Written comments received from members of the SAB Expert Elicitation Advisory Panel

Table of contents

Comments from Dr. William Ascher 2
Comments from Dr. John Bailar 6
Comments from Dr. Mark Borsuk 8
Comments from Dr. Roger Cooke 9
Comments from Dr. Scott Ferson 14
Comments from Dr. Christopher Frey 18
Comments from Dr. Katherine Walker 24
Comments from Dr. Thomas Wallsten 28

Comments from Dr. William Ascher

Bill Ascher's comments

A. Comments on the report in general

i. Points from throughout the report: There is an inconsistency in the statements about the focus of the report. It starts out with a very strong emphasis on quantitative, probabilistic outputs, but then at the end there is an endorsement of qualitative approaches, for example, to identify the the most important risks, in order to guide further research.

The conclusion that EE has to cost a fortune seems to be based on only one of the approaches (essentially the Morgan-Henrion approach). This overly limits the application of EE. Other approaches, such as some Delphis, have been done at much lower costs than the report says are needed in general. And a good Delphi does not require a post-elicitation workshop in order to discover the experts' bases of judgment, because the Delphi elicits the reasons why experts' estimates diverge as part of the process. It is worth commissioning someone to do an analysis of whether less expensive and/or quicker approaches produce similar results to the more elaborate approaches to EE (maybe it already exists, and just a literature search is required).

The treatment of the Delphi technique seems inconsistent. Early in the report, it seems to be treated dismissively (e.g., p. 30 claims that Delphi is susceptible to social persuasion), but gets a more positive treatment later on, recognizing that it can be done remotely. The premise that Delphi seeks to foster convergence is not quite right—convergence often occurs, but the method does not force convergence.

It should be emphasized that experts are plugged into the empirical research; therefore the separation between EE and empirical research is not as sharp as the report implies, and the opportunities for blending the two are neglected.

In the early sections of the report there is so much emphasis on casting EE as a way of characterizing uncertainty it loses sight of the other purposes, such as estimating means.

The report makes many suggestions about the importance of interactions between the elicitor and the expert, to prepare the expert to do the best possible job. However, these interactions can impart biases that are difficult to detect and even more difficult to convey transparently. For example, the report recommends exposing the experts with The synthesizer may impart a bias? Providing particular journal articles can impart a bias.

ii. Comments pertaining to specific sections:

p. 11: 2nd para. 4th line: "…bring technical expertise as well as expertise" must be a typo.

p. 17: reference to Pope at al. and Laden et al. is confusing; it seems to imply that they committed some sort of statistical error.

p. 22: implies that Bayesian approaches are equivalent to subjective probability approaches. Bayesian approaches are most precisely understood as the mathematical framework for using subjective probabilities, but they are not equivalent.
Q-method is cited as an expert elicitation method, but it really is not; it typically focuses on preferences and attitude structures. If it is to treated as an expert elicitation method, more explicitness in how it can be used is needed.

p. 25: The assertion that "a description of our state of knowledge of the world" does not involve the creation of new data or knowledge is problematic. The "data and knowledge" are about the degree of agreement or disagreement, levels of uncertainty, etc. If you want to say that it is not the creation of "new empirical data," that can be said more explicitly.

pp. 26-27: last sentence on p. 26: is this about finding the parameters that are truly most uncertain, or those that are believed to have the greatest uncertainty?

p. 33: end of first full para: the assertion that probabilities are not acceptable for describing evolutionary systems because "probability contains an inherent assumption of stability within a given system" is a non sequitur, or at least unclear.

p. 45: the "5 Morgan and Henrion criteria" list only lists 4 criteria.

pp. 51-52: Section 4.3 on other methods to characterize uncertainty is not terribly useful.

p. 65, section 5.1, state that the white paper focuses on EEs involving individual interviews of the experts, but in fact there is a fairly well balanced treatment of both individual and group processes.

79: The Q method is not an EE method, though it should be mentioned that the Q method can be used to identify different factors among experts as indicators of "schools of thought" as a way of coping with the aggregation challenge; i.e., instead of averaging everyone, even if there are clusters of experts from the same schools of thought.

p. 82: The aggregation issue should not be based on the overlap criterion discussed here, but rather—if possible—by gathering the responses of experts who can be clustered as belonging to the same school of thought.

p. 91: Two of the bullets seem problematic: 1) "possible correlations with non-elicited components of the overall analysis or policy question"—it is not clear how to get this and who would be capable of doing so; 2) "text or graphics…" this is a tall task, which not only raises the price, but also could impart the biases of the elicitation team.

p. 94 on: Many of the examples in the last section of the white paper need more explanation of what is going on; the figures or tables in themselves are not very useful without more explanation.

B. Comments on the Specific Issue of Transparency

Transparency is important to 1) identify the range of expertise; 2) identify the experts' assumptions (i.e., "basis of judgment"), in order to decide when the effort is obsolete; 3) evaluate strengths and weaknesses of the study in the future; 4) assign degrees of credibility, but also to increase credibility by demonstrating that the approach was applied rigorously; 5) withstand litigation.

In terms of *what* should be transparent, it should be methodology (how the problems were framed, how judgments were solicited and combined), confidence intervals, characteristics of the experts, the assumptions they hold, and the specific interactions between elicitor and expert.

The draft white paper does a good job in urging transparency in methodology, confidence intervals, and characteristics of experts (this does not mean naming which expert made which estimates). However, the white paper should say more on how to capture the assumptions/"basis of judgment" held by the experts. It should also discuss how deepening the interactions between the elicitors and the experts make it more difficult to be fully transparent. In other words, the report should assess each method in terms of how much of the process is intrinsically a black box. E.g., the Morgan-Henrion approach of sending a person in to interview each expert may be the best approach in terms of getting the most accurate estimates, but the interaction between the elicitor and the expert is really difficult to chronicle, and constitutes somewhat of a black box. E.g., if the elicitor has to point out to the expert that the expert's responses are incoherent, this interaction and how the expert shifts his or her ground is very difficult to chronicle. A remotely-conducted Delphi, in contrast, whether or not it is the most effective approach, is much easier to describe with full transparency. It would be useful to consider creating a table that lists the aspects that can be easily conveyed transparently and those that cannot. That is, put transparency in the context of tradeoffs. In short, the more help that the experts get, the more difficult it is to document and convey all of the interactions.

It should also be emphasized that putting huge resources into chronicling the interactions may not be worthwhile in light of the impossibility, in many cases, to be fully transparent. The suggestions for documenting seem extreme in terms of time and costs involved.

In terms of how to enhance transparency, the report does a good job of urging that experts be selected by the criteria of being capable of articulating their basis of judgment and to think consistently in terms of quantitative probabilities. For parallel reasons, it is also important to select methods that are fairly straightforward and comprehensible, which are also capable of illuminating bases of judgment.

Enhancing the transparency in presentation, reportage should retain enough information to clarify bases of judgment, differences, inconsistencies.

The report covers many aspects of the information necessary to reveal, transparently, the characteristics of the participants, the procedures, the analysis, etc.

Although the report makes the correct point that the “basis for judgment” (read: the experts’ assumptions) be made transparent, the report does not explain how this can be done under conditions of fairly limited resources.

An issue related to reporting the “characteristics of the experts”: do you report just the “credentials”, or do you also report on the school of thought of each expert? Perhaps some of this can be conveyed in the description of how the experts were chosen, but the question remains.

Comments from Dr. John Bailar

EPA Draft White paper
Expert Elicitation Task Force
Comments by John Bailar re charge question 1.
February 26, 2009

There is much good material in this draft on expert elicitation (EE), and the Task Force deserves a lot of credit for it. My comments are meant to help make a good thing even better.

A committee wrote this draft, and its origins show. There is a great deal of redundancy, in places the tone is pompous, or it talks down to the reader, some seems to be an undigested agglomeration of whatever the author has been reading, and the style is highly verbose. Thorough, end-to-end rewriting is needed. This should be by a single person, or perhaps two people working elbow-to-elbow. The target should be a report of half the present length.

There should be a shift in point of view from ordinary frequentist statistics (what happens if some hypothesis is true) to a more Bayesian view (what is the probability that the hypothesis is true). This has profound implications, and would bring the draft fully in line with EE.

The authors could strengthen the draft by adding appendices with case studies (perhaps 2 or 3 of them, each 10-20 pages long) on completed EPA EEs. These should be chosen to illustrate problems rather than to display how well the process can be made to work.

A critical issue that needs more attention is the identification of elicited opinions that are simply not compatible and should not be combined (though the full range of EEs should of course be reported).

Is there any empirical evidence comparing group vs. individual EEs? If so, that evidence should be cited. If not, say so, and discuss whether EPA should support research to develop such evidence. For example, one might carry out a dozen EEs, each with 10 experts divided randomly into five who would work individually and five who would work as a group. One might learn little from each EE, but patterns over the set of EEs might reveal a lot.

Similarly, there is a need to discuss any empirical evidence comparing EEs vs. actual observations or situations. This might have to be limited to measurements rather than other matters studied by EES, but even that much would be helpful. If such evidence is scanty, say so, and consider whether EPA should support research on the matter.

There is no sharp dichotomy between experts and mere mortals (see page 31 and elsewhere), and one can find “experts” with any desired opinion. Problems in choosing experts are discussed, but need more attention.

EE processes seem to be designed to separate the expert opinions from objectified conclusions derived from them. This is wrong – the opinion-based nature of EEs should be preserved through the final results.

Finally, EE is clearly a possible way to learn about unknown matters, but it should also be presented as a good way to organize and understand whatever is already known about the matter and to identify what remains to be studied.

Comments from Dr. Mark Borsuk

Points with respect to Charge Question 3C: Problem Structure
Mark Borsuk

1. The importance of being explicit about the structural relationship among system variables and parameters should be addressed in the White Paper. It can be suggested that this might be best accomplished through the use of graphical models such as influence diagrams or belief networks. Figure 6.1 on page 96 is a poor example of an influence diagram and should be replaced. The book "Risk Communication: A Mental Models Approach" by Morgan et al. has plenty of good examples. Better problem structuring serves at least three important roles:
    a. Makes the underlying structural assumptions of the model explicit, including the level of decomposition chosen.
    b. Makes it clear what the relationship is between quantities being elicited and other variables in the system. In particular, it makes it explicit what other variables the elicited variables need to be conditioned on.
    c. Makes it clear if an elicited variable will be used to infer the values of other variables of interest, rather than being of primary interest itself.
2. The role of existing empirical data in the EE process needs to be clear. Should care be taken in the process to be sure that all experts have access to the same relevant data? Should elicitations be made with explicit reference to the existing data, including reasons for departures or inconsistencies?
3. Some discussion should be included as to what types of variables are appropriate for elicitation. The existing text on page 50 is a bit vague in this regards, and I would argue that Scenario Uncertainty and Decision Rule Uncertainty are not amenable to EE. Rather, these are ambiguities that should be resolved before undertaking the EE. Also, as we discussed in the SAB repeatedly, the focus of an EE should be on observable quantities whenever possible.
4. Consideration should be given for the occasional (or even frequent) need to allow experts to express probability judgments as ranges which represent their own imprecision or ambiguity with respect to a probability. In my experience, experts often want to do this and it is information that is important to record.

Additional References (not specific to charge question C.3):

Uncertain Judgements: Eliciting Experts' Probabilities
by Anthony O'Hagan, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow.
Wiley (September 14, 2006) ISBN-10: 0470029994, ISBN-13: 978-0470029992

Comments from Dr. Roger Cooke

Written comments on EE white paper

Roger Cooke
Resources for the Future
Feb 25, 2009

General:

Plus

- That the document appears at all
- Recognize the importance of EE in uncertainty quantification
- Recognize that subjective probability is the primary formalism for encapsulating EE
- Findings chapter 7.

Min

- Literature is very old
- Insufficient focus on applications
- No attention for performance validation
- No attention for dependence between variables
    - How to elicit dependence information
    - How to represent dependence mathematically
- Combination given very short shrift
    - No mention of scoring rules
    - No mention of EE data and expert performance
- The process of transferring expert distributions on observable quantities to parameters of a model (probabilistic inversion) is very important in applications and is missing
- Costing is unrealistic

There are more Mins than Plusses, but the Plusses are more important.

Detailed:

p.22 "The goal of an EE is to characterize, to the degree possible, each expert's beliefs (typically expressed as probabilities) about relationships, quantities, events, or parameters of interest." Rational decision requires utilities and probabilities. Utilities (values) are usually the province of stakeholder preference, but may also be elicited from experts. The techniques for this are different from those discussed in the white paper. EE is mostly employed for uncertainty quantification. I refer to quantification of uncertainty via expert judgment throughout as EJ. Uncertainty is degree of belief wrt the true state of the world. Things like "relationships" and "parameters of interest" are suitable objects for uncertainty quantification only insofar as they can be given operational meaning. We cannot use EJ to assess the specific heat of phlogiston, regardless of our interest in phlogiston. There are three plausible goals of an EJ exercise:

1. Census (survey of opinions)
2. 'Political' consensus (equilibrium of interests)
3. Rational consensus (pre-commit to method consistent with scientific principles – see documentation on classical model).

Pick one, and follow a method that is consistent with that goal

p.28 “Using mathematical methods to combine expert opinions relies on an assumption that the individual expert opinions are independent” This statement remains meaningless until a definition of “expert dependence” is given. Suppose two experts each give distributions to express their uncertainty about given uncertain quantities. How do you define the dependence or independence here? Do you envisage a correlation between distribution functions (as opposed to correlation between random variables)? How is that defined? Perhaps you mean the correlation between the experts’ mean values or median values? Well, these better be correlated, since the experts are assessing the same uncertain quantities. Perhaps you mean the *conditional* correlation of means or medians *given* the true values of the quantities in question. Perhaps you mean the correlation of the percentile realized by the true value in each experts’ distribution. Something like that makes sense, but it can only be assessed if the true values are known. The TU Delft EJ database allows expert dependence to be assessed, an article giving results is included with the submitted written material. The link between (in)dependence and combination is quite obscure. Clemen proposed a model for this, and its performance in terms of statistical accuracy was poor. (Kallen, M. J., Cooke, R.M. (2002) “Expert Aggregation with Dependence” in *Probabilistic Safety Assessment and Management* E.J. Bonano, A.L. Camp, M.J. Majors, R.A. Thompson (eds),Elsevier, 2002; 1287-1294.)

The EU-USNRC suite of expert judgment studies described in supporting material involved 69 experts spread over 9 panels, and totaling 15,422 expert-variable elicitations. To get the final results we must use distributions from each panel. If we did not combine the experts in each panel, should we consider each expert in panel A with each expert in panel B, etc? That would be 67 million combinations. (Cooke, R.M. and Kelly, G.N. Overview of Joint EU/USNRC Project on Uncertainty Analysis of Probabilistic Accident Consequence Codes (written for EE SAB))

p.29 “However in addition to their knowledge, each expert brings different biases (and experience) to the question of interest. Therefore EE practitioners must be cautious about aggregating expert judgments and presenting combined conclusions about EE results”. No one is against caution; however the slur on experts is unwarranted. How do you know that they are “biased” if you don’t know the true values of the uncertain quantities? We use expert judgment because things really are uncertain, and because they are uncertain, the experts will and should disagree. What justifies the word “biases”?

p.30 “According to Kieth (1996), combining judgments could be problematic because the methodological assumption is that the experts chosen for the elicitation represent the entire continuum of ‘truth’ with respect to the technical question” . If this means anything, it is wrong; there is no such assumption. How does the true value of an uncertain quantity become a ‘continuum’? If you know the ‘truth’ then why do EJ?

“…the fraction of experts used for the elicitation cannot be assumed to be proportional to the probability of that view or opinion being correct.” What does it mean to say that a subjective probability, or probability distribution is “correct”?

p.33 “In EE the expression of expert judgment as probabilities assumes that experts understand all alternatives so that their judgments can be compared.” The questions should be clear, and the experts should understand them – does the quote mean any more than that?

p.34 “Cooke(1990)” is Cooke(1991)??

p.35 Again the issue of “biased” Figure 3-1 fails to convey statistical accuracy; statistical accuracy can only be judged relative to a set of assessments with realizations. In statistics, an estimator of a parameter is ‘unbiased’ if, on the assumption that the samples are randomly drawn from the assumed distribution type, the expected value of the estimator equals the true value of the parameter. This interpretation is not available here, and hence only the perjorative senses of ‘biased” remain.

p.40 “Poor performance when the number of variables increases can be caused by many different factors’” How do you define and measure performance?

p.49 Table 4-1. No indication is given of the size of the EJ studies. The EU-USNRC study mentioned above involved 69 experts and 2036 uncertain quantities, and cost $4M (1990), including 15$K remuneration per expert. That study carried a peer review burden at least as great as anything at EPA. The cost figures in table 4-1 seem totally out of touch with the reality on the ground. A typical EJ study involving circa 100 variables should cost less than $100k. A recent EPA study of economic impacts of invasive species in the Great Lakes cost about half that (9 experts, 100 variables)

p.51 “For example an EE could be conducted to estimate the magnitude of a cancer slope factor (including that number’s uncertainty)”. The cardinal rule in our work is *ask experts only about outcome of possible (but infeasible) measurements/experiments.* What would you measure to estimate a CSF? Ask experts about the outcome of such measurements.

P 52 “other methods” “fuzzy methods” do not represent uncertainty. They are not a middle ground between frequentist and subjective methods. The fuzzy membership function has no operational meaning, and the combination rules are palpably wrong for representing uncertainty. At most they may claim to represent ambiguity.

p.52 sensitivity analysis – the discussion here is too brief to make sense. No reference to Saltelli?

p.54 “…there are suitable approaches to adjust data so that it is sufficient to support a specific decision”. I have no idea what this means, but it sounds like it’s against the law.

p. 62 Table 4-2. Same remark as for Table 4-1. Spending more money entails higher quality only if the possibility of wasting money is excluded.

p.63 “… alternatives to face-to-face elicitation may reduce costs” I strongly discourage that. If it is not important enough for the analyst’s time, it’s not important enough for the expert’s time. Tele-interviews are appropriate for surveys, not for EJ.

P65. “training” it is mentioned (p.74) affirming that it is good, but nowhere do you say what it is or how it is to be done. You *must not* use almanac items (how long is the Nile river) for training, it alienates experts, as it does not engage their expertise and does not emphasize the value of informativeness. The old SRI school believed that the purpose of training was to make experts less confident. Do not do that, it will also irritate experts. Training variables must be from the experts field of expertise. You cannot do training without knowing what constitutes a ‘good assessor’. That means saying how expert performance is defined and measured. Once you say that, it is very difficult to explain why you pay no attention to performance validation in an EJ study, or in the White Paper.

p.67 “clairvoyance test” . “This demands that all of the significant assumptions and conditions that could impact the expert’s response are well-specified”. The intention is right, but the articulation is wide of the mark. Better than saying what value an omniscient being would assign to the uncertain quantity, it is better to describe the measurement which YOU the analyst would do to determine the value of the quantity. This forces the analyst to operationalize his concepts and conditionalize them in a way appropriate to the purpose. It is very important that ALL uncertain quantities be conditionalized consistently. The notion of a clairvoyance test does not bring out this conditionalization aspect.

Example

An example from the EU-USNRC study, dry deposition velocity (ddv): It is known that ddv depends on at least 80 physical parameters ranging from the mean free path of Brownian motion to the mixing layer of the atmosphere; but it is not know *how* it depends on all these. There are many models, all of them wrong. The study in question distinguishes ddv’s according to

- Chemical species
- Surface (grass, meadow, urban, peat, forest, skin)
- Aerodynamic diameter
- Windspeed

An expert is NOT asked to build a model for ddv. He is asked something like *suppose we measure the ddv of aerosols with aerodynamic diameter = 1μ on grass with windspeed 2 m/s, what will we get, please state quantiles of your distribution*?

The expert is asked to conditionalize his uncertainty in a way conformable to the model whose uncertainty is being quantified. We all know that ddv can vary by an order of magnitude according to the species of grass. The expert is not told the species of grass, rather he is told that uncertainty arising from species of grass should be ‘folded into his distribution’. What about electrostatic potential, humidity, temperature, surface roughness, insolation, wind profile, etc, etc,etc? Any of these may be important, the omniscient being knows, but we don’t. Hence if you really demand ”that all of the significant assumptions and conditions that could impact the expert’s response are well-specified” then you might as well abort the whole thing before you begin. On the other hand, it is very important, and very difficult, to maintain a consistent conditionalization across a large study. When we elicit atmospheric dispersion and wet deposition, the conditionalization must be consistent with that for dry deposition. All of this problem gets swept under the rug by talking about ‘clairvoyance’.

How do expert's answer such an elicitation question? In very different ways. One may build a model, another may consult data, another may draw on his experience, another may average published models etc etc. In the end they all quantify uncertainty wrt a measurable quantity. We are combining these uncertainties, NOT combining their models.
P 68 – 84, The references for this 'how-to' part of the white paper are very old and of dubious relevance. For example the Clemen and Winkler reference (1985) concerns Winkler's normal error model, according to which expert's biases are 'corrected' by the decision maker, and expert unbiased estimates are combined as if they were point observations from a joint normal distribution – extremely unrealistic. "The practical implication is that, the more different the experts are, the more experts are needed"(p72) – don't understand that. The situation is more complicated, negative dependence is helpful, very high positive dependence with unequal variances can be helpful – but this model is not relevant for applications. A few other remarks

- The correlation coefficient ranges from -1 to 1, not 0 to 1 (note 12, p 72).
- There should be no discussion of the Delphi method without referencing (and reading) Sackman, H. (1975) *Delphi Critique, Expert Opinion, Forecasting and Group Processes,* Lexington Books, Lexington, Mass.
- The discussion of Cooke(1991)'s principles for rational consensus is garbled. "…the goal of accountability requires each judgment to be explicitly associated with a named expert" (p.81). No. Experts must be named, and the association of names and distributions must be preserved for competent peer review but need not be part of the open literature (see p. 82 of Cooke (91)). The reason is to shield experts from cherry-picking by interest groups in juridical proceedings, or from other interest conflicts. See EU Procedures Guide (p.27). The issue of use of expert names deserves more attention.
- The question of how and why to aggregate expert opinions reads like a section that didn't want to be there. Why not define the alternatives, compare their performance in practical applications?
- Measuring expert performance is necessary for performance based combination, but even without combination, good scientific practice demands that performance be defined and measured. Without this, EE does not rise to the level of good science, in my opinion. Spending lots of resources on peer review is not a substitute for performance validation.

Completely left out:

- Dependence assessment, it was a big deal in the EU-USNRC studies and has only become bigger since.
- Performance measurement, how, why, experience to date
- Probabilistic inversion: experts assess only results of possible – if impractical – observations. Pulling this uncertainty back onto the parameter space of a model was a big deal in the EU-USNRC studies and has only become bigger since.

The state-of-the-art anno 2000 can be retrieved from the EU-USNRC reports, but this is an active area and the baseline has moved; see more recent articles.

Comments from Dr. Scott Ferson

Scott Ferson
Applied Biomathematics
1-631-751-4350
scott@ramas.com

Answer to Charge Question C.4 and C.5

The charge question asks whether the findings and recommendations are supported by the document. If the point of this review exercise is to consider whether guidance on EE should be promulgated, however, it seems to me that the more relevant question is whether the findings and recommendations are prudent, sound, complete and sufficient.

One concern is about the tone of chapter 7. EE is one of many tools needed in risk assessments. Enshrining the method in a special guidance document seems to elevate the method to a status that is inappropriately highlighted. This runs the risk of making EE seem like a gleaming golden hammer, which, of course, makes a host of problems all seem like potential nails. There are dangers associated with such an outcome. The document does place EE in the context of peer review and the public comment process, which is helpful and entirely proper. It would be helpful to extend this contextualization to explain that EE is a tool along a spectrum of methods that address the problems arising from the paucity of empirical data. These include statistical methods for censored data, surrogacy model in general, meta-analysis, model averaging methods, and bounding methods. We wouldn't need a review of these methods, but appropriate mentions that expand and enrich the context would be useful.

There are several critical issues that were left out of this chapter (and the document as a whole). One is the issue of consistency of the elicitations from an expert. Luce (<<ref>>) quipped that, if the study of decision making in humans has proved anything, it has proved that humans have incoherent probabilities. We can expect that experts will likely exhibit inconsistent beliefs manifested in incoherent probabilities. Experts are also surely susceptible to the cognitive biases that affect human perception. A variety of checks can be used to detect and correct inconsistencies in EE. These include elementary units/dimensions checking (are multiple experts giving answers in the same units?). They also include mathematical checks that the elicited probabilities are coherent and that related quantities are mutually consistent. Elicited distribution moments, for instance, must satisfy certain consistency constraints (see Karlin and Studden 1966; Smith 1990). Correlations must be positive semi-definite. This condition is a logical requirement for meaningful risk assessments, but it is actually easy to inadvertently violate it. Birge (<<ref>>) showed that when the number of variables goes up, the probability tends to one that a matrix of correlations coefficients constructed by separate pairwise estimates (even from real data) will violate positive semi-definiteness. In addition to mathematical and logical constraints, checks based on physical constraints may likewise help to improve EE. Roger Cooke mentioned that aerial dispersion always increases downwind, so elicited beliefs that imply the opposite are probably in error. I think that logical, mathematical and physical checks such as these on elicitation results are an essential element of a best-practices program. Such checks

should probably be applied interactively and in consultation with the experts, rather than as filters that alter or omit expert statements.

The literature on cognitive biases is much richer than is indicated in the White Paper. In addition to the famous decision biases such as loss aversion, ambiguity aversion, zero-risk bias, anchoring, and availability heuristics, there are biases relating particularly to uncertainty perception such as probability misperception, the conjunction fallacy, pseudocertainty, overconfidence, base rate fallacy, and neglect of probability and a variety of other biases such as clustering illusions, gambler's fallacy, framing, regression toward the mean, the ludic fallacy, primacy, recency, and hyperbolic discounting, all of which may distort perceptions of experts and lay people alike. There is little or no evidence that people can overcome these biases, even with training, so strategies to elicit proper expert judgments correctly or to account for these biases deserves some discussion in the White Paper. For example, would Alexander Shlyakhter's discounting be appropriate for correcting for overconfidence? This field of psychometry is in flux, and some suggest an 'illusion illusion' in that these biases are to a large extent themselves misunderstandings. Gigerenzer appears twice in the reference list, but is not actually cited in the White Paper. (And a Gigerenzer paper is misattributed to Cosmides.) Glimcher's (2003) work is not even mentioned, and this is unfortunate because it looks like it will revolutionize our understanding of how humans appreciate risk and uncertainty. There is a now growing body of literature (e.g., Hsu et al. 2005) that strongly suggests that humans have two separate processors in our brains, one for handling variability that closely mirrors Bayes' rule and one for handling uncertainty that seems to focus more commonly on the worst case. Tucker et al. (2008) considered the import of some of this recent literature on risk perception and communication.

There is virtually no discussion in the White Paper of whether or how expert judgments would be combined or aggregated. This is a serious technical omission because it will determine the crux of how expert elicitations will be used in practice. Users will need guidance, for instance, on whether it is reasonable to combine probability distributions from different experts with linear pooling (a.k.a. mixture or vertical averaging of probability values) or a related method that tends to erase the disparities among experts. It is not necessary to reduce the different opinions to a single probability distribution. An alternative to linear pooling is forming the envelope of the various distributions. Such envelopes sidestep the problem of addressing the issue of dependence among experts (or canalization). They allow a robustness analysis that retains a full expression of the uncertainty among experts without forcing on analysts calculations of combinatorial complexity.

Answer to Charge Question A

The White Paper has an inordinately positive tone. For instance, it discusses the "uses and limitations" of EE as though all we care about it what one can do and what one cannot do with the method. Of special interest at this point is also *whether the method should be used*. The Peer Review Panel itself had maximal variation on the question of how often EE should be used in practice, literally ranging from "it should be used on essentially all assessments" to "it should be used only as a last resort". A balanced consideration is needed to support a broad discussion of whether and in what circumstances EE should be endorsed by the Agency. On page 114 is a list

of "advantages and limitations", as though disadvantages were excluded from the White Paper. A frank discussion of the potential drawbacks of using the method would be more appropriate and more useful for this document. There are two reasons. First, as policy consideration, we'd like to see what drawbacks there might be from endorsing this approach. Second, as fodder for future guidance, users are going to need to know about the criticisms they will likely face, and which criticisms are knee-jerk complaints and which are serious issues deserving reflection and careful justifications. There are a variety of questions that should be considered in the White Paper. For instance, is there a risk that using EE will diminish confidence in the integrity of analyses because of its controversy? How could such blowback be mitigated? Can intervariable stochastic dependencies be elicited in practice? Can concerns about inter-expert dependencies (canalization) be put to rest?

Some text to flesh out preliminary points #10 and #11

10) *The White Paper should address how EE will account for model uncertainty.*

The White Paper mentions model uncertainty in the second bullet of page 59 in a quote from the *Draft OMB Risk Assessment Bulletin*, but doesn't anything else about the subject. Because accounting for model uncertainty is such an important use of expert elicitation, and because methods available to make this accounting are limited and controversial, the White Paper should address this issue explicitly. It should emphasize that this will be a significant use of expert elicitation, and outline how it will work.

11) *EPA can use EE when it addresses two issues: (a) reconciliation EE's transparency restrictions versus FACA and related rules, and (b) reconciliation of EE with EPA's understanding and practices regarding uncertainty and variability.*

The White Paper should address two issues that may appear to potential users to represent discrepancies between EE best practices as described in the White Paper and other long-standing EPA policies. The discrepancies may well be only apparent, but the issues should be considered explicitly to establish this fact.

The first apparent discrepancy is between the general provisions for openness and transparency in rule-making that EPA observes and the uses of anonymity for experts under some elicitation protocols. Different protocols vary with respect to their degrees of transparency. Insofar as the public and peer reviewers will not be able to learn which experts made which pronouncements and how precisely the elicited probabilities were developed from individually identifiable opinions, such elicitations restrict transparency. Although the elicitation protocol and process are intended to be transparent and subject to peer review, the *beliefs* elicited from experts may (will?) also be insulated from peer review. Most members of the Peer Review Panel consider these restrictions on transparency in some elicitation protocols to be compatible with a scientific approach and some feel they are crucial to the success of the elicitation process. If the Agency uses elicitation protocols that restrict transparency, it will be promising anonymity to the experts involved. Is this anonymity compatible with FACA and related rules? Can the Agency provide legally defensible assurances to potential experts that their frank opinions expressed during

elicitation will not be traceable to the individual experts? How could assurances like this be protected in practice? Will such assurances withstand FOIA requests? Will they withstand discovery rulings in case of litigation? Must communications between experts and elicitors be marked with general-purpose disclaimers such as "privileged, draft, pre-decisional"? These questions should be resolved to the satisfaction of the authors of the White Paper and EPA staff.

The second issue that seems essential for the White Paper to address in order to decide *whether there should be EPA guidance* is whether EPA's understanding and practices regarding uncertainty versus variability can be reconciled with EE methods. Most methods to combine probability distributions from different experts reduce them to a single probability distribution (such as by linear or logarithmic pooling). This has the effect of averaging potentially incompatible theories from the different experts, as though one expert is correct some fraction of the time, and another expert is correct some fraction of the time, and so forth. Disagreement among experts is usually considered a form of model uncertainty and, as such, it is usually not handled in this way (NRC <<ref>>; Finkel <<ref>>). This apparent dichotomy should be resolved to the satisfaction of the authors of the White Paper and EPA staff.

References

Glimcher, P.W. 2003. *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics.* MIT Press/Bradford Press.

Hsu, M., M. Bhatt, R. Adolphs, D. Tranel, and C.F. Camerer. 2005. Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310:1680-1683.

Karlin, S. and W. J. Studden. 1966. *Tchebyshev Systems: With Applications in Analysis and Statistics*. Interscience, New York.

Smith, J.E. 1990. *Moment Methods for Decision Analysis*. Ph.D. Dissertation, Stanford University, Stanford, California.

Tucker, W.T., S. Ferson, A. Finkel, and D. Slavin (eds.) 2008. *Strategies for Risk Communication: Evolution, Evidence, Experience*. Annals of the New York Academy of Sciences, Volume 1128, Blackwell Publishing, Boston.

Comments from Dr. Christopher Frey

Comments by H.C. Frey, 3/2/09
General Comments:
I support that structuring should occur BEFORE meeting with experts as much as possible, and should be evaluated in a pilot elicitation. However, despite all preparations, I would expect occasionally an expert may not want to accept the structure posed by the elicitor, and may want to propose an alternative structure. There should be some flexibility to accommodate this when appropriate, as long as the Agency is still able to infer the model outcomes that it ultimately cares about (e.g., if the goal is to estimate avoided excess mortality, then one could accept alternative approaches to characterizing dose-response).
I recommend that elicitations focus on empirical quantities that are in principle measureable or observable, even if not so at the current time. Ideally, such quantities should be model independent. I don't want to be too prescriptive about this, in case there might be some reasonable exception.
The report should elaborate on the "clairvoyance test" and alternative ways of conceptualizing it. One notion is that it should be a quantity that is measureable or observable and not dependent on a model.
The EPA report states that uncertainty in "relationships" is within the scope of EE, but doesn't really address it. I recommend that EPA focus on eliciting empirical quantities and then use that information to inform model selection and parameter estimation, rather than asking experts to provide judgments regarding parameter values. Any exceptions to this approach should be discussed.
The issue of dependency among quantities should be further discussed e.g., elicitation of conditional distributions.
I agree that it is difficult/impossible to elicit correlations directly.
I strongly recommend that EPA prepare a more extensive glossary of terms that are jargon related to the topic area. The goal here is to let the reader know that the EPA means by these terms in the context of this document. Examples include "representativeness," "data gap," and many others (see list on last page of these comments). It is fine if EPA adopts glossary definitions from existing documents.
The draft white paper would benefit from copy editing to

- enforce consistent use of terms throughout
- eliminate repetition
- enforce consistency in scope (e.g., EE for uncertainty, and data gaps?)

Specific comments:
p. 3 mention the most recent NRC (2008) report on risk assessment methodology.
p. 5 one could characterize judgments regarding data gaps, but can they be “filled” by expert elicitation?
The word “parameter” is commonly misused. A parameter of a model is usually a constant that is determined through a calibration process. E.g., for a linear regression $y = a x + b$, a and b are parameters. x is the input, not a parameter.
What is “objective probability data”? The term “objective” is never defined in this document.
p. 7. ABA is not credible to weigh in on how science should be conducted, just as scientists are not credible to practice law.

p. 10. Seems to stray from expert judgment into judgments by non-experts about values/preferences. Could be more clear about this.
Figure 2-1. "expert judgment to characterize science vs. integrate values and preferences" implies expert judgment about values. Is this the intended meaning?
Figure 2-4. Define what the cryptic labels mean. In a document that talks about transparency and communication, all figures should be self-documented.
p. 18. Is the result of a probabilistic analysis considered to be based on EE even if it is not sensitive to the EE?
Figure 2-5. The experts made judgments about the C-R relationship, not about avoided premature mortality. This should be more clear.
p. 19. Second bullet – also add unknown future events.
Bottom of same page – hasn't EE been used for years nuclear power plant safety assessments?
Section 2.7. also add citation to IPCC (2006) guidance on uncertainty in national greenhouse gas emission inventories, which includes guidance on expert elicitation.
Section 3.1, p. 22: choice of methods, test criteria, choice of data sets, etc. are subjective even when using so-called "objective" methods.
p. 23. EE should not be forced to produce consensus among experts if that is not representative of their judgments. Decision options or the decision is not really in the realm of EE.
p. 24 "Most scientists…" how do the authors know? "They lack understanding…" – seems pejorative. What is the "likelihood of the nature of an unknown quantity"?
Section 3.3.1: it should be stated that the goal of EE is to represent the expert's state of knowledge.
Section 3.3.3: not sure this is a widely accepted view of science. An alternative is to pose hypotheses that are tested based on empirical data in an ongoing process… also, there is an element of making predictions, not just describing a state of knowledge. A lot has been writing on "Science" and at least a citation could be made.
p. 26. What if results are not sensitive to the inputs obtained using EE? Or if empirically-based distributions are based on assumptions that data are a random, representative sample when they are not? Does "principles or rules" refer to a "model"? What is meant by "additional certainty for sparse data"? Perhaps one can make the most efficient use of such data, but one cannot make them more certain than they are.
Section 3.3.4 – mention that one can seek to reflect the range of credibility scientific judgments not for individual experts necessarily but by capturing judgments of multiple experts.
Bottom f p. 26. "factors" might be better here than "parameters"
Top of p. 27 – cross reference another section?
2nd paragraph, the point is to represent expert judgments. If experts disagree, then so be it. One can try to explain/interpret the disagreements, but should not seek to force a false consensus.
p. 29: first line. This assumes that the results from the seed domain are somehow relevant to results from the expert domain. How is that verified?
End of 3.4.3.2. Why would the goal be to obtain consensus views? This would seem to bias the elicitation and force experts into positions they might not really support individually.
p. 30, last paragraph – a common definition might not be an advantage to an individual expert who prefers a different structure and level of aggregation/disaggregation.
Section 3.5 Many of the subsections do not seem to fit under this header's topic.
p. 33. Here there seems to be internal tension in the definition of EE. Earlier, it was stated to be for uncertainty, but here it seems to apply to data gaps.

p. 34. The definition of a good expert seems a bit weak. "Understanding" is a vague and unmeasurable term. Perhaps one could say that an expert has the ability to identify and describe relevant data, models, theories, etc., and their applicability, strengths, limitations, etc.
how does one know if the judgment is "reasonable" – what is meant here? (too vague).
Informativeness is certainly useful to a decision maker but it does not seem to be an appropriate way to judge the goodness of an expert judgment. Does this imply a normative requirement for overconfidence? Or is it just not stated accurately here?
p. 35. "A good expert should not have bounds that are too broad…" this seems to be a dangerous concept if it is meant to be prescriptive. What does this have to do with the goodness of the expert?.
p. 36. The bulletized list is not parallel. Not all items are heuristics. Nor does it actually cover the scope of the rest of this section.
p. 39. Does the phrase "fifty-fifty" have a particular role here. Is the main point to avoid this phrase when encoding judgments?
p. 40. Conditional probabilities is undefined here. End of same section – this argues for multiple experts. Section 3.5.5.7 – is this in the context of eliciting individual judgments regarding specific quantities?
p. 41, next to last paragraph. Needs more explanation.
p. 42. Rather than control, perhaps better to say reduce the impact of these heuristics. Are "EE practitioners" also known as "elicitors"?
are all judgments about technical quantities likely or necessarily value laden? E.g., what about an elicitation of uncertainty in July daily peak air temperature at a specific location?
p. 44 – how are major sources of unc. Known prior to the elicitation?
p. 45: extrapolations – are these based on models?
p. 46: what is "full probabilistic"
p. 47 – the definition of EE seems to be given repeatedly in the document, leading to confusion for the reader.
Public comments are not necessarily from experts. Many such comments are from advocates who are highly biased.
p. 48 – this sentence is very unclear.
Table 4.1 – public comment is not a method for expert judgment – while experts can respond, many of these comments are from stakeholders and advocates, not experts. Is this table really needed?
p. 50: list the bullets in the order of scenario, model, input, and decision rule.
Section 4.3. The title implies that this section will deal with topics other than EE, but it starts with EE. Scenario analysis is not typically probability based. Other non-probabilistic methods are not probability based.
p. 52: be careful about strawman arguments. With regard to model choice, one could use frequentist methods to assess model performance using various diagnostic checks and/or by estimating precision and accuracy. "less capable" than what? In what situation?
Isn't EE a supporting method for Bayesian approaches, and not parallel to it?
Sensitivity analysis can and should be used in combination with probabilistic analysis. See, for example,
Mokhtari, A., and H.C. Frey, "Recommended Practice Regarding Selection of Sensitivity Analysis Methods Applied to Microbial Food Safety Process Risk Models," *Human and Ecological Risk Assessment*, 11(3):591-605 (2005)

Frey, H.C., and S.R. Patil, "Identification and Review of Sensitivity Analysis Methods," *Risk Analysis*, 22(3):553-578 (June 2002).

Here is an EPA-relevant example:

Mokhtari, A., H.C. Frey, and J. Zheng, "Evaluation and recommendation of sensitivity analysis methods for application to Stochastic Human Exposure and Dose Simulation (SHEDS) models," *Journal of Exposure Science and Environmental Epidemiology*, 16(6):491-506 (Nov 2006)

See also the text by Saltelli

p. 53-54: the issue of representativeness or lack thereof is one that should be defined and explained more, since this is typically argued as a key problem that EE can address.
p. 54: is poor level of detection a study design issue? It seems to be a limitation of instrumentation.
p. 55: separating variability and uncertainty, etc., is not a problem if one has a random representative sample. If data are not a random representative sample, then one needs to consider EE as an alternative regardless of sample size.
4.3.2.3: the use of sens anal. Prior to focusing EE needs a bit more discussion – how does one specify ranges for these sensitivity analyses prior to doing the EE?
4.3.2.4. What is meant by whether uncertainty is "acceptable"? what is its relevance to EE?
p. 91: need to list the structuring aspects and level of aggregation, dependencies, etc.
how can it be known how well extreme values are likely to be represented or existence of bias?
Table 6.3: EE questions need to include structuring.
What is meant by "mental models." Is this the subjective probability distribution function? Or is it something else?
p. 93, section 6.5.1, since the defined endpoint of EE is a subjective probability distribution, why is this section needed or what is the context for it?
Figure 6.1 is terrible – many undefined terms. This is perhaps the worst example of an influence diagram and is far from transparent or self-documented.
Figure 6.3 stands of trees? Avoid jargon specific to a particular domain if the authors want a general reader to follow it.
Figure 6.4. What is the header for the labels on the vertical axes? What does each data point represent? This seems overwhelming and yet unclear.
Figure 6.5: how were the judgments combined? Mixture distribution based on equal weights? Or an "average"?
Figure 6.7. Define RCS.
Table 6.6: give a header for column 1 – are these experts?
Table 6.6: is this a table of ratios? Not very clear. Are these mean values only? Where is the uncertainty? Why is this a relevant example for communicating uncertainty?
Table 6.6 (p 103) – label columns with headers. What do the letters mean in col. 1?
Table 6.7: seems like a lot of detail. What is the bottom line inference from this information? The table itself is a bit overwhelming and the key point is not clear.
Section 6.6. spell out EEs in the header.
p. 106, near bottom. Use weighting, not averaging.
p. 109. Some of the document mentions data gaps, but this is not mentioned in 7.1.1
What is "scenario analysis for ranking"?
p. 57: how is EE used to support "scenario analysis."?

What are the "suitable approaches to adjust it"?
p. 58. "subjective judgment" is there another kind of judgment?
p. 59. MCDM and MAUT do not seem to be mutually exclusive to EE – i.e. EE can support these in terms of uncertainty in outcomes for alternatives, even if not for utility functions or weights.
Why is stakeholder input critical to EE?
p. 61: why does Table 4.2 focus only on pre-elicitation?
p. 63: somewhere it needs to be said more clearly as to who does the elicitation? E.g., an "elicitor"?
p. 65: bottom, there is more than one EE protocol. The one given is more or less okay, but more clarity is needed that there are alternatives.
Figure 5.1: the elicitation session should include structuring, in which the conceptual and analytical model is reviewed with the expert and can be adjusted.
p. 67: problem definition shouldinclude prioritizing key sources of uncertainty based on prelim uncertainty and/or sensitivity analysis.
p. 69: 2nd paragraph under 5.2.4.1 – what would be differently in these two examples?
p. 71: the number of experts should be first and foremost related to the study objectives, with resource implications of the decision taken into account. The resource limitations of the assessment itself is of course a practical factor but should not be the first item listed here.
p. 72: why are more experts needed if the goal is to characterize central tendency and not just the range?
p. 73: step 3 needs to include structuring with the expert. Ad a step 5 on verification.
p. 73: "substantive expert" does not seem like the right term. Suggest "domain expert" later, "subject matter expert" is used, which is okay. Use terms consistently.
p. 75 "encoding process" - should this refer only to the encoding step, with "elicitation" used to refer to the process in all of its steps?
p. 76. Some experts want to draw their own distributions.
p. 76, section 5.3.3. earlier, I thought I read that Delphi method was considered outside the scope of EE as defined in this document.
p. 79. Section 5.3.4: sometimes it helps to be able to draw diagrams when meeting with the expert, for which only telephone is not adequate. Written surveys are easy to set aside.
p. 80. "combine" experts – should this really refer to weighting the experts?
p. 81. 2nd paragraph, middle. NRC committees do not attribute specific content to specific experts.
p. 82: what is a "paradigm" of an expert? Errors in the process are different than different choice of informational basis or inference options. Errors should be in a separate paragraph.
P. 83
Figure 5.3: use weights for each expert to create a mixture distribution, not averages that smear out the experts. What is meant by "aggregate"? does this mean to weight multiple experts?
p. 87 the use of the term "literate" implies that others are 'illiiterate', which seems unnecessarily pejorative in its implication. Perhaps another term or phrase would be better.
p. 89 – how is this discussion relevant?

Terms to add to the gloassary and to use consistently throughout the document

Accurate
Aggregation
Assumption
Assumptions
Availability
Averaging
Bias
Conditional Probability
Data gap
Data quality
Decision options
Dependence
Domain expert
Elicitation
Elicitor
Encoding
Estimates
Event
Extrapolation
Heuristics
Input
Model
Model choice
Objective
Overconfidence
Paradigm
Parameter
Precision
Quality
Quantity
Relationship
Representativeness
Robust
Subjective
Subjective Judgment (?)
Subjective Probability
Weighting

Comments from Dr. Katherine Walker

K. Walker: Charge question, reordered –

> **1)What potential implications of having such guidance should be considered?**
>
> **2) Do the topics and suggestions covered in the White Paper regarding selection, conduct, and uses of this technique provide a constructive foundation for developing 'best practices'?**
>
> **3) As EPA considers the future development of guidance beyond this White Paper what additional technical areas should be addressed?**

Question 1: What potential implications…

Having heard John Bailar's strong statement at the meeting that EPA should move forward with formal guidance, I want to make more explicit my reservations which are implicit in the recommendations I made below. While I understand the desire for guidance, I am not a big fan of guidance –I think it has the tendency to stifle rather than promote a scientific process in which new methods are allowed to be proposed, to be tested, and if they work, to be implemented.

- Need to frame the role, objectives, boundaries of guidance
  - Make clear the broader context of other guidance that is applicable, or under development. Concern expressed that pulling EE in particular out of the lineup of possible tools gives it an exaggerated emphasis relative to other methods.
  - Consider a role for EE that is integrated into discussions about future research directions, value of future research to decisions at hand, rather than just for questions about particular quantities late in the process
  - Think about it as true <u>guidance</u>, a set of goals, criteria for evaluating success that can be met by multiple approaches vs. something that will be used as a checklist (cookbook)
  - Discussion of when EE is appropriate/not appropriate will be challenging
- Remain open to new research, applications that demonstrate benefits
  - Evolving field,
  - methods would benefit from additional research, innovations, improvements to existing approaches
  - Avoid tendency of guidance to lock methodologies in time.

Question 2

Generally heard yesterday that the topics covered seemed appropriate, but not comprehensive. Needs to be built on, updated. Document needs sharpening of concepts, definitions, and careful resolution of conflicting redundancies, before it is put forward as a basis for guidance.

Question 3: Additional areas that need to be addressed.

From Charge Question A:

- Consider literature on how to gather rationales, (e.g. survey research.) Could help with peer review and transparency
- Update literature on cognitive biases and how they influence how probabilities, quantitative judgments should be elicited
- Consider other ways to encode judgments that are not quantitative probability encoding
- Literature on performance measurement, scoring is missing from document
- Protocols for ex-post evaluation (could have tools for evaluation as part of the elictation process) of judgments to evaluate consistency, statistical coherence. Include provisions for longer term follow-up of experts' performance (i.e. if relevant data become or are likely to become available)

Charge Question C.1. Selecting Experts

- Consider what the goal of the elicitation is (consensus, range of views, etc)
    - See for example, Cooke's concepts of
        - Survey
        - Political consensus
        - Rational consensus
    - Goals factor into evaluation of what "balance" means (e.g. have we got the right range or distribution of views, expertise to estimate the quantity of interest)
    - Also factor into decisions about whether or how to combine
- Need to clarify impact of OMB's paperwork reduction act on numbers of experts
    - Does that trump all decisions about greater than 9 experts?

Charge Question C. 2 Expert Aggregation/combination

- Generally needs more comprehensive, accurate discussion of alternative methods, their strengths and limitations
- Literature needs broadening - (eg Tetlock (hedgehogs v. foxes); Cooke performance based combinations, Copulas?) and updating; Bayesian model averaging (not sure what we concluded on this)?
- Need more careful discussion of when it is even appropriate to combine experts' distributions.
- Need more careful discussion of issues dependence, independence of experts

Charge Question C. 3 Problem Structure

- Clearer discussion of what is appropriate for elicitation: definitions of quantity parameter, relationship (i.e. don't generally elicit parameters, dealing with the issue of second order uncertainty, dealing with resistance to giving point estimates (preferences for ranges)
    - Was general agreement that the quantity should be in principle measurable
    - Add clearer discussion of epistemic v. aleatory uncertainty, uncertainty v. variability

- Need discussion of importance for expert of understanding the context and for the variable being elicited, what it is conditioned on, dependencies between variables
    - Presentation of the model into which the elicited value is going into, for example for regulatory impact analysis
    - influence diagrams,
    - mental models
    - Bayesian belief networks
    - other
- Role of stakeholders in development and structuring of questions – many around the table felt that stakeholders needed a role in order to gain credibility and acceptability.

Charge Question C. 4 findings and conclusions

- Needs fuller discussion of limitations of the methods, not just the benefits.
    - Consider opportunity costs of taking on a full EE, and the alternative approaches that might be taken depending on what problem at hand
- Although there are concerns about development of a cookbook, there are also benefits potentially of a more streamlined process.
- Consider alternative tools for characterizing uncertainties, their strengths and limitations (or at least make clear reference to relationship of this guidance to other activities at EPA)
- Update literature on cognitive biases - lots of new literature
- Needs fuller discussion of how to evaluate consistency, coherence of judgments
    - Nature, role, and appropriate use of feed back information/tools
    - Want reality checks, but ones that use analogy (e.g. comparison of estimate to better known or studied values (in children v. adults; causal likelihood of ambient PM related mortality v. environmental tobacco smoke or tobacco use.) Don't allow experts to see what the implications of their estimates are for the values relevant to public policy (minimize temptation to game the system).
- Append or reference specific examples of EE, other concepts

Charge question B – transparency

There was a suggestion that there might be one section on transparency. It is currently discussed in several places in the document. At the same time, that is because issues of transparency applies to:

- Selection of experts
    - Characteristics of experts – How defined: schools of thought v. academic title
    - Anonymity of experts –tradeoffs between strict scientific accountability vs. honesty/willingness to participate
- Development of protocol:
    - Inclusion of body of evidence from survey research methods could be valuable for developing clear questions
- Methods for documenting assumptions and basis of judgment; these were not clearly defined in the white paper.

- How one evaluates the interactions between elicitor and experts – did the process really go as advertised?
    - This gets to the issue of whether one can only evaluate the process not the output or opinions of the experts. General agreement was that should not critique the opinions. At the same time, the elicitation does not always prevent experts from saying things that make no sense – should but doesn't always happen.

- White paper should at least refer to risk communication skills and principles

General agreement that there are degrees of transparency, difficult to achieve full transparency, and that there are various kinds of tradeoffs between transparency and efficacy (cost, time) of the process, need for anonymity of experts.

Are we ready for guidance? What are the steps?

EPA should re-draft the white paper. The white paper is meant, in EPA's words, to lay the scientific foundation for making decisions about what the guidance should be. Thus at this stage, EPA needs to have a systematic basis for including literature that it includes in the development of the foundation. The agency shouldn't use policy grounds to discount certain literature (e.g. on performance measures and scoring as a methodology for weighting experts' judgments) at the white paper stage. If the foundation is incomplete, decisions on how to build the guidance may be poorly supported or miss important insight from literature that has not been considered.

An organizational thought –

As I was thinking about it after the meeting, it occurred to me that EPA should follow the general principle of separating risk assessment (science) from risk management (policy and judgment about its use) in its white paper. The document, or alternatively each section or issue, could be divided into two components:

1. Scientific review and critique of the literature (consistent with expectations for published review articles)
2. Discussion of EPA policy considerations:
    a. Regulatory needs
    b. Feasibility of methods and/or resources for implementation
    c. Need for transparency in the process
    d. Role for stakeholders
    e. Other…

Ultimately, my vote would be for "performance" guidance rather than "technology" guidance -- that we look for characteristics that define a good process and performance of that process, but not prescribe or proscribe the processes that can be used.

Comments from Dr. Thomas Wallsten

Notes for Charge Question C.1 (from pp 69-71)
Thomas S. Wallsten
Mitchell Small

a) Does the White Paper adequately address the different criteria and strategies that may be used for nominating and selecting experts?

The report sets out the following points regarding the process for nominating and selecting experts and the criteria by which the panel should be selected. We have bulleted the points here gleaned from the report; and suggest that they be bulleted in the report. They are as follows.

Process for selecting experts should be:

- Explicit and reproducible
- Reasonably cost-effective
- Straightforward to execute

Criteria:

- Multiple disciplines/technical balance
- At the forefront of their fields
- Recognized leaders
- Institutional or stakeholder balance
- Potential conflicts of interest not a barrier

Approaches to nominating experts:

- Literature counts [Beware here of high literature counts due to criticisms leveled at the work.]
- Ask scientists who have published to nominate people
- Use literature counts to identify people who then are asked to make nominations
- [Here, we would add inviting stakeholders to make nominations of scientists who meet the criteria specified above.]

Should sponsor conduct the expert selection process or contract that job out? (p. 71)

- Contract out – "benefit of greater objectivity" [We don't see why that necessarily is so.]
- Do in-house – more active control and ultimately is responsible for the quality and credibility.
- [We would add that the process will be judged on its merits regardless of whether or not it is contracted out. We strongly suggest further that the process be reviewed and approved before the actual selection process begins.]

b) Should additional technical aspects about this topic be included?

We would suggest:

- The process for selecting the experts be reviewed and approved before it commences. The review should be by EE experts, relevant EPA staff, and stakeholder representatives.
- Distinguish criteria for establishing the pool of experts from criteria from selecting from the pool given that it is larger than the number of experts required.