimates of uncertainty that are more defensible than those in the general scientific community anyway.

d. Was the number of experts appropriate given the topic covered by this elicitation and the number of studies and experts on the topic?

The number of reviewers is substantial for this expert elicitation. In a expert elicitation, the uncertainty distribution is viewed as a reflection of experts' understanding based on available evidence and examination of each other's judgments. Therefore, as per Dr. Crawford-Brown, increasing the number of experts more than twelve may only provide very marginal benefits in terms of variability in the PDFs, which is not warranted.

3.2 Design of the Elicitation Protocol

Overall, the reviewers found that the protocol was of high quality and properly elicited judgment regarding PM and mortality. The protocol is well reasoned and justified and conforms to the best practices. The protocol included important and relevant topics and explained the issues appropriately. However, as we discussed earlier, reviewers argued that concepts of separating the probability of causality and conditional PDFs were unclear to the experts.

a. Did the elicitation cover all the topics relevant to PM mortality?

The reviewers agreed that EPA/IEc did a commendable job covering topics relevant to PM mortality. Additionally, experts had the opportunity to include topics that were not originally included in the protocol if the expert felt the need.

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

Reviewers agreed that the topics were adequately described to the experts, especially considering their expertise. However, Dr. Stieb commented that the separation of the probability of a causal relationship and the elicitation of the quantitative distribution of risk appeared unclear to experts even in the post-elicitation workshop. Dr. Crawford-Brown also agrees with him.

c. Do you think that the word choice, structure, or the order of the questions affected the quality of the results?

The reviewers, in general, believed that the form of the elicitation process did not unduly influence the judgment of uncertainty or the results. All these elements have been tested in a pilot expert elicitation and discussed in a workshop. Also, the experts were given options for characterizing the uncertainty. However, Dr. Wallsten argued that the primary question—estimating the true percentage change in annual, all-cause mortality resulting from 1 $\mu g/m^3$ reduction in annual average PM concentration—is too complex for the participants to think about carefully. Answering this question requires the experts to take into account a variety of complicated issues, unlike a simpler question related to C-R

functions. Dr. Wallsten noted that, in general, a simpler question, and thus a simpler judgment, will yield more useful results. However, he also recognized that both simple and complex encoding of uncertainty have advantages and disadvantages. Dr. Evans, on the other hand, believed the elicitation protocol worked the experts up to the above question carefully in the context of the extensive background material available and the experts' own knowledge of the topic.

d. Did the protocol design adequately control for heuristics and biases in the process?

Reviewers agreed that the protocol controlled for heuristic and biases in the process as well as possible. For example, the protocol guided experts to think about all perspectives on the issues, evaluate the effect of their assumptions, consider a range of evidence, and provide a theoretical basis for their judgment. However, Dr. Stieb wished for a more specific discussion of how particular elements of the protocol addressed particular heuristics.

3.3 Background Materials/Briefing Book

Overall, the reviewers were impressed with the quality and the content of the briefing book. In addition, they dismissed the possibility of any serious bias in the materials.

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

Dr. Evans said that the briefing book was well organized and written, and that he was impressed with the care that has obviously gone into the preparation of these materials. Other reviewers resonate similar feelings. At the least, the reviewers agreed that all relevant topics were touched on, several of them covered adequately, in the exhaustive review of the scientific literature. No reviewer recommended excluding any material. Dr. Stieb listed four additional studies that could have been considered (see Dr. Stieb's review in Appendix A).

b. Were any biases introduced given the set of materials provided to the experts?

The reviewers found that there is no readily apparent bias introduced by the set of materials provided to experts. Dr. Stieb argued that the experts would ultimately gravitate toward the same basic set of materials regardless of whether they were included in the briefing book so that a bias is unlikely.

3.4 Elicitation

The reviewers lauded the elicitation and a couple of them even recommended it as a model to emulate for conducting future elicitations. The qualitative review of the available evidence provided for more structured discussion that resulted in efficiency in the interview process and understanding the experts' thought processes. The web-based software to visualize distributions during the elicitation was a commendable tool. The pre-elicitation workshop and the elicitation

itself was very carefully done and well structured. The flow of topics was logically and systematically developed. The post-elicitation feedback and subsequent workshop were very helpful. The quality of briefing book and the training process was also commendable. Reviewers, in general, find the interview summaries very informative and a useful addition to the report.

a. Were expectations of the elicitation process effectively communicated to the participants prior to the interview process?

The pre-elicitation workshop, well-structured process, appropriate duration of the process, and post-elicitation workshop all helped prepare the experts and effectively communicated the expectations for the exercise. Dr. Crawford-Brown believed the text materials should have been adequate for the experts to understand the task for them. However, the actual conduct of the information and discussion sessions ultimately plays a role in determining whether and how well the experts understood the expectations, to which he cannot attest.

b. Was adequate training provided for the participants prior to the elicitation?

All experts approved of the process and structure used to train participants and found the training to be exhaustive. Dr. Crawford-Brown was not sure that whether the training included an exercise to calibrate the experts' judgment. This exercise could have helped them understand how judgment of uncertainty differs depending on the quality of the available data. On a related topic, he noted that there are systematic problems in getting even the experts to widen confidence intervals appropriately to reflect nonstatistical sources of uncertainty. He recognized that this expert elicitation process is in the lines with best practices. However, he wondered whether there was any training on the above aspect because the confidence intervals were narrower than what he would have expected.

c. Was the pre-elicitation workshop properly conducted (based on the description provided in the report)?

Based on the materials provided and description of the process in the report, the workshop was conducted properly. However, Dr. Crawford-Brown underlined that the actual in-class participation and presentation were equally important and cannot be evaluated without attending the presentation. Dr. Evans felt that the pre- and post workshops were very useful features of this expert elicitation and that are in the lines with the best practices.

d. Was the length and format of the interview appropriate?

The interview provided the right balance between (1) providing enough time for the expert to reflect on the relevant considerations and (2) providing so much time that the expert might be swayed from an already firm judgment. Dr. Evans pointed out that the quality of information collected toward the end of a day-long interview may be

compromised, but a day-long interview is often needed to elicit an informed judgment regarding complex questions. He states that this tension is typical and constant in expert elicitation and not unique to this exercise.

e. Were the tools used during the interview process acceptable (i.e., use of a domain expert and expert in elicitation methods, Web link to two additional observers/recorders, transcription and summary of the interview, cards for key questions, electronic visual of expert's distribution provided as immediate feedback to allow for adjustments, etc.)?

Overall, the reviewers were impressed by above tools which were all worthwhile additions to the protocol relative to the pilot. These tools would lend additional validity to the elicitation. Dr. Crawford-Brown found that Venn diagrams to explain causality was confusing, perhaps because the question of whether there is causality depends on the level of exposure.

f. Was the interaction and feedback after the elicitation appropriate?

All the reviewers agreed that the interaction and feedback were appropriate, subject to Dr. Crawford-Brown's observation that his opinion is based on the report, rather than on actual observation of the process.

g. Were the summaries of each interview adequate and appropriate?

Most of the summaries clearly described the thought process each expert used in formulating his quantitative estimates and adjusting for various factors. However, Dr. Crawford-Brown pointed out a possibility of post hoc adjustment or rationalizing of the findings elicited during the interviews. For example, when an interviewee sees a transcript, he may "improve" it in hindsight, but this "improvement" may not reflect the actual reasoning that went into production of the PDF. Dr. Stieb could not reproduce the numerical estimates in some of the cases, and this, expected more clarification on such cases. He provided a few examples to this effect.

h. Was the post-elicitation workshop properly conducted (based on the description provided in the report)?

Based on the report, the reviewers found that the workshop was well conducted. Dr. Crawford-Brown again emphasized the importance of actual execution. He acknowledged the suggestion of some experts to keep separate the estimates of uncertainty due to causality between $PM_{2.5}$ and mortality and the conditional uncertainty in the C-R function. However, for more practical use by EPA, he recommended using the aggregate PDF that jointly considers the uncertainty due to causality and the conditional uncertainty.

3.5 Summary of Findings and Final Study Report (IEc, 2006)

Reviewers found the report to be very informative and well organized and the results well presented. Dr. Evans specifically appreciated the sensitivity analyses and EPA's and IEc's own assessment of the strengths and weaknesses of the study. The report clearly satisfies EPA's needs for an analysis of the uncertainty in the C-R function. However, reviewers made additional suggestions to further inform the reader.

a. Are all the essential elements included in the report?

All essential elements are included in the report and the report is very clear in understanding the process and results. The tables and figures in the report are adequate in number and very informative. The reviewers suggested additional information that could further improve the report. For example, Dr. Stieb recommended adding a discussion on the relationship between quantifying the magnitude of mortality risk and probability of a causal association. Dr. Wallsten recommended including more information on the encoded C-R functions and how to use them to estimate the answers to the main question of interest.

b. Is there adequate information in the report to understand how the interviews went and issues that were addressed during the interview?

The interview summaries provide excellent information on what was addressed during the interview and the expert's thought processes in providing judgments in the form of probabilities. Experts appreciated the detailed description of the interview. Dr. Stieb noted a lack of clarity in some specific aspects of the interview summaries as discussed earlier when he could not reproduce some of the steps. Dr. Crawford-Brown brought to our attention that some insights are only gained and retained during the discussion sessions and cannot be reported in a report which is more structured than an interview.

c. Can you suggest other analyses that could have been done with the data?

Overall, reviewers find that the analyses presented in the report adequately meet EPA's needs. Some additional analyses can be of further interest. For example, Dr. Stieb recommended displaying C-R functions with inferred confidence intervals so that readers can get a sense of the overlap in differently shaped uncertainty distributions. Dr. Crawford-Brown recommended displaying how variability-specific percentiles of the uncertainty distribution differ across the 12 experts – a 2-dimensional surface. Dr. North recommended calculating the expected decrease in mortality using the uncertainty distributions. He believed that such analysis would provide a useful perspective on the extent of the differences among the experts. Dr. Evans found the analysis of the results fairly limited. However, he liked the sensitivity analyses that explored alternative ways to combine the experts' judgments.

d. Can you suggest other ways to present results?

Experts agreed that the box and whisker plots are informative and well established in the literature. Most reviewers find results complete and easy to understand. Dr. Morgan recommended comparing summary results or box plots of the experts in a single diagram because one has to read through a great deal of information to do a comparative analysis. Dr. North also expressed similar concerns. Dr. Morgan recommended a discussion of the obvious differences in these plots in the light of differences in the viewpoints of the experts. Some of the box plots represent a flat but multimodal distribution, which the experts may not have meant to imply. EPA may need to check such plots with experts. He also recommended reporting upper and lower bounds, as well as 5th and 95th percentiles of the uncertainty distributions judged by the experts.

3.6 Responsiveness to Reviewers

Reviewers based their comments mainly on the summary information provided in the report on the peer review of the pilot elicitation. Overall, they found EPA's and IEc's response to issues raised in the peer review of the pilot elicitation to be satisfactory. However, they recommended more discussion about the peer review of the pilot elicitation and the resultant changes to the current elicitation.

a. Did the elicitation adequately address the concerns and comments from the peer reviewers of the pilot elicitation?

Overall, all the reviewers agreed that the report and elicitation adequately addressed the comments on the pilot elicitation. However, the reviewers have not evaluated this in detail. For example, Dr. Crawford-Brown only verified whether his comments on the pilot elicitation were adequately addressed. He believed that the methodology in this elicitation is consistent with the general concerns raised by pilot peer review. Dr. Stieb relied on the brief summary provided in the report. He recommended including more explanatory material on how earlier comments on anchoring and adjustment bias are addressed in the current elicitation. Dr. Evans appreciated asking for the combined impact of changes in short-term and long-term $PM_{2.5}$ levels in a single coherent question rather than in two separate questions, as was the case in the pilot expert elicitation. He found this modification as one noteworthy outcome of the peer review of the pilot expert elicitation.

b. Were any biases introduced given the changes made to the protocol as a result of the pilot?

The reviewers found that the elicitation process included reasonable measures to avoid biases and thus is not biased as a result of any revisions prompted by the pilot peer review. However, Dr. Stieb argued that, in the pre-and post-workshops, it is possible for more persuasive experts to influence the opinions of other experts, irrespective of the validity of their arguments.

Appendix A: Original Reviews

Review of Expanded Expert Judgment Assessment of the Concentration-Response Relationship Between PM 2.5 Exposure and Mortality

Review by Douglas Crawford-Brown
University of North Carolina at Chapel Hill

General Comments

I begin by saying that I was impressed with the work done here. While I am not generally a believer that subjective judgment, however expert, is a proper sole basis for characterizing uncertainty, this study has done as good a job as it is possible to do in eliciting these judgments. One of the chief causes of systematic underestimation of uncertainty in science is the tendency to ignore some of the key weaknesses in causal claims (the claim that PM causes mortality at all; the claim that there is or is not a threshold). The resulting uncertainty is usually then totally conditional: the slope factor uncertainty distribution becomes conditional on the claim that there is any causal connection to begin with. The present study avoids that problem by confronting the assessor with the claim that there is a causal connection and with the claim that there is no threshold, and then factors these causal judgments into the composite uncertainty PDFs.

I was also impressed with the way in which the assessors were given flexibility in reflecting this uncertainty as to causality and threshold. The decision to either have the assessor "bury" the causal judgment inside the uncertainty distribution, or to have this judgment as a separate component and then combine it with the conditional distribution (conditional on the causal claim) was a wise choice. Research makes it clear that some people prefer one way or the other of aggregating these kinds of uncertainty, and so providing a choice allows each assessor to decide which approach is best for themselves. In the approach using the conditional distribution, I felt the use of the Monte Carlo method for sampling was entirely appropriate.

The one point here where I might disagree is that I would have found it more informative to have each expert (or at least some of them who were willing to do so) produce the PDF under BOTH approaches, and then to compare these as a kind of sensitivity analysis. This would let us know whether individuals truly were separating the causal claims from the conditional estimates of the risk coefficient.

The use of what I might call a Delphi Method, in which the assessors were confronted with the judgments of the other assessors and then given the opportunity to adjust their views, was also a good feature. What made the approach especially rich is the fact that the individuals were asked not only for their distributional information, but for a narrative justification for that distribution. This aspect of elicitation often is missed, and it is important in the Delphi Method to understand not only the judgment made by others on the team, but the evidential reasoning used in forming those judgments. The reason for a particular assessor to change a judgment of uncertainty should not be that another assessor has produced a different PDF, but rather the soundness of the argument given by

the latter to support his or her PDF. I didn't sit in on any of the sessions where these narratives were produced, and so I can't attest to the rigor of those narratives (i.e. whether the narratives had any structure to them that caused the assessor to consider modes of evidential reasoning, concepts of the epistemic basis for judgments, etc). However, the categories of questions asked of each assessor (as reported in the Appendices) are at least the ones one would have expected to raise these issues of evidence and reasoning. I might have wished to see a little more philosophical structure to the questions, helping the assessor understand better what makes for reliable and unreliable modes of reasoning, but I am not convinced this would have significantly altered any results.

All in all, then, I consider this to be a defensible study and a reasonably reliable basis for eliciting the subjective judgment of uncertainty in the PM 2.5. concentration-response relationship.

Specific Comments on Questions

1. Selection of Experts

1A. Choosing the experts. The criteria for choice here are consistent with what I take to be best practice. This is always a contentious step, since sociological studies indicate that results can depend on the institutional affiliations of the respondents. Part of this has been dealt with by having a nomination process. The criteria on Pages 2-11 and 2-12 offer a rich and nuanced set of considerations for anyone making a nomination. I didn't, however, fully understand the terminology of "nominator" in Exhibit 2-3. Are these really people who nominated someone else, or are these people who were nominated by someone? I had thought the latter, since the list of Exhibit 2-4 is drawn mostly from the list of 2-3. But I suppose the experts themselves may have been asked for their own views of appropriate team members. I just wasn't clear about that point. And I also wasn't clear whether the nomination process was open to organizations we might consider "causal skeptics" with respect to PM. It seems to me it was not, since it focused largely on individuals on who had published studies (see my later comments on publication bias). I worry that this might either be perceived as "stacking the deck" or might be the equivalent in climate change of omitting skeptics from a panel. I still think the Agency has a reasonable sample of experts used, but perception will an issue to be considered.

1B. Are the relevant fields represented? I am comfortable with the people, and the scientific backgrounds, of the final panel. I did find it dominated by epidemiologists and would have preferred to see a few more people from a clinical background and perhaps one from modeling. But the epidemiologists selected are certainly the people one would expect to have the best insights into the available data, especially given that issues of confounding and other epidemiological limitations are core causes of uncertainty. And several of the epidemiologists selected have broad experience that makes their work touch on clinical and modeling work. So in the end, I felt the mix was correct.

1C. Did the experts reflect the views of other scientists? This is a bit tougher to answer, since there are two issues buried in the question. It might refer to the views of other scientists on the risk estimates, or it might refer to the views on uncertainty. I do think the team reflected the range of views on the former issue. Given that the exercise itself helped the assessors understand, clarify and codify their uncertainty, I believe the selected team, in the end, produced judgments that are more defensible than those of the general scientific community, and so I am unconcerned that their judgments might not reflect those of other scientists who did not pass through this process. I am, however, convinced that the judgments obtained reflect what would have been the judgments of similarly qualified scientists had those others been in the process.

1D. Was the number of experts appropriate? The answer here depends on how you view such expert elicitations for uncertainty. If you believe there is a "true" uncertainty distribution, and the various experts are samples for getting at this true distribution, then 12 is probably an insufficient sample size, especially for estimating the tails of the distribution. I don't, however, believe there IS a true distribution. Instead, the distribution emerges in part from the reflection of thoughtful and experienced individuals confronted with available evidence, and in part from the social process by which they examine each other's judgments. Under this conception of uncertainty (which is the one I prefer), 12 is an adequate number of experts.

I think it is also important to bear in mind at this point the nature of uncertainty assessment. Uncertainty is not an objective property of the world (the world is not uncertain- WE are uncertain). Confronting and characterizing uncertainty is rooted ultimately in judgment, and judgments are fluid, imprecise. It is important to be as precise as possible in creating the PDFs, but also to recognize that there is a limit to the rigor with which these PDFs can be generated. Any further refinements of the ways in which the PDFs were generated in this study, including increasing the number of experts sampled, would to me be of very marginal utility. They might produce slightly different PDFs, but I am not convinced those PDFs would be any better than the ones produced here.

2. Design of the Elicitation Protocol

2A. Did the elicitation cover all topics of relevance? No, but I believe this would not change the results. For example, there are number of uncertainties associated with the deposition and clearance of particles in the lung, and the behavior at carinal ridges, that we know are relevant to risk but were not part of the discussion. And yet, this absence doesn't concern me. There is nothing about the aspects of uncertainty left out that would have been different in nature from the aspects that were included. After considering a few of the causes of uncertainty, people tend to converge (individually) onto a pretty stable PDF, so long as the considerations included cause them to think more generally about issues of causal claims, thresholds, sensitive subpopulations, extrapolations, etc. These issues are raised by the factors that WERE considered, and so I am comfortable that the experts reflected the range of uncertainty appropriately. The reason I am so comfortable with the range of topics considered is that I believe such topics are simply vehicles for

confronting general causes of uncertainty (the possibility of thresholds, etc); it almost doesn't matter which vehicle is used so long as the general issues arise within those vehicles.

2B. Were the topics described adequately? Yes, given the pre-existing expertise of the experts. The materials they were provided at least raised the relevant questions and pointed to the relevant bodies of data.

2C. Did word choice, structure, etc, affect results? No. This can be the case when eliciting perceptions of the risk of a specific situation relative to other situations, or selecting a course of action or inaction, or eliciting a willingness to pay for some amenity. But I don't see any sense in which the form of the elicitation process here would unduly influence the judgment of uncertainty. And the fact that the experts were given several options for characterizing that uncertainty seems to me to avoid this problem.

2D. Did the protocol adequately control for heuristics? I am not an expert at conducting such elicitations, and so I can't be sure of the answer here. I'm sure another reviewer can answer this better.

3. Background Materials

3A. Were materials complete? I went through the materials with an eye towards "reasonable completeness". By this, I mean whether the materials at least touched on the relevant topics, rather than being an exhaustive review of existing scientific studies. I believe they did cover all of the relevant topics, including the classic problems associated with epidemiological studies. The articles/papers provided to the experts constituted a very good review of the literature, although I doubt the experts would each have read all of the papers provided (there simply were too many). I can see no reason to exclude any of the materials, so long as the sample wasn't biased (see 3B).

3B. Was there any evident bias? If there is a bias, it is due to publication bias rather than an inappropriate selection of articles that appear in the literature. I think it is safe to say that papers tend to be published with lower probability if the result is negative than if the result is positive, and so this would tend to skew the literature base towards studies that show a causal link between PM exposure and mortality. But given this publication bias, I believe the sample provided the experts was appropriate and unbiased. What is important here is that many of the articles discussed the major limitations in drawing conclusions, and so the experts should have had their attention drawn to the full range of limitations even with this abbreviated sample.

4. Elicitation

4A. Were expectations effectively communicated? This is impossible for me to answer, since the term "communicated" involves as much the reception on the part of the expert as the quality of the materials developed for presentation. I did an exercise in which I imagined myself to be one of the experts selected, and then I read through the material

they were provided and asked myself whether I understood the task before me. If I had been in their position, I believe the materials would have prepared me for the task. But I was not in on the information/discussion sessions, and so I can't attest to how well these were conducted and whether the experts therefore "received" the message as intended.

4B. Was adequate training provided? Again, I have only the materials presented, and a description of the training. Two professors can work from the same materials and the same course outline, and yet differ dramatically in how well they teach and therefore train their students. But as in 4A, I would say that the materials provided and the structure of the exercise lead me to believe that training was adequate. This is particularly true given the quality of the experts taking part in the study, who I suspect had already considered issues of uncertainty in these slope factors.

Perhaps I missed it in the writing, but expert elicitations often contain a step in which the experts are "calibrated" in some way. They might, for example, go through a simple exercise in which the PDF is elicited when there is a lot of very good data and causal clarity, and one in which both of these are poor. The goal there is to be sure that all individual have collectively experienced what the data look like in these two cases so they can be sure they would have produced similar judgments of uncertainty in these special cases. I can't see where such an exercise was part of the training process. Again, I may have simply missed it, and I don't think this would have significantly affected the results here since the tasks are so well laid out, but it did strike me as a missing step.

Of particular interest to me in this regard is the fact that the confidence intervals in Figures 3-10 and 3-11 are only marginally wider than those of the two primary original studies. The confidence intervals in the latter reflect only measurement error (I believe) and not the full source of uncertainty (which includes conceptual uncertainty, issues of bias and confounding, etc). For expert elicitation of uncertainty to be most effective, experts need to be confronted with examples of how confidence intervals are usually greatly underestimated. This can be done by showing some historical examples in which subsequently improved measurements of some property (such as slope factor for a particular compound) shifted over time, with later measurements often falling outside the confidence bounds of earlier measurements. This isn't to say that the expert elicitation here is incorrect, or doesn't follow best practice (it does follow it), but rather that there are systematic problems in getting even experts to widen confidence intervals appropriately to reflect non-statistical issues.

4C. Was the pre-elicitation workshop properly conducted? My answer is the same as in 4A and 4B: that the write-up suggests a properly conducted workshop, but what is said on paper and what took place in the "classroom" may differ.

4D. Was the length and format of the interview appropriate? Yes. I believe the length found the right balance between providing enough time for the expert to reflect on the relevant considerations without provided so much time as to sway the expert from an already firm judgment. What I mean by the latter statement is that if you overly analyze a judgment, providing too much time at the end for revision of the judgment, an individual

can feel compelled to make changes to a judgment that was already good enough. There is a point in analysis beyond which an expert is simply making rather random changes in the judgment without necessarily improving that judgment.

4E. Were the tools used during the interview process acceptable? Yes. I was quite impressed by the tools. The one exception is the part on Venn diagrams. Perhaps the interviewer was able to explain this issue better to the expert than did the materials, but I had some trouble understanding fully what the pictures of Venn diagrams were intended to provide by way of help in moving the expert towards a judgment of causality. I suspect that the treatment of short-term and long-term effects remained somewhat confusing to the experts, and am not sure how they sorted through this in their own minds. This in turn leads me to wonder whether they fully understood the implications of this distinction on the uncertainty PDFs they generated. After several readings of the materials, I am still not sure I would have understood how to use this distinction in forming my PDF.

Whatever tools were used to address causality claims, it is clear that these still are causing confusion on the part of the experts. This appears to be because the question of whether there is a causal connection is too generic. Experts felt more comfortable with a claim of causality at different levels of exposure. This could have been dealt with by having them produce a plot of exposure level versus level of confidence in causality. My concern is that absent such nuanced information, the expert may have been interpreting the question as: do PM exposures cause mortality at SOME level of exposure (however high that might be). This would produce high levels of confidence. The information on Page 3-21 suggests this may have been going on.

4F. Was the interaction and feedback appropriate? Here, I have the same answer as in some previous questions. The STRUCTURE of the feedback, including the opportunity for feedback and the opportunity to adjust PDFs in light of results from other experts, was appropriate and laudable. What I can't comment on is whether the actual CONDUCT of that feedback in the exercise was appropriate. I assume it was given the structure, but I have been involved in proper structures where the specific elicitor was the source of the problems. I just can't answer this question in any more detail without having been part of the process myself.

4G. Were the summaries appropriate? The summaries of the discussions by the experts seem to me appropriate, although since I was not there for the actual interviews, I can't vet the text of the report as being an accurate reflection of what was actually said. The discussions appear to have been rich, and so I encourage the Agency to keep these summaries for use by others conducting research on expert elicitation (perhaps after removing any identifiers).

I also wasn't clear as to the nature of the editing that went into the summaries. Were the interviewees given the opportunity to review and then provide edits? An issue here is that when an interviewee sees a transcript, the tendency is to "improve" it in hindsight. But this "improvement" may not reflect the actual reasoning that went into production of the PDF. It might instead be a post-hoc rationalization, or suggest a different PDF should

have been generated in the first place. I cannot tell from the text how this issue was resolved.

4H. Was the post-elicitation workshop properly conducted? Again, the answer is yes in structure, and indeterminate with respect to actual execution. It strikes me that the workshop would need to be conducted in part by someone who knew how to dissect lines of reasoning (perhaps with some philosophical training) and to spot the difference between a valid and invalid line (by which I don't mean "true" and "untrue", just whether a conclusion follows formally from the evidence given to support it). I agree entirely with the experts who said they were somewhat confused about the role of the causal judgment given that they were first asked to give a conditional PDF based on the assumption of causality. I also would have been somewhat confused at first, and I assume the workshop addressed this confusion.

I am in less agreement with the experts on the desire to keep the aggregated PDF out of the report. They may have felt uncomfortable combining the results in the way suggested (i.e. combining the causal judgment and the conditional PDF), but such a combination is a natural consequence of the way in which the exercise was conducted. However, I do think one or more of them offered a classical, philosophical approach: to treat the conditional PDF quantitatively and the causal judgment qualitatively. I can understand why the Agency is choosing not to take this approach, as the qualitative part would undoubtedly get lost when the actual benefits uncertainty analysis is conducted. But the experts were drawing attention to a valid issue of misplaced concreteness (i.e. a false sense that the judgment of causality can be given a truly quantitative interpretation). Still, I think the Agency would be justified, at least formally, in combining the two parts of the distribution into an aggregate distribution, even if the experts were uncomfortable with it.

5. Summary of Findings

5A. Are all essential elements included in the report? Yes. I was able to find everything I needed to understand not only the results, but how they were generated.

5B. Is there adequate information to understand how the interview went and issues that were addressed? This is always a difficult question to answer. While the interview was structured, it also had an open-ended aspect to it (which was an appropriate design). Whenever an interview is open-ended, there is a need by the author to do a bit of post-interview rationalizing in summarizing discussion- making it appear perhaps more orderly than it was to bring some clarity to it for the reader. This is not, after all, a legal document with a virtual transcript of the conversation. I think everything is available in the report and supplement to help the reader understand the issues raised and the thinking process by the experts. But it will take some additional work to tease further insights out of the discussions so the reader can fully understand the concerns being raised by the experts and how these were reflected in the judgments and PDFs produced. Having said this, though, I believe this is additional work is not needed to convince me that the judgments and PDFs were properly formed by the process. Instead, this might be interesting follow-on work for someone else.

5C. Can you suggest other analyses that might be done with the data? None that would be relevant to the needs of the Agency in conducting this study in the first place. What I might recommend is that some thought be given to a variability-uncertainty "surface", showing how variable specific percentiles of the uncertainty distribution are across the 12 experts. This could be a form of sensitivity analysis. But many of my comments at the beginning of this review might place this suggestion into the category of being an "over analysis" of the results.

5D. Can you suggest other ways to present results? No. There are a lot of different ways to present these results, but box and whiskers diagrams are as informative as any of these, and are now well established in the literature. There is no sense in complicating matters by having a different visual format.

6. Responsiveness to Reviewers

6A. Did the elicitation address concerns of prior reviews? All of my initial concerns were addressed. I will not comment on those of other reviewers, as only they know the intent of their comments. But I can say that the methodology adopted here appears to be consistent with the concerns raised by other prior reviewers, at least those concerns with which I agree, or the Agency has responded appropriately in their summary of the comments.

6B. Were any biases introduced? No, I can see no sense in which the final methodology was biased due to any revisions prompted by prior reviewers.

7. Overall Comments. These are all included in my General Comments section at the beginning. My summary judgment is that this study represents best practice and will hold up to scrutiny at least as well as other expert elicitations I have seen and participated in. I also believe this exercise has begun a process of developing expertise in elicitation methods within the Agency- expertise that can be applied in a range of other settings. The one major limitation I see remains the problem of experts producing PDFs that are too narrow based on their significant reliance on statistical confidence intervals from empirical studies. More work is needed at the Agency to explore the historical development of confidence intervals for various pollutants so experts can be given a better sense of the degree of match between subjective confidence intervals seemingly tethered to statistical intervals and the degree of uncertainty caused by conceptual issues.

20 September 2006

Carol Mansfield, Ph.D.
Senior Economist
Environmental and Natural
Resource Economics Program
RTI International
3040 Cornwallis Road
PO Box 12194
Research Triangle Park, NC 27709-2194

By electronic mail to carolm@rti.org

Dear Dr. Mansfield,

It is with pleasure that I submit this review of EPA's Expanded Expert Judgment Assessment of the Concentration-Response Relationship between $PM_{2.5}$ Exposure and Mortality (Peer Review Draft, 25 August 2006). I am sorry that my recent illness and subsequent hospitalization has delayed my review and want to express my gratitude to you and the EPA for allowing me a few extra days to complete the work.

For the record I should note that I was the doctoral advisor of Katherine Walker when she was a student at the Harvard School of Public Health. I also have conducted consulting work from time to time for Industrial Economics and was the person responsible for recruiting Michael Huguenin to Harvard as Executive Director of the Harvard Center for Risk Analysis, once he left his position as President of Industrial Economics. I assume that you and the EPA are fully aware of these relationships.

My overall reaction to the EPA/IEc work is that it is quite well done. I reach this conclusion based on my review of – (i) the methods used to select experts; (ii) the design of the protocol, briefing book, and expert workshops, both pre- and post-elicitation; (iii) the approaches used in the individual elicitations; and (iv) the quality of the analysis and reporting. Below I provide more detailed discussions of my reaction to each of these components of the project.

Expert Selection

The selection of experts is always a difficult matter. If we knew who knew the most about the question of interest we would simply select them. Of course, we do not. And so, we have to rely on indirect measures of knowledge. There are also questions about how much to focus on expertise and whether and when to balance expertise against other desirable attributes of expert selection – balance (either disciplinary, institutional, or political); degree of sponsor control; economy; etc.

The EPA/IEc team selected experts using a two-stage peer nomination process. In the first round a group of nominators was identified based on publication counts. In the second round the nominators identified experts who they thought could best answer the question of interest. This two-stage peer nomination process is clearly in line with best practice in the field.

The EPA/IEc team went a bit further. They split the group of nominators into four subgroups and asked each subgroup to use slightly different secondary criteria in their identification of experts. I am not confident that a lot was learned from this variation but commend the EPA/IEc team for their effort to innovate. This approach produced a panel of 9 experts.

In response to concerns that the panel did not adequately represent toxicologists, the Health Effects Institute provided the EPA/IEc team with a list of ten toxicologists/clinicians. Using a random process to order the list, the EPA/IEc team augmented their original panel of 9 experts with 3 toxicologists chosen from this HEI list. While this process is a bit unusual, I understand the argument in favor of including toxicologists in the expert group.

Protocol, Briefing Book and Expert Workshops

The EPA/IEc protocol was developed quite carefully. The final protocol was based on the protocol used in the EPA Pilot Study, but was modified to reflect input from both the EPA SAB and the EPA Symposium on Expert Judgment, and then further modified after pre-testing the revised protocol with two EPA internal experts.

One of the key modifications, which was responsive to EPA SAB concerns and also to questions raised in the EPA Symposium, was to ask about the combined impact of changes in short-term and long-term PM levels in a single coherent question rather than in two separate questions. I was delighted to see this change as I had not been satisfied with the approach used in the Pilot Study to address the relationship between short-term and long-term impacts.

In any elicitation of structured expert judgment, one of the first issues faced is to determine whether to use an aggregated or disaggregated approach to elicitation. The EPA/IEc team presents a clearly reasoned rational for largely using an aggregated approach, while allowing experts to disaggregate their answers about "the probability of causal interpretation" and "the existence and location of a threshold." I believe that the rational offered for this decision is sound.

There are decision analysts who would argue that disaggregation of the "causal inference" question is inappropriate. And personally I do not believe that there is adequate evidence upon which to base judgments about the location of a population threshold. But at the same time I recognize that there are many somewhat arbitrary decisions which must be made in the design of any protocol and I believe that the EPA/IEc team's decisions are well reasoned and justifiable.

The basic three part design of the protocol is in keeping with best practices in the field. It is important to get problem definition and assumptions out first, followed by a

qualitative exploration of evidence and rationale for interpretation of evidence, before coming to the quantitative questions.

The protocol and briefing book are both quite clearly organized and written. I am impressed with the care that has obviously gone into the preparation of these materials and with the depth of background information on health, demographics and pollution levels that was provided to the experts.

The EPA/IEc team's decision to convene both pre- and post-elicitation workshops provides further evidence of the quite thorough approach that they have taken in this work. I have always found that pre-elicitation workshops are helpful in developing the protocol, educating the experts about common biases in human judgment and ways to minimize them, and in ensuring that all of the experts have a shared understanding of the available evidence. I have always felt that a post-elicitation workshop would be helpful as well and am convinced by the experience of the EPA/IEc team that this was a fruitful addition to the study design. The fact that the EPA/IEc team did not use these workshops to encourage consensus, but instead to promote reasoned discussion of the evidence is consistent with good practice in the field.

Elicitations

The EPA/IEc team elicited the experts individually using a team of elicitors including a normative expert, Dr. Katherine Walker, and a substantive expert, Professor Patrick Kinney. Both of these professionals are highly experienced and well-qualified to conduct this work. As is common in structured elicitation of expert judgment, when eliciting probability distributions the elicitors encouraged the experts to begin with the tails and work toward the median.

During the qualitative review of the available evidence and its strengths and weaknesses, the team asked each expert to write down the factors which impacted the interpretation of key studies on index cards and then to arrange these cards in order of

importance. This approach undoubtedly ensured a more structured discussion of these matters; encouraged efficiency in the interview process; and must have been helpful in understanding, and later describing, the experts' thought processes.

The EPA/IEc team also provided experts with access to web-based software which would illustrate the impact of their answers on mortality in the US and also helped them visualize their distributions during the elicitation. This approach is to be commended.

The individual interviews required a full day's effort. There is always a tension in the design of elicitation protocols between being thorough and compromising the quality of information obtained toward the end of the interview – which is usually when the quantitative elicitations occur. We have dealt with this same issue in several previous structured elicitations of expert opinion – but have never found a way to conduct reasonably thorough interviews of complex questions such as this without spending nearly a full day with each expert.

Analysis and Reporting

The EPA/IEc team followed the advise of the EPA SAB and others in their emphasis on the 12 individual elicitations and the expert's reasoning underlying each of these individual results. This focus is entirely appropriate as it encourages (perhaps forces) analysts and decision makers using the results to see the entire spectrum of expert opinion.

The report is clearly organized and well written. The executive summary does a nice job of summarizing the approach and main results of the work and provides an appropriate balance of quantitative results and qualitative interpretation. The body of the report complements this with a carefully crafted exposition of the approach, the underlying evidence, the experts' rationales, the strengths and weaknesses of the work, and the conclusions reached. Tables and figures are presented only where necessary and are clear and concise. Technical appendices include the protocol, the briefing book and other information essential for a complete understanding of the work.

The analysis of results is fairly limited, but includes sensitivity analyses intended – (i) to explore the impact of individual experts on an equal weighted combination of results; (ii) to learn whether it mattered whether experts provided parametric uncertainty distributions (e.g., normal with mean μ and standard deviation σ) or directly gave the 5th, 25th, 50th, 75th and 95th percentiles of their subjective uncertainty distributions; and (iii) whether participation in the pre- and post-elicitation workshops influenced the results.

One aspect of the report which I find particularly valuable is the EPA/IEc teams own analysis of the strengths and limitations of their work. I find that I agree almost entirely with their self-evaluation.

Summary

The purpose of expert elicitation is to produce a more complete characterization of the current state of knowledge and uncertainty about the mortality impacts of exposure to PM. The use of structured expert judgment promises to produce a synthesis of the state of knowledge and quantitative characterization of uncertainty obtained from experts, chosen in a transparent and reproducible manner, using techniques designed to minimize well-known biases and heuristics in human judgment. It is important that when we review the results of expert judgment we keep these objectives in mind.

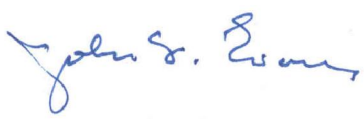
In my view the current EPA/IEc exercise has clearly fulfilled these goals. A set of experts, chosen for their ability to answer the question of interest, has been elicited using a well-designed protocol (produced after extensive external review, revision and pre-testing) with methods that are state of the art (normative expert, substantive expert, and real-time web-based computer feedback) and has produced a set of 12 individual probabilistic characterizations of the answer to the question of interest. The question itself – "What is the true percent change in annual, all-cause mortality in the adult US population resulting from a permanent 1 $\mu g/m^3$ reduction in annual average ambient $PM_{2.5}$ across the US?" – was carefully developed and, when considered in the context of the extensive material about background concentrations, population characteristics,

and other assumptions, passes a clairvoyance test. Pre- and post-elicitation workshops were conducted with the goals of – (i) informing the experts about biases and heuristics in human judgment, (ii) development of the final protocol; (iii) promoting thoughtful discussion of the evidence and its strengths and weaknesses; and (iv) reviewing the elicitation results. The results were presented in a clear and concise manner which emphasized the 12 individual distributions and the experts' rationales underlying these.

There are aspects of the work which I might have conducted differently – for example, I am a bit uneasy about the disaggregation of the "causal inference" question and I believe that the evidentiary basis for determining the location of a population threshold for PM mortality is too weak to imagine that there are "experts" on this question. However I believe that on balance the EPA/IEc team has done extremely well in designing and conducting this expert elicitation. It is inevitable that in the design of a protocol for such a complex question that many somewhat arbitrary decisions must be made and, while I disagree with certain choices, I believe that the EPA/IEc teams choices are well reasoned and defensible.

In closing, let me say that as a member of the NAS committee which urged the EPA to improve its treatment of uncertainty in dose-response, as a consultant to the EPA SAB panel which reviewed the EPA Pilot Study, and as a scientist who has conducted several expert elicitations I believe that the EPA should be quite proud of the excellent work reflected in the report Expanded Expert Judgment Assessment of the Concentration Response Relationship between $PM_{2.5}$ Exposure and Mortality. It is absolutely first rate and sets a fine example of what can be done with this approach. There is no question that the characterization of knowledge and uncertainty about PM mortality effects provided by this work is far superior to any previous analysis of this topic.

Most sincerely,



John S. Evans, Sc.D.

One Lookout Lane ● Portsmouth, New Hampshire ● 03801

p.s. – In future elicitations I would encourage the EPA to reconsider their view about "calibration." While I understand their initial reservations, my own experience and that of anyone who has conducted expert elicitation is that while all experts may be roughly equivalently well grounded in the science of interest, some are much better at providing quantitative representations of their knowledge. Unless some effort is made to capture these differences by using calibration questions, there is the risk that any combination of judgments (whether formal or informal) will be inappropriately influenced by judgments provided by experts who are not able to provide good probabilistic characterizations of their own uncertainty.

One Lookout Lane ● Portsmouth, New Hampshire ● 03801

2006 September 22

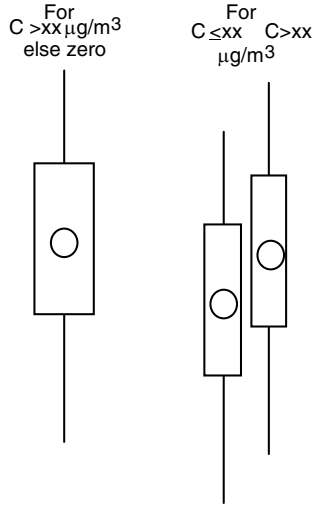
Review by Granger Morgan of Expanded Expert Judgment Assessment of the Concentration-Response Relationship between PM2.5 Exposure and Mortality

This is a first-rate piece of work, done very carefully, and very much at the same level of procedural and methodological care as the best work in the field. In addition to finalizing this report, EPA should encourage the authors, in the strongest possible terms, to publish the results as a refereed paper. I would recommend the peer-reviewed policy section of *Environmental Science & Technology* (*ES&T*). Alternatively *Risk Analysis* would also be appropriate, although more environmentally-oriented readers would see it in *ES&T.*

The basic procedure outlined in Exhibit ES-1 is very appropriate. It is particularly good that it was possible to get (most of) the experts together both before and after the face-to-face interviews. This has rarely been the case in other elicitations and is an excellent development.

The survey and background briefing materials reproduced in the appendices look to me to be of very high quality.

In presenting the summary results, I think one could do more to get them all on to the same plot. I discuss below one strategy for dealing with getting experts in Groups 1 and 2 on the same plot. For box plots that involve a threshold one could place that right on the upper or lower end of the box. Similarly, one could display the piece-wise distributions side by side:

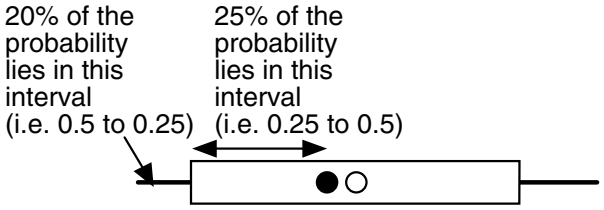


As it is, one has to read a bunch of fine print in footnotes and compare across figures.

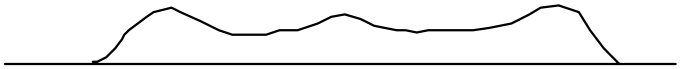
While there is a discussion of the views of different experts, it is hard to relate them to some of the big differences in the plots. It might be nice to identify a few of the more obvious disagreements in the plots and discuss them explicitly. For example, why do experts E and most others have little or no overlap? The same is true of experts K and G who hardly overlap, etc. Is this the result of very different readings of the same science, different views about how the biology works, etc? Perhaps it is all in the detailed discussion (or the meeting transcripts), but I did not find it easy to pull it out.

It looks like there is a problem with at least a couple of the box plots. If the distribution were flat, the distance from the median to the end of the box should be 25/20 times the distance from the end of the box to the end of the whisker. Since the distribution is presumably not flat, then the whisker should be even longer.

Consider the plot for expert B that looks roughly like:



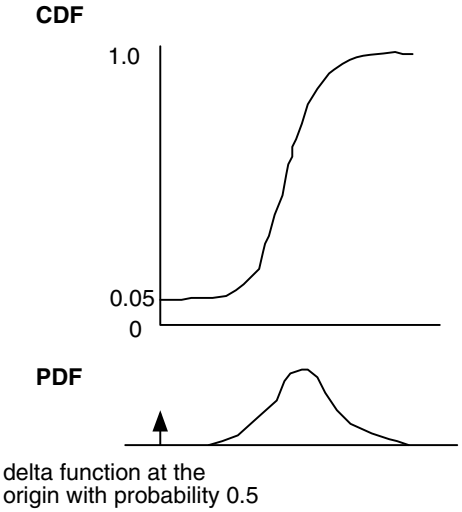
That implies a PDF that looks roughly like this:



This is probably a consequence of how the distributions were elicited. We recently fell into the same trap when (for the first time) we asked experts to just give us box plots.

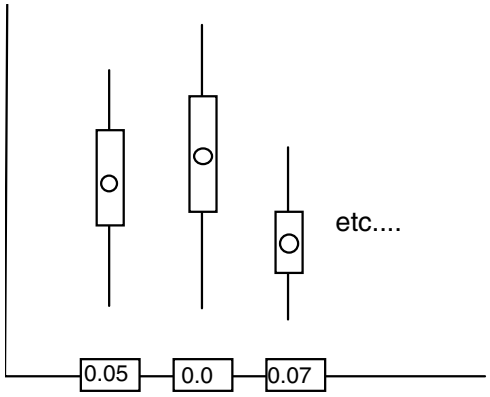
In this case, it looks like the problem is serious only for a few of your experts. Summary box plots for B and L look problematic. See especially the bottom of the left plot for B in Exhibit 3-12. You could simply go back to them, point out the problem, and ask them to adjust either the length of their whiskers or boxes since they probably did not want to imply a multi-hump distribution.

I am not confident that I understand how experts "incorporated the likelihood of a causal relationship directly in their distribution." If I think the probability that there is NOT a causal relation is 0.05 and that there is a causal relation is 0.95, then presumably my total distribution looks like this, with a delta function of strength 0.05 at the origin:



What does it mean to incorporate the 0.05 or the 0.95 in the right hand distribution?
If I am absolutely confident that there is a causal relation, presumably my distribution gets a bit higher (the integral must grow by 0.05). I guess that means that the location of the ends of the box and ends of the whiskers move a bit (although the median and mean should not change). Is that what is being talked about? Looking at the values in Exhibit 3-6, I guess this should hardly make a noticeable difference for any expert except G and K, both of whom gave conditional distributions anyway.

Perhaps one could recompute the boxes for those that are "combined" and plot all 12 with the same set of box plots by simply reporting how much probability is at the origin in a box at the bottom, e.g.



Or, am I simply not understanding correctly?

It is interesting to note that Expert G, who is the most confident (i.e., tightest distribution), made the fewest literature references in the conditioning equations (Exhibit 3.3). Expert F, who was the next most confident, made only 4 literature citations. In contrast, Experts A and B who were the least confident (i.e., widest distributions) offered 9 and 10 literature citations respectively.

The interview protocol actually asked experts for upper and lower bounds as well as for 5 and 95% confidence intervals. Unless there was a problem with these data, it might be interesting to report them as well.

I have not had time to review the 12 individual summaries on the CD. Sorry but I have been swamped.

IN SUMMARY:
Despite these minor problems, most of which I believe can be easily fixed, this is a first-rate piece of work. Both EPA-OAQPS and IEc deserve warm congratulations for a job well done, which others across the Agency would do well to emulate.

Review of: "Expanded Expert Judgment Assessment of the Concentration-Response Relationship between $PM_{2.5}$ Exposure and Mortality," Draft of August 25, 2006, prepared by Industrial Economics Incorporated (IEc) for the Office of Air Quality Planning and Standards, US Environmental Protection Agency

Review by:

D. Warner North,
President and Principal Scientist, NorthWorks, Inc.,
1002 Misty Lane, Belmont, CA 94002-3651,
and
Consulting Professor, Department of Management Science and Engineering,
Stanford University, Stanford, CA 94305.
Tel: 650-508-8858, Fax: 650-591-2923,
e-mail: northworks@mindspring.com

September 25, 2006

Summary:

This reviewer commends EPA for commissioning an innovative and valuable exercise in formal elicitation of expert judgment on an important and complex environmental health issue. I also commend the contractor team led by Industrial Economics, Incorporated (IEc) for its fine work in carrying out their assignment and for preparation of an excellent report.

There are a number of aspects of this work that deserve careful evaluation and interpretation. Assessment of expert judgment in the form of probabilities is no panacea for dealing with uncertainty – it is rather a means to describe uncertainty in quantitative terms, and to explore the extent of agreement and disagreement among experts and the rationale for these areas of agreement and disagreement. The application to the concentration-response relationship for $PM_{2.5}$ is especially challenging, because of the complexity of particulate matter, the time-dependent nature of exposure patterns, the problem of assessing individual human exposure from outdoor air monitors, the large and complex set of epidemiological investigations carried out to date, and the suggestive but not yet conclusive information on the biological mechanism(s) linking $PM_{2.5}$ exposure to increased mortality. The material in this report should serve to stimulate further dialogue and discussion within the expert community, and between this community and decision makers in the US EPA and other federal agencies involved, such as OMB and Congress. It may be especially valuable in motivating additional research to improve understanding of the concentration-response relationship between $PM_{2.5}$ and mortality. Additional research leading to better understanding should enable better decisions in managing human exposure to $PM_{2.5}$ in ambient air.

Responses to Questions Posed to Reviewers:

1. Selection of Experts.

A. *Was the method for choosing the experts consistent with standard practice?*
Response: For a public policy problem of this degree of importance and scientific complexity, I do not think there is enough application experience to assert that there is a "standard practice." In this reviewer's opinion, the selection process used was reasonable but might have been improved by additional effort to expand the diversity in views among the experts selected. (Some effort to do this is clearly described in the report, on page 2-14.) Related questions are discussed below.

B. *Are the relevant fields represented?*
Response. Yes, to first order. But in this reviewer's judgment, there was too much emphasis on epidemiology, compared to toxicology and general medicine relevant to biological mechanism(s) for PM-induced mortality.

C. *Did the set of experts selected reflect the views of other scientists in the field?*
Response. This reviewer's expertise is primarily in decision and risk analysis and in assessment of probabilities, rather than on the substance of the science involved. However, this reviewer's previous extensive involvement with PM and related air pollutants suggests that the views of scientists in the relevant scientific disciplines are complex, and there are significant disagreements among the knowledgeable scientists. Further review and discussion within the expert community should be encouraged in order to answer the question of how representative the views are of these twelve experts, compared to the larger scientific community.

I have considerable concern with selecting experts based on who has the most/most cited publications in a selected area of published scientific research, such as adverse health impacts from ambient PM exposure. This may result in a bias toward selecting those who believe these health impacts are large, versus those who are more skeptical of the research results obtained to date, and particularly the attribution of causality. The scientific tradition of peer review tries to avoid having scientists judge their own work, or having a small group of scientists draw conclusions on the importance of their collective work. Such conclusions may be disputed by others not involved in the work, and the review process attempts to set up checks and balances on making such conclusions, to assure that these conclusions are supported by available scientific information.

In a corporate context, the problem of selecting experts can be referred up to senior management, such as the CEO, members of the Board, and corporate officers. They may not always agree. But there is strong motivation to do a good job in selecting the sources of judgment on which the corporation makes

decisions, especially when these decisions will have a big effect on its financial future. Often the corporation uses both in-house experts and experts brought in from outside. Experience has shown there is great value to including skeptics and outliers who will challenge an internal "company party line." Failure to include outsiders often results in overconfidence, sometimes by a process of "groupthink" where internal experts may reinforce each other's poor judgment and/or try to please senior management, telling the senior managers what they want to hear, and failing to convey disagreements and information on potential problems.

In a public policy context, the problem of expert selection becomes harder, because no senior executive has ongoing overall responsibility (our federal government has lots of checks and balances among its three branches, and the leadership for two of these branches changes via elections.) Further, the outcomes from decisions involve much more than financial gain or loss. In such situations resolving of differences among experts is often done through discussion and debate, sometimes within an agency, sometimes within the expert community, and sometimes in a noisy and emotional political, or a legal, public process. Which experts do you trust – and what policy based on the selected experts' judgment would you like your government to follow? Our democratic society has many areas of disagreement on matters of public policy, especially where the informational aspects are complex and uncertain.

I would have preferred fewer from the group of epidemiologists well-known from their studies on PM, and more experts on potential mechanisms, such as cardiovascular specialists and toxicologists with knowledge of how substances cause inflammation and consequent cardiovascular damage. I would have preferred at least one scientist who is well-known to be a skeptic and critic of viewpoints broadly held within the group of experts that were selected. I am concerned that the degree of agreement in this exercise may not be representative of the diversity of viewpoints within the broader scientific community. I would have liked to see more independent reasoning, such as by expert K. But I regard this exercise is an excellent start, and that it has documented a scientific basis for judgment that even low concentrations (7 $\mu g/m^3$) of $PM_{2.5}$ are very likely to result in significant increased mortality. Scientific progress often occurs from critical review by scientists outside a particular effort, both in the journal peer review process and in the response of other scientists after publication occurs. This critical review process should be encouraged here, both from experts in the air pollution health effects and from experts (such as myself) with a background in risk analysis and probabilistic elicitation.

D. *Was the number of experts appropriate given the topic covered by this elicitation and the numbers of studies and experts on the topic?*
Response: Twelve experts seem like an appropriately manageable number for an exercise of the type that has been conducted. The goal of the exercise should be to characterize the diversity of responsible and well-informed viewpoints within

the relevant disciplinary areas – on this extremely complex topic of great importance for public health. What has been done seems commendable as a major step forward in advancing EPA's incorporation of uncertainty into risk assessment for an important air pollutant, as recommended in [1]. But further review and involvement of additional experts seems highly desirable.

2. Design of the Elicitation Protocol

A. *Did the elicitation cover all the topics relevant to PM mortality?*
Response: This reviewer believes that the team did a commendable job in organizing the very complex set of topics – especially in separating short-term and longer term exposure effects -- to assist the experts in making their judgments about the relation of exposure to mortality. Perhaps some might judge that the selection of topics and especially the emphasis on long-term (annual average) exposure should have been done differently. Again, further review and discussion may lead to improvements from what has been done in this exercise. Comparison with other exercises such as [2] will be useful.

B. *Were the topics adequately described to the participants (eliminated ambiguity?)*
Response: Again, this reviewer judges that the team did a commendable job, especially in avoiding ambiguities and potential failure of the "clairvoyant test" in the way that questions were presented to the experts. Further review and discussion among the expert community may identify areas in which further improvements should be made.

C. *Do you think that word choice, structure, or the order of the questions affected the quality of the results?*
Response: Of course the word choice, structure, and order of questions affected the quality of the results, as is the case in any communication process. This reviewer is impressed that the team did a commendable job on a challenging problem.

D. *Did the protocol design adequately control for heuristics and biases in the process?*
Response: Adequate control over heuristics and biases is extremely difficult, especially in a situation as complex as this one. This reviewer believes the team accomplished a commendable effort, given the time and resources available and the experts with whom they were working.

3. Background Materials/Briefing Book

A. *Were any materials missing that should have been included in the Briefing Book? Should any materials have been excluded?*
Response: The process of building a briefing book of relevant materials is an important one, and whatever is done may be criticized as having made errors of

omission or as including unimportant or inappropriate material. The experts chosen and others in the community must be the judge of the adequacy of this process. In my judgment the team did a commendable job, especially in documenting what they did and the interviews with the experts.

B. *Were any biases introduced given the set of materials provided to the experts?*
Response: Again, the selected experts and their scientific colleagues must be the judge of how well the materials selected for the Briefing Book summarized the very large amount of relevant scientific information available on the relation of $PM_{2.5}$ exposure to mortality. The careful documentation of the basis for the experts' judgments seems a highly commendable aspect of the process that was carried out. Others can then judge whether there were important omissions, or an over-reliance on materials thought to be of questionable validity.

4. Elicitation

A. *Were expectations of the elicitation process effectively communicated to the participants prior to the interview process?*
Response: This reviewer judges that the team did a commendable job in what is always a difficult process with expert participants, many of whom had not been through a probability elicitation exercise of this type before.

B. *Was adequate training provided to the participants prior to the elicitation?*
Response: As with question A above, training expert scientists in making judgments in the form of probabilities can be difficult, in large part because expert scientists with advanced training in statistics often assume they are already good at making such judgments. These experts may not have the patience and willingness to learn through training so that they become better in making such judgments. Experience with the elicitation process and critical review from colleagues may improve their performance over time. This reviewer judges that the team did a commendable job given the circumstances under which they had to operate.

C. W*as the pre-elicitation workshop properly conducted (based on the description provided in the report)?*
Response: As with many of the questions above, including A & B in this section, there are no absolute standards. In this reviewer's judgment, the team did a commendable job.

D. *Was the length and format of the interview appropriate?*
Response: Once again, this is a judgment call, and tradeoffs must be made given the time and resources available – and the expert participants' patience, as well as their time. The test of adequacy and appropriateness should be that each expert believes that his/her judgment has been adequately captured in the probability distribution resulting from the interview. This test appears to have been met, and the basis for the expert judgment has been documented in write-

ups of the interviews. This attention to documentation of the interviews is especially commendable.

E. *Were the tools used during the interview process acceptable (i.e., use of a domain expert and expert in elicitation methods, web link to two additional observers/recorders, transcription and summary of the interview, cards for key questions, electronic visual [aids] of expert's distribution as immediate feedback to allow for adjustments, etc.)?*
Response: The interview is a communication process leading to a summary of expert judgment in the form of a probability distribution. Whatever tools and visualization aids help this process are acceptable, providing they do not introduce misunderstanding or bias – and none of that seems apparent here. Documentation of the interview can be very helpful in understanding the reasoning leading to the expert's probability judgments, and in understanding why different experts might disagree. The team has done a commendable job especially on this aspect in the interview documentation. Even more discussion summarizing the similarities and differences in the experts' reasoning would be a useful and informative addition to the main report text.

F. *Was the interaction and feedback after the elicitation appropriate?*
Response: Yes. In this reviewer's experience it is always appropriate to check with the expert after completion of the interview process and passage of some time, so that the expert can verify that he/she agrees with the resulting probability distribution.

G. *Were the summaries of each interview adequate and appropriate?*
Response: While it is hard to make this judgment accurately without being present at the interview, this reviewer is impressed with the set of interview summaries he has read.

H. *Was the post-elicitation workshop properly conducted (based on the description in the report)?*
Response: Again, this reviewer believes the team did a commendable job in a first-of-its kind exercise (at least in recent years) within the air pollution health effects community advising EPA, on a particularly challenging and important problem.

5. Summary of Findings and Final Study Report (IEc, 2006)

A. *Are all of the essential elements included in the report?*
Response: This reviewer has not identified important omissions.

B. *Is there adequate information in the report to understand how the interview went and the issues that were addressed during the interview?*
Response: The interview summaries provide excellent information on what was addressed during the interview and the expert's thought processes in providing

judgments in the form of probabilities. The main body of the report summarizes this information in a way that will be appropriate for most readers, but some readers will wish to have access to the interview summaries. As noted above, even more effort to mine the insights out of the interviews on similarities and differences in the experts' reasoning would be useful to readers, who must otherwise dig this out through diligent reading of a great deal of appendix material.

C. Can you suggest other analyses that could have been done with the data?
Response: I would like to have seen some overall calculations of expected decreases in mortality with the individual distributions. If not all 12, I would have like to compare at least the extremes, such as those for experts E and K. I also would have liked to see illustrative value-of-information (VOI) calculations using EPA standard methods used in regulatory impact analysis (RIA) for valuing mortality (or, increases in human life span) plus alternatives for reduction in emissions with an estimate of the cost. These calculations showing implications of the distributions from the individual experts would give useful perspective on the extent of the differences among the experts. I did not find the sensitivity analysis by removing one expert at a time very instructive on the extent or importance of the individual differences. (Reference: last bullet, ES page ix; bottom paragraph 4-2, second bullet, page 5-2, Appendix C)

D. *Can you suggest other ways to present the results (e.g., other than box and whiskers?)*
Response: Box and whiskers plots seem to communicate well to non-technical readers. Plots of probability distributions (in cumulative form or as probability density functions) are less effective – and these are in the interview summaries. Given the volume of results for 12 experts being presented, I think what has been done is adequate. Others may have some good suggestions for how to improve the communication of insights from this exercise. I lean towards a better summary in words of how the available scientific data and judgment about mechanisms for mortality support the numerical judgments, rather than focusing solely on presentation of the numerical judgments. I note the advice I and others offered several years ago from HES cited in the quote on the top of page 2-24. I reiterate part of it with emphasis added: "to carefully examine the set of individual judgments **noting the extent of agreement and disagreement, to thoughtfully assess the reasons for any disagreement**, …" This has been done well in the document, but it might be even better.

6. Responsiveness to Reviewers

A. *Did the elicitation adequately address the concerns and comments from the peer reviewers of the pilot elicitation?*
Response: I did not observe that there were important concerns or comments that were ignored. I support the comments from the pilot peer reviewers that it was inadvisable to combine the distributions from different experts into a

composite distribution, and I endorse the decision made by the project team described at the bottom of page 2-22 "not to incorporate a calibration component in this study." A related discussion is on page 2-24 below the quote, and I concur with that discussion also.

There are other points of view. Some in the field of probability elicitation prefer to use calibration questions to evaluate how well experts can judge uncertainty in situations where the answers are known, and use this information to weight expert judgment. Such methods have recently been used by John Evans and colleagues in assessing the mortality impacts of PM from oil fires in Kuwait, based on an elicitation exercise involving six European air pollution health experts [2]. The result of these two elicitation exercises should be compared, recognizing that the combustion-generated particulate matter from the oil fires is a different composition than the $PM_{2.5}$ from a mixture of sources in the United States.

B. *Were any biases introduced given the changes made to the protocol as a result of the pilot?*
Response: I did not note any that I thought were a matter of significant concern. My greatest concern is the selection of the experts, discussed above.

7. Overall Comments

A. Overall, how does EPA's elicitation compare to best practices/acceptable practices for a defensible expert elicitation?
Response: I believe the IEc contactor team has carried out its assignment well, and that the report is an excellent product and a good step forward in EPA's process of risk assessment for air pollutants, responsive to the recommendations in the 2002 National Research Council Report [1], especially its Chapter 5.

I think that further work and a great deal of discussion is needed within the broader scientific community and among the stakeholders in air pollution risk management before EPA begins to use such elicitation results as the basis for big decisions, such as setting National Ambient Air Quality Standards (NAAQS) – with multibillion dollar implications for the US economy and similarly large implications for public health. Probabilistic assessments may be very useful in explaining to Congress and the public that there are not sharp thresholds for onset of health effects, and that the value of reducing uncertainty through lengthy and costly targeted research programs may be far greater than the large cost of these programs. So I am concerned about the word "defensible" in the question – against whom and in what context? I would not like to see EPA take this report into a federal courtroom and cite it as a principal basis for the EPA Administrator's decision in setting a NAAQS for $PM_{2.5}$.

A few days ago, on September 21, the Administrator's decision was announced on setting both a daily and an annual average standard for $PM_{2.5}$.

Such decisions are controversial: one decision overrode the recommendations of EPA staff and scientists on EPA's Science Advisory Board – and from past history, litigation of these decisions should be expected.

B. *What are the strengths and weaknesses of this elicitation?*
Response: This exercise has been done commendably well, and the documentation in the report is excellent in explaining what was done and why. In my view, the potential weakness of this exercise and others like it is that it can be taken out of context and asserted to be more than it is – an exploration of uncertainty based on the judgment of twelve experts, selected from the air pollution health effects research community, to provide guidance to EPA and others concerned about risk assessment and management of $PM_{2.5}$. I am particularly concerned about efforts to use the probability numbers from this exercise without careful consideration of the scientific information and judgment that lie behind these numbers. I would like to see calculations of expected mortality reductions from changes in $PM_{2.5}$ levels, and VOI calculations, for each of the 12 experts – these calculations may provide illuminating summary information for those not yet persuaded of the importance of airborne $PM_{2.5}$ as a public health problem in the US and much of the rest of the world.

However, I would like to be sure I am on record as urging resistance to the temptation toward making specific decisions based on this type of cost-risk-benefit numerology without shared understanding of what the numbers represent. We need to improve the understanding of our leaders and our citizens about hard problems, where there are opportunities for decreasing risk with large consequent health benefits, but lots of cost needed to realize these benefits. Cost-risk-benefit thinking can be a useful **framework** to achieve improved understanding and to build consensus on what we as a society should choose to do. But it will be highly controversial and maybe even counter-productive if cost-risk-benefit calculations are offered up as a **formula** for decision making in situations involving uncertainty, complexity, and ambiguity -- as is the case with regulation of PM. We need to pursue shared improved understanding, not complex calculations. The calculations are a means toward achieving better shared understanding.

PM is a complex mixture of different chemicals from a multiplicity of sources – and this complexity is not yet reflected in the elicitation, but rather all types of $PM_{2.5}$ from the various sources are lumped together and treated as equivalent. Hopefully, information will emerge from research that will allow future elicitation that will distinguish PM, based on particle size, chemical composition and perhaps other characteristics that affect PM's impact on morbidity and mortality. And morbidity impacts from $PM_{2.5}$ have not been included in this present elicitation exercise.

I shall conclude by quoting from the recently released report from the National Research Council on Health Risks from Dioxin and Related Compounds [3]. On page 40, this report states "…EPA should continue to treat risk assessment as a process. In this context EPA should expect to continue to iterate and improve on the assessment over time as new information becomes available. However, instead of producing and continuing to add to massive reports, EPA should consider a … structure that will allow it to focus its reports on new information that drive the quantitative estimates of risk, rather than on cataloging all information." This advice to EPA seems equally applicable to $PM_{2.5}$ and to expert elicitation as a means of developing quantitative estimates of risk.

More Detailed Comments, Minor Corrections Needed.

This is a short list. The report is already very well edited and largely free of grammatical and typographical errors.

Page 2-22. last sentence, first paragraph. This sentence is confused and seems wrong as written. Perhaps material was inadvertently omitted. What is the "expected 'true' median of the predicted value?" Was the intention to describe "unbiased" as implying that that the mean of median estimates of values should lie close to the median of "true" values, over a set of many predictions of values uncertain to the expert but known to the elicitor, like the example of the height of Mt. Everest?

Page 3-22, 4th line from the bottom (excluding heading): "expert" should be plural: "… most expert**s** indicated …"

Page 3-23, last paragraph. Who is the other expert who expressed stronger reservations about the plausibility of causality, besides K? Could you identify this expert by letter? Is it G? It would help a reader to find the interview record.

References:

[1] National Research Council, *Estimating the Public Health Benefits of Air Pollution Regulation*, Washington, D.C.: National Academy Press, 2002.

[2]. Jouni T. Toumisto, Andrew Wilson, John S. Evans, and Marko Tainio, "Uncertainty in Mortality Response to Airborne Fine Particulate Matter: Elicitation of European Air Pollution Experts," *Reliability Engineering and Systems Safety*, special issue on expert judgment, in press for 2006.

[3] National Research Council, *Health Risks from Dioxin and Related Substances,* Washington, D.C.: National Academy Press, 2002.

Review of:

Expanded Expert Judgment Assessment of The Concentration-Response Relationship Between PM2.5 Exposure And Mortality

Prepared by:

Dave Stieb
Healthy Environments and Consumer Safety Branch
Health Canada
269 Laurier Ave. W.
3rd Floor, 3-029 PL4903c
Ottawa, Ontario K1A 0K9
Phone: (613) 957-3132, Fax: (613) 948-8482
dave_stieb@hc-sc.gc.ca

September 18, 2006

I first provide answers to the specific questions itemized in the charge to reviewers, and conclude by highlighting specific strengths and weaknesses.

Selection of Experts

Was the method for choosing the experts consistent with standard practices?
The selection method was appropriate and transparently described in the report.

Are the relevant fields represented?
Epidemiologists and toxicologists were suitably represented and it was reasonable to expand beyond the original list of nominees when it was evident that toxicologists had been excluded. One field which was not represented was science assessors/evaluators i.e. government staff who are responsible for the review and evaluation of evidence in the development of standards. This group could have brought an additional perspective, and a broad familiarity with the evidence, to the exercise.

Did the set of experts selected reflect the views of other scientists in the field?
This is somewhat difficult to evaluate. There was certainly a range of views, but there appeared to be a wider range, particularly towards the low side, during the pilot, even though the number of experts was smaller. As indicated in the report, this could also be attributable to the emergence of new evidence.

Was the number of experts appropriate given the topic covered by this elicitation and the number of studies and experts on the topic?
The numbers seemed reasonable.

Design of the Elicitation Protocol

Did the elicitation cover all the topics relevant to PM mortality?
It would not be possible to cover all the topics, but the most important ones in terms of quantifying the risk and its uncertainty were covered. There was also sufficient opportunity for the participating experts to raise topics which were not pre-identified as part of the protocol.

Where the topics adequately described to the participants (eliminated ambiguity)?
All topics seemed to be adequately described, with the exception of the separation of the probability of a causal relationship and the elicitation of the quantitative distribution of risk. Even in the post elicitation workshop, there seemed to be remaining ambiguity in the experts' minds about this issue (see general comments at the end of this review).

Do you think that the word choice, structure, or the order of the questions affected the quality of the results?
All of these elements of the protocol appeared to have been thoroughly considered and debated, and benefited from the EPA workshop and pilot testing to make further refinements.

Did the protocol design adequately control for heuristics and biases in the process?
In general, I believe the protocol would deal with these effectively, but the report would benefit from a more specific discussion of how specific elements of the protocol dealt with these issues.

Background Material/Briefing Book

Were any materials missing that should have been included in the briefing book?
Obviously there's a limit to the number of items which can be included in the briefing book. Also, simply providing more material will not necessarily result in a more informed elicitation. However, consideration might have been given to the following additional items:

McMichael AJ, Anderson HR, Brunekreef B, Cohen AJ. Inappropriate use of daily mortality analyses to estimate longer-term mortality effects of air pollution. Int J Epidemiol. 1998 Jun;27(3):450-3.

Burnett RT, Dewanji A, Dominici F, Goldberg MS, Cohen A, Krewski D. On the relationship between time-series studies, dynamic population studies, and estimating loss of life due to short-term exposure to environmental risks. Environ Health Perspect. 2003 Jul;111(9):1170-4.

Rabl A. Interpretation of air pollution mortality: number of deaths or years of life lost? J Air Waste Manag Assoc. 2003 Jan;53(1):41-50.

Stieb DM, Judek S, Burnett RT. Meta-analysis of time-series studies of air pollution and mortality: update in relation to the use of generalized additive models. J Air Waste Manag Assoc. 2003 Mar;53(3):258-61.

Should any materials have been excluded?
No.

Were any biases introduced given the set of materials provided to the experts?
I think the experts would have gravitated towards the same basic set of materials as the sources of their responses, whether or not they were included in the briefing book. So I think it's unlikely that any biases were introduced by these materials.

Elicitation

Were expectations of the elicitation process effectively communicated to the participants prior to the interview process?
There appeared to be ample preparation of the experts, particularly with the addition of the pre-elicitation workshop, compared to the pilot elicitation. The post elicitation workshop also provided you of an opportunity to further communicate expectations, and provided an opportunity for clarification.

Was adequate training provided for the participants prior to the elicitation?
See previous question.

Was the length and format of the interview appropriate?
Yes.

Were the tools used during the interview process acceptable?
The web link, cards for key questions, and use of electronic feedback were all worthwhile additions to the protocol relative to the pilot. These would appear to lend additional validity to the elicitation.

Was the interaction and feedback after the elicitation appropriate?
Yes.

Were the summaries of each interview adequate and appropriate?
Most of the summaries clearly described the thought process employed by each expert in formulating his quantitative estimates and adjusting for various factors. For instance, in most cases it was clear that source x provided a quantitative estimate of y percent, and then this was adjusted in a particular direction by z fold. It was also informative to see examples of how the electronic tools were used to assist the experts in constructing their distributions. In a few cases, I could not reproduce some or all of the precise numerical steps in the elicitation. Expert A, for example, drew from the NMMAPS as a lower bound on a 95% confidence interval. However, the result quoted, 0.4% per 10 µg/m3, is based on PM 10, and it's not clear if or how this was adjusted to PM 2.5. Also, it is reported that "He ultimately identified the NMMAPS estimate, 1.004, and the 1.37 relative risk from the Laden et al. 2006 as plausible lower and upper bounds, respectively on a 95 percent confidence interval for the mortality effect of a 1 µg/m3 change in $PM_{2.5}$.", which should actually indicate that these are both for a 10 µg/m3 change. Further, if I take ln (1.37)/10 µg/m3 = 0.032, I get a 3.3% increase in mortality per µg/m3, as the upper 95% confidence interval, whereas the reported value is 2.9. As another example, the summary indicates that, "For the 5th percentile, Expert B adjusted downward from the 50th percentile to account for a real SO_2 effect and some residual confounding. Expert B determined the 25th percentile for Range 2 by reducing the 50th percentile effect estimate downward for the SO_2 effect alone." The values for the 5th, 25th and 50th percentiles (0.2, 0.5 and 1.2) are reported in a table, but the exact thinking behind adjusting from 1.2 to 0.5 and 0.2 is not specified and it doesn't conform to round adjustments such as reducing by 50%.

Was the post elicitation workshop properly conducted?
Apparently yes.

<u>Summary of Findings and Final Study Report</u>

Are all the essential elements included in the report?
Overall, yes, however additional discussion is needed of the relationship between quantifying the magnitude of mortality risk and probability of a causal association,

particularly as it pertains to applying the elicitation results to subsequent benefits assessments (see below). Is it technically accurate to characterize a piecewise log-linear function as a spline (page 3-27), which has a specific functional form?

Is there adequate information in the report to understand how the interviews went and issues that were addressed during the interview?
Generally yes, with the exception of the issues noted above regarding instances of a lack of clarity in the interview summaries.

Can you suggest other analyses that could have been done with the data?
No.

Can you suggest other ways to present results?
The box and whisker plots were informative, as were the probability density functions and cumulative density functions provided in the summary for each expert.

Responsiveness to Reviewers

Did the elicitation adequately address the concerns and comments from the peer reviewers of the pilot elicitation?
I have not reviewed the pilot reviewers' comments in detail, and relied on the brief summary in this report. This elicitation addresses the issues of increased communication among experts and avoiding the pooling of expert judgments. However, it's unclear to what extent the issues of anchoring and adjustment bias have been addressed. The authors of the report might consider adding some material to explicitly detail how they feel they have addressed this.

Were any biases introduced given the changes made to the protocol as a result of the pilot?
I believe the protocol was developed and applied in a manner which takes all reasonable measures to avoid bias. However, while on balance the addition of the pre-and post workshops is most likely a positive, theoretically it's conceivable that interaction of the experts could have resulted in the undue influence of those who might express their views more persuasively, whether or not they are more valid.

Overall Comments

Overall, how does the EPA's elicitation compare to best practices or acceptable practices for a defensible expert elicitation?
The elicitation compares favorably with accepted practices.

What are the strengths and weaknesses of this elicitation?
The elicitation protocol was very thoroughly thought through and benefited from extensive consultation and pre-testing. The addition of real time feedback measures was a particular strength compared to the pilot.

In the future, consideration should be given to participation from non-research experts, such as scientific assessors/evaluators.

The issue of separating the quantitative distribution from the probability of a causal relationship proved to be problematic, particularly since some experts chose to combine the two quantities. This makes it impossible to disentangle the two for that group, and results in ambiguity in interpreting and applying the estimates elicited from each group. These concepts were clearly interpreted inconsistently by the experts. My understanding is that on the one hand, one wants to quantify the magnitude of the association between PM and mortality and on the other, the probability that this association is causal. If this is the case, then I believe the experts need to be forced to quantify them separately. Whether these two quantities are independent also requires further elaboration. Is it conceivable that an association could be measured with limited precision, but have a high probability of being causal based on associated evidence? On the other hand, could it be measured with considerable precision, but have a low probability of causality? At least some of the experts evaluated the consistency of their distributions of the magnitude of the association relative to their assessment of the probability of causality, indicating that they conceived of these two quantities as being dependent. Four of the experts put forward 80% as a lower bound on the probability of a causal association, and three expressed a value lower than that. Thus, the interpretation of this question could have a non-negligible impact on a benefits assessment. The resolution of this issue is important and warrants further discussion in the report particularly as it pertains to application of the findings from the elicitation to subsequent benefits assessments.

REVIEW OF
IEc August 25, 2006, Peer Review Draft of
Expanded Expert Judgment Assessment of the Concentration-Response Relationship between $PM_{2.5}$ Exposure and Mortality

Prepared by
Thomas S. Wallsten

This review is structured according to the detailed topics specified in the Questions for Reviewers. Before proceeding to those, it is important to state that overall, the report is well written, well organized and highly readable. The described work is of very high quality and is carefully and thoughtfully done. I do have some concern, which I will describe below, regarding the relative emphasis placed on judgments about percent decrease in mortality per 1 $\mu g/m^3$ $PM_{2.5}$ versus on the overall concentration-response (hereafter C-R) function.

It is also important to state that I am not expert in the substantive domain of study. Rather, my expertise, such as it is, is in the area of human judgment, subjective probability encoding, risk perception and assessment, and related topics. Thus I cannot comment on issues that require knowledge of the domain, but can comment on matters of procedure and methodology.

Selection of Reviewers. The care exhibited here is noteworthy. Particularly impressive were the methods used to identify experts to provide peer nominations, the use of group-specific criteria to assure the nomination of scientists on the basis of a broad array of considerations, and the method of aggregating those nominations for the purpose of soliciting participants. This method virtually assured that the names of the most highly respected scientists would emerge. The project team (hereafter PT) also was sensitive to the disciplinary split among the nominees and took care to assure that multiple disciplines were represented.

Whether the number of experts whose judgments are encoded is sufficient depends in part on how many recognized experts there in the field and on much agreement there is among them. Twelve certainly is a substantial number. Exhibits 3-10 and 3-11, which show probabilistic judgments regarding percent decrease in mortality per 1 $\mu g/m^3$ $PM_{2.5}$, suggests considerable overlap in opinions, as does Exhibit 4-1, based on those judgments. One would like to know, as well, the overlap in judgments regarding the C-R function. Unfortunately, that is not easily deduced from Exhibit 3-9, which summarizes those results.

I cannot comment on whether the relevant fields are properly represented nor on whether the views of other scientists are sufficiently reflected.

Design of the Elicitation Protocol. The elicitation protocol is particularly impressive. I cannot comment on whether all the topics relevant to PM mortality

were covered, but I can say that the PT's efforts to encourage the experts to think broadly and carefully are noteworthy. The PT guided the experts to think about all perspectives on the issues, to consider carefully the assumptions underlying their views, to attend to the full range of evidence, and to explain the theoretical bases of their judgments. They used software to show the experts the implications of their judgments or to aid them in quantifying their judgments. These are all important steps to reduce overconfidence and to obtain accurate, internally consistent judged probabilities. Heuristics, biases, and other contaminants were controlled as well as one could possibly expect.

My one concern is that the primary quantitative question of interest is a very hard one to think about. The question was phrased as

> What is your estimate of the true percent change in annual, all-cause mortality in the adult U.S. population resulting from a permanent 1 μg/m3 reduction in annual average ambient PM2.5 across the U.S.? In formulating your answer, please consider mortality effects of reductions in both long-term and short-term exposures. To characterize your uncertainty in the C-R relationship, please provide the 5th, 25th, 50th, 75th, and 95th percentiles of your estimate. (Appendix A, page 7)

In order to provide the requested estimate as well as judged percentiles around that value, experts had to attend to and integrate a vast number of considerations, as outlined in the two paragraphs following the question. Fortunately, before responding to that question, the experts were asked to encode their judgments about the shape and parameters of the C-R function. While not an easy judgment itself, this is a much simpler one, requiring less mental integration. It appears that this judgment was encoded in preparation for the more complex one to follow, but in my opinion it may ultimately be the more useful one.

The authors provide a compelling discussion early in the report on the trade-off involved in encoding judgments about multiple issues or endpoints and mechanically combining them to yield the output of interest versus encoding judgments about the outcome directly. There are advantages and disadvantages in both directions. I happen to think that the simpler judgments may be more useful in this case, but also see the reasons for focusing on the more complex one.

Background Materials/Briefing Book. I cannot comment on the PM material included here.

Elicitation. In many ways, this elicitation was a model of how such elicitations should be conducted. The pre-elicitation workshop appears to have been excellent. The elicitation itself appears to have been very carefully done and well structured, with good use of software. The elicitation took a good amount of time, but was not unduly long. The flow of topics was logically and systematically developed. The post-elicitation feedback and subsequent workshop all look to be

very well done. Noteworthy to me was how the C-R function was encoded. The experts were not forced to assume a particular function shape, but were encouraged to select a form based on their theoretical understanding of the literature. They then provided probabilistic judgments about the parameters of the C-R function as they conceived of it. As I indicated above, these judgments seem like ones that scientists who think deeply about and work regularly with C-R functions can make.

Summary of Findings and Final Study Report. Overall, the report is well written and informative. However, I think it would be useful to include more information about the encoded C-R functions and to use them to estimate the answers that were provided directly to the main question of interest. In addition to Exhibit 3-9, the authors might consider displaying judged C-R functions with inferred confidence intervals around them so that readers can get a sense of the overlap in judgment across the variously shaped functions.

Another suggestion, perhaps infeasible due to lack of data, is to estimate answers to the primary question from the judged C-R functions. As this is not my area, I may be off-base here, but it would seem that if concentration distributions are available, they can be convolved with the judged C-R functions (i.e., with the functions corresponding to selected judged percentiles of the function parameters) to yield estimated probability distributions over the percent change in annual, all-cause mortality in the adult U.S. population resulting from a permanent 1 $\mu g/m^3$ reduction in annual average ambient $PM_{2.5}$ across the U.S.

Responsiveness to Reviewers. The report seems excellent in this regard.

Overall Comments. I provided these art the beginning of this review.