

March 13–14, 2006

Highlights of the Expert Judgment Policy Symposium and Technical Workshop

Roger M. Cooke and Katherine N. Probst

1616 P St. NW
Washington, DC 20036
202-328-5000 www.rff.org

Highlights of the Expert Judgment Policy Symposium and Technical Workshop

ROGER M. COOKE AND KATHERINE N. PROBST



AN RFF CONFERENCE SUMMARY

Introduction

Resources for the Future (RFF) hosted a two-day workshop in March 2006 on the theory and practice of expert judgment in risk and environmental studies. The School of Engineering and Public Policy at Carnegie Mellon University and the Center for Risk Analysis at Harvard University co-sponsored the workshop with RFF. The workshop brought together some of the most experienced researchers in expert judgment, senior policymakers from federal agencies, academics, and private consultants to discuss the role of probabilistic uncertainty analysis in public policy analysis, and the contributions and limitations of expert judgment to this analysis. This workshop would not have been possible without start-up funds from the Lounsbery Foundation, and contributions from RFF and Harvard's Center for Risk Analysis.

Uncertainty is inherent in most decisions faced by regulatory agencies: uncertainty exists in the assessment of health and environmental benefits as well as in the costs of regulatory options. Both the Office of Management and Budget (OMB) and the National Academy of Science (NAS) have repeatedly called for better efforts to quantify uncertainty for regulatory decisions and rulemaking. This workshop was one response to that call, and all participants were keenly aware of the potentially far-reaching consequences of revamping the way federal agencies deal with uncertainty.

The goal of the workshop was not to reach consensus on methods for employing expert judgment but to provide information to government managers about ongoing activities in the wider research community and to stimulate discussion on issues and concerns regarding the use of expert elicitation as one method for quantifying uncertainty. The first day of the workshop focused on presentations by prominent practitioners and users. The second day was a hands-on exercise in expert elicitation. Most of the presentations are available on the workshop website, www.rff.org/expertjudgmentworkshop. The purpose of this report is to present brief descriptions of each of the presentations and also to summarize some of the major issues that were discussed. Our goal is to give someone who was not at the workshop the key highlights from the meeting.

Keynote Address

Dr. George Gray, Assistant Administrator for the Office of Research and Development (ORD) of the U.S. Environmental Protection Agency (EPA) delivered the keynote address. A number of issues related to how EPA currently addresses uncertainty were raised. A particular area where more effort in uncertainty quantification is needed is IRIS (Integrated Risk Information System), the risk database developed and maintained by EPA. Gray noted that this database is used not just by EPA in regulatory decisions, but supports a wide circle of users both in the United States and abroad. He mentioned the following issues:

- improving the transparency and inclusiveness of the IRIS chemical evaluation process,
- including formal quantitative uncertainty analysis in assessments to make the information effective for the wide range of uses that IRIS supports, and
- developing guidance for incorporating uncertainty analysis in decisionmaking.

Gray noted that EPA and ORD are committed to developing capabilities in uncertainty and sensitivity analysis. This means not only factoring uncertainty into rulemaking and resource allocation, but also communicating effectively with the public about these uncertainties.



Above: Keynote speaker George Gray. Below: Gray answering questions



Why Quantify Uncertainty?

Presentations by Granger Morgan and Roger Cooke highlighted the need for quantifying and displaying experts' uncertainties in addition to simply summarizing their views for policymakers.

Granger Morgan, Professor and Head, Department of Engineering and Public Policy, Carnegie Mellon University

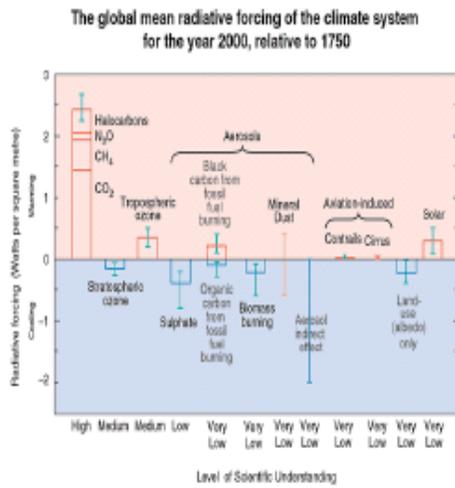
Morgan's opening talk, "Regulation under Uncertainty," focused on the characterization of uncertainty and decisionmaking under uncertainty, and gave some summary guidance on reporting and analyzing uncertainty. The bottom line was: "Without at least some quantification, qualitative descriptions of uncertainty convey little, if any, useful information." He noted that the climate change community is learning this lesson.

This point was illustrated with recent work in the field of climate change. A recent study conducted by Morgan and colleagues focused on mechanisms by which aerosols influence the rate of climate change (called aerosol forcing). Twenty-four experts were involved, and Morgan showed how the experts individually assessed the uncertainty in key physical variables. Striking differences existed among the experts. He compared the spread of the experts' uncertainties with the aggregated uncertainties of the International Panel on Climate Change (IPCC). The uncertainties can be expressed as box and whiskers plots, shown in the following figure. The vertical axis is watts per square meter. The "box" indicates the central 50 percent of an expert's uncertainty regarding the direct aerosol effect, and the "whiskers" indicate the 90-percent confidence band. On the left hand side, we see aggregated uncertainties from the IPCC, where the aerosol-related factors are covered by a bracket.

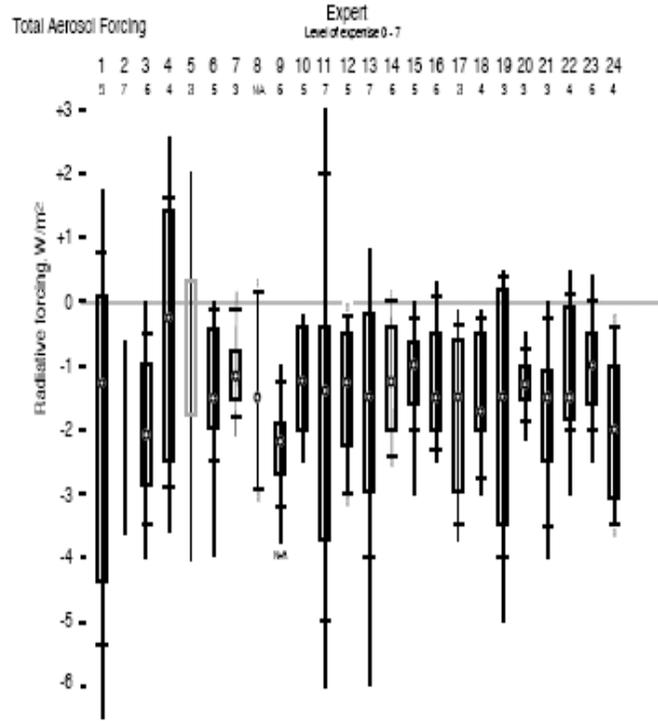
The lack of conformity among the experts with regard to this important driver of climate change is captured by these pictures and communicates something that policymakers and the general public should bear in mind. The 90-percent confidence bands of the experts range from +3 to $-8\text{W}/\text{m}^2$. However, the IPCC aggregated bounds range from +1 to $-2\text{W}/\text{m}^2$. Evidently, the aggregation of opinion performed by the IPCC masks this diversity of judgment within the expert community. This example underscores the importance of not brushing scientific uncertainty under the rug.

Carnegie Mellon University

Comparison with IPCC consensus results



Sources: IPCC TAR WG1
Morgan et al, *Climatic Change*, in press.



Department of Engineering and Public Policy

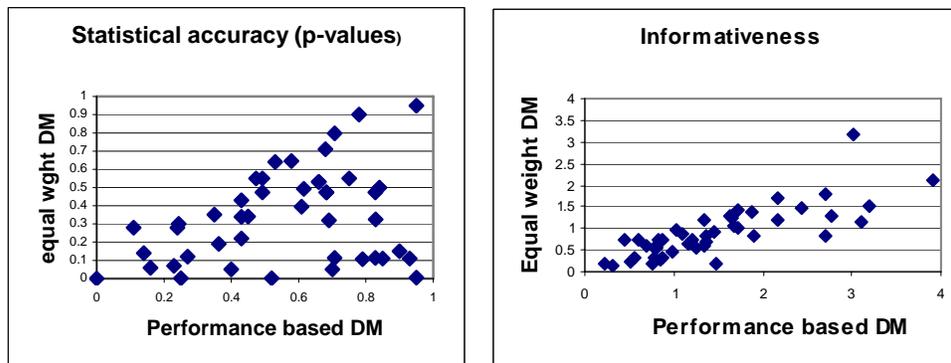
Roger Cooke, Chauncey Starr Senior Fellow, Resources for the Future

Cooke followed with nine theses on expert judgment:

1. Expert judgment is not knowledge.
2. Experts can quantify uncertainty as subjective probability.
3. Experts don't agree.
4. Experts are not always overconfident.
5. We can do better than equal weighting.
6. Citation-based weights do not perform better.
7. Experts like performance assessment.

8. “Uncertainty from random sampling ... omits important sources of uncertainty.”¹
9. The choice is not whether to use expert judgment, but whether to do it well or badly.

Cooke stressed that knowledge and scientific consensus are the result of the operation of the scientific method; expert judgment methods cannot pre-empt that role. The role of structured expert judgment is to quantify uncertainty, not remove it. Cooke and colleagues at the Delft University of Technology and elsewhere have gained extensive experience with structured expert judgment. In total, 45 full-fledged studies involving some 67,000 individual expert elicitations have been performed to date. Characteristic of these applications is that, in addition to the variables of interest, experts are also asked to assess their uncertainty on variables from their field with values that are known post hoc. These “calibration” or “seed” variables enable experts to be studied as statistical hypotheses. Two performance measures are of interest, statistical accuracy (often called calibration) and “informativeness.” Several examples from this reservoir of data showed that experts were sometimes overconfident and sometimes remarkably accurate in their probabilistic assessments. This motivates the scoring of expert performance and using these scores to form performance-based combinations. One participant added that her analysis of many of these data sets points to a clearly discernible expert effect. All experts are not equal in their ability to quantify their uncertainty. This motivates a performance-based combination of experts. This idea is illustrated in the two graphs shown below.



Each point represents one of the 45 studies. The graph on the left compares the statistical accuracy of the equal weight combination (vertical axis) with that of the performance based combination (horizontal axis). The graph on the right makes a similar comparison with regard to informativeness. The fact that the points tend to lie beneath the diagonal points to the advantage of performance based combination.

¹ Committee on Estimating the Health-Risk-Reduction Benefits of Proposed Air Pollution Regulations, Board on Environmental Studies and Toxicology, National Resource Council, *Estimating The Public Health Benefits Of Proposed Air Pollution Regulations* (Washington, DC: The National Academies Press, 2003).

Cooke also noted that experts' prestige is not always a good predictor of performance. He illustrated this with comparisons in which the combination of experts was based on their scientific citations. The citation-based combination gave performance roughly comparable to the equal-weight combination.

One graph from the recent Harvard-Kuwait expert judgment study on the mortality effects of fine particulate matter (also the topic of John Evans' presentation, summarized later in this report) drove home a point made repeatedly by the National Academy of Science in their 2003 report, *Estimating The Public Health Benefits Of Proposed Air Pollution Regulations*: "The probability models in EPA's primary analyses incorporate only one of the many sources of uncertainty in these analyses: the random sampling error in the estimated concentration-response function." The following table compares the effect on overall mortality of a one microgram-per-cubic-meter increase in ambient concentration of PM_{2.5} as reported by the American Cancer Society study,² the Six Cities Study,³ the equal-weight combination, and the performance-based combination from the Harvard-Kuwait⁴ study. The first two studies report uncertainty ranges based on random sampling. The experts acknowledge other sources of uncertainty, as reflected in the wider uncertainty ranges.

Estimates of the Percent Increase in Overall Mortality per 1 µg/m³ increase in PM_{2.5}

	Parameters Directly from Study		Estimates by Expert Judgment – Harvard Kuwait Project	
	ACS	Six Cities	Equal Weight Combination	Performance Based Combination
Median or Best Estimate	0.7	1.4	1.0	0.6
Ratio of 95% to 5% Estimates	2.5	4.8	260	60

² Pope, C. A., M. J. Thun et al. (1995). "Particulate Air-Pollution as a Predictor of Mortality in a Prospective Study of US Adults." *American Journal of Respiratory and Critical Care Medicine* 151 (3): 669–674.

³ Dockery, D. W., C.A. Pope III et al. (1993). "An Association between Air Pollution and Mortality in Six U.S. Cities." *The New England Journal of Medicine* 329 (24): 1753-1759.

⁴ Discussed in Dr. Evans's presentation, and in supporting documents for this workshop.



Drs. Cooke and Morgan field questions

Expert Judgment: Recent Applications

The second panel included presentations on the use of expert elicitation in two quite different contexts. William Aspinall, a world-renowned specialist in volcanoes and earthquakes, is frequently consulted when a volcano becomes active for advice on matters like alert levels and evacuation that can directly affect the lives of thousands or even millions of people. The emphasis on acute rather than chronic health threats distinguishes these applications from those that supply information to health regulators. Granger Morgan's second presentation is an example of the latter.

William Aspinall, Consulting Scientist, Aspinall and Associates

Aspinall has applied structured expert judgment with performance measurement and performance-based combination of experts extensively. As a volcanologist, he has been called in to support decisionmakers when volcanoes become active, and difficult decisions must be made regarding alert levels, travel restrictions, and evacuation. One such incident was the eruption in the British protectorate of Montserrat in 1997 pictured below, with deadly pyroclastic flows. These flows of hot heavy gas rush down at speeds of 150 km/hr destroying everything in their path.

When called upon by decisionmakers, Aspinall and his colleagues put in place a formalized procedure for providing scientific advice using the software system EXCALIBUR @TUDelft. This program for performance-based combination of expert judgments (used in the second day of the workshop) was developed with support from the European Union. More recently, Aspinall has applied the software in a European project, EXPLORIS, that studies the risks and risk management of the Vesuvius volcano. Mt. Vesuvius is not only potentially dangerous but it also has a population of several million in the immediate vicinity, as shown below. This is the highest-risk volcano in the world.



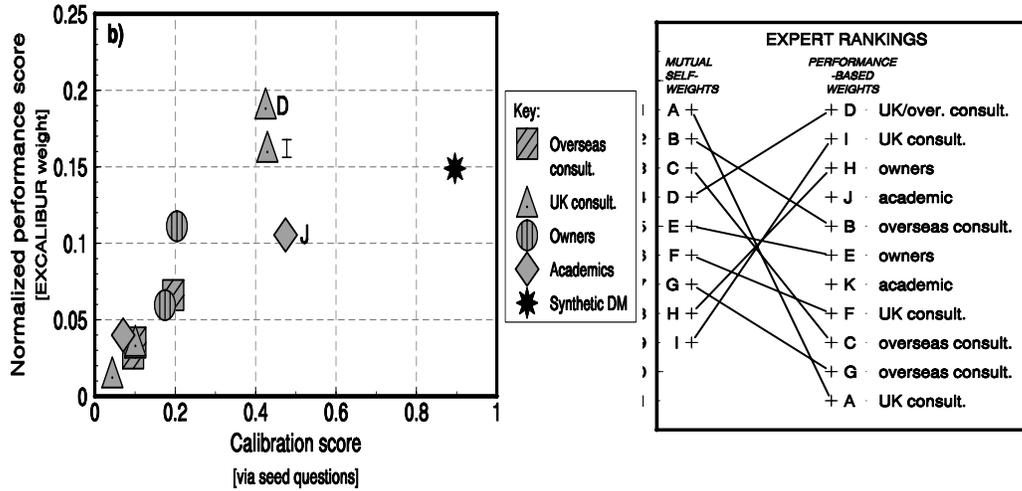
Mount Vesuvius with a population at risk of several million people



Other studies in which Aspinall has elicited expert judgments are in the areas of airline safety, dam safety, and SARS protection for healthcare workers. One interesting vignette points to a feature that was echoed in several other presentations, namely that prestige, or self-assessment is a poor predictor of expert performance.

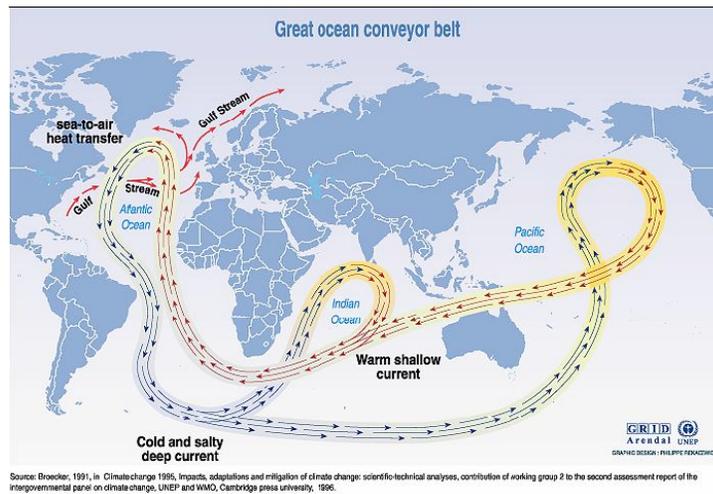
To illustrate, in one dam risk project, experts were asked to rank each other's expertise; these ranks were then compared with the experts' performance when assessing variables for which the true values were known post hoc. Performance is measured in terms of calibration, or statistical accuracy, and informativeness. The following graph from an extensive dam safety study shows the calibration scores and performance-based weights

of various experts. The “synthetic DM” is the decisionmaker formed with performance-based weighting.



Granger Morgan

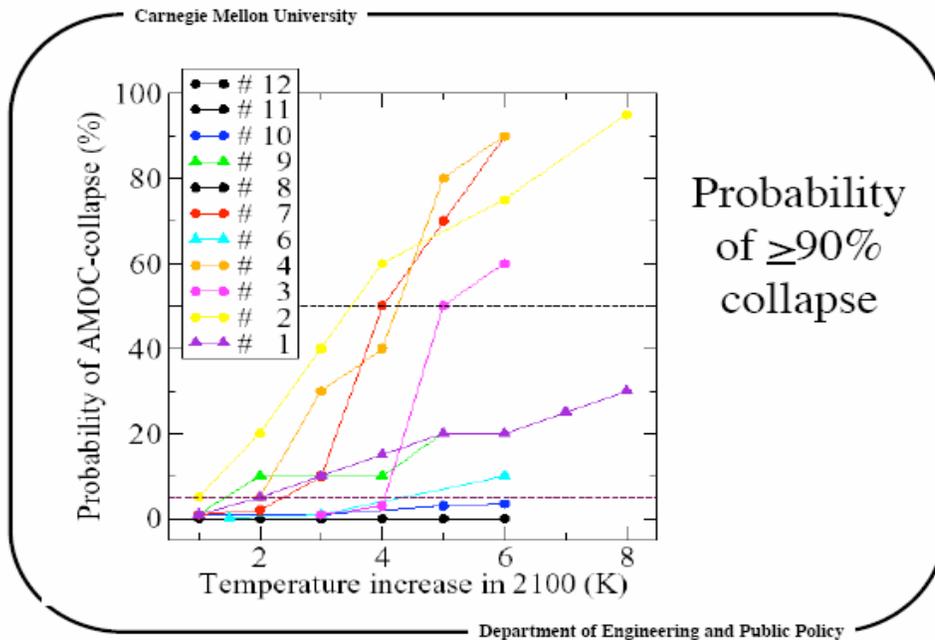
Morgan elaborated on experts’ judgments about uncertainties in climate change, focusing on aerosol forcing and on the “Great Ocean Conveyor Belt”—that is, the Atlantic Meridional Overturning Circulation (AMOC).



His study examined the present and future fate of the AMOC as viewed by experts from the Potsdam Institute for Climate Impact Research, Carnegie Mellon University, and University of Calgary. Detailed examples of the elicitation protocols showed how, in this case, the elicitation of the experts’ uncertainties was sensitive to each individual expert’s modeling of the AMOC. This contrasts with other studies where the questions for which

uncertainty quantification is requested, are fixed beforehand and are the same for all experts.

Perhaps the most dramatic graph on this subject is one depicting 12 experts' assessments of the probability of at least 90 percent collapse of the Great Ocean Conveyor Belt as a function of temperature increase in 2100. Whereas one group of experts placed this probability below 10 percent if the temperature increases by 6 degrees Kelvin, a substantial group envisioned a much higher probability.



Fine Particulate Matter PM_{2.5}

Recently, two structured expert judgment studies have been conducted on the health effects of fine particulate matter, PM_{2.5}. One study was conducted by the Harvard School of Public Health under the leadership of John Evans for the Kuwaiti government as part of their compensation claim for the 1991 oil fires and the other was performed by Katherine Walker and Henry Roman of Industrial Economics, Inc., Patrick Kinney of Columbia University, and Lisa Conner, Harvey Richmond, Robert Hetes, Mary Ross, and Bryan Hubbell, all of EPA. In the third panel of the day, these two studies were discussed.

John Evans, Senior Lecturer on Environmental Science, Harvard Center for Risk Analysis

Evans's talk, "What Experts Know about PM Risks" focused on the 1991 Kuwaiti oil fires, which were deliberately set in what appears to be the first instance of large-scale environmental terrorism. The key facts regarding the fires are shown below:



- More Than 700 Fires
- First Fires:
 - Air War ~ 17 January 1991
 - Ground War ~ 23 February 1991
 - Liberation ~ 28 February 1991
- Last Fire: 6 November 1991
- Oil Burned ~ 4×10^6 barrels per day
- PM Emissions ~ 3×10^9 kg
- PM10 levels – typical 300 ug/m³, sometimes 2000 ug/m³ – about 50 ug/m³ due to smoke.

According to Evans, the key issues surrounding the assessment of the dose-response relationship necessary to evaluate the oil fires' health impact were:

- **Exposure Level and Pattern**
 - Kuwait oil-fire exposure level and pattern is different than those of interest for most regulation, with background PM₁₀ levels in Kuwait of 200 or 300 $\mu\text{g}/\text{m}^3$, and PM_{2.5} increment due to fires averaging 10 $\mu\text{g}/\text{m}^3$ with spikes of several hundred $\mu\text{g}/\text{m}^3$.
 - Therefore, should time-series or cohort studies be used to estimate risk?
- **Composition of Smoke**
 - Oil-fire smoke differs in composition from typical urban aerosols in the United States and Europe.
 - Therefore, should an adjustment be made for differential toxicity?
- **Age-Structure of Population**
 - Kuwaiti population is far younger than U.S. or European populations.
 - Thus, do relative risk estimates from U.S. and European studies apply directly?
- **Do Epidemiological Studies Reflect Causal Associations?**
 - And if so, do the estimates need to be adjusted to reflect the possibility that the studies are confounded?

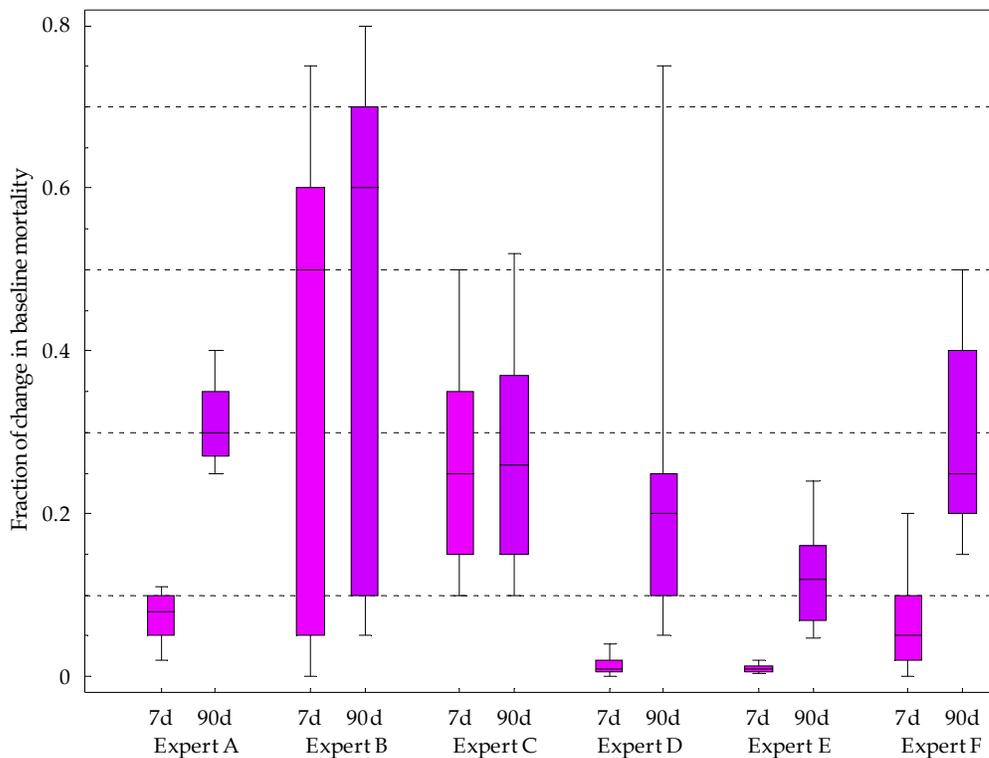
The elicitation questions focused on long-, short-, and intermediate-term effects. An example of a long-term question from the elicitation protocol is:

<i>Question</i>	<i>Setting</i>	<i>Exposure (Effect Interval)</i>	<i>Change</i>	<i>Pollutant</i>	<i>Composition</i>	<i>Baseline</i>
1	US	Long-term	1 $\mu\text{g}/\text{m}^3$	PM _{2.5}	Ambient	18 $\mu\text{g}/\text{m}^3$

What is your estimate of the true, but unknown, percent change in the total annual, non-accidental mortality rate in the adult U.S. population resulting from a permanent 1 $\mu\text{g}/\text{m}^3$ reduction in long-term annual average PM_{2.5} (from a population-weighted baseline concentration of 18 $\mu\text{g}/\text{m}^3$) throughout the U.S.? To express the uncertainty associated with the concentration-response relationship, please provide the 5th, 25th, 50th, 75th, and 95th percentiles of your estimate.

5% : _____ 25%: _____ 50% : _____ 75%: _____ 95%: _____

The answers to this first question were relatively consistent from expert to expert. However, substantial differences in opinion were observed on the questions relating to the timing of the effects. The graph below depicts the box-and-whisker plots for the percent of the long-term effect that is expressed within two time periods after exposure—one week and three months.



The estimates of the mortality impacts of the oil fires, which formed the basis for the State of Kuwait's claim before the United Nations Claims Commission (UNCC), are summarized below:

- **Kuwait's Basic Estimate of Risk**
 - ~ 35 deaths (based on ACS Study),
 - 0 (if studies are not causal) to 120 deaths (based on Six Cities Study)
- **Expert Judgment**
 - Best estimates of number of deaths by individual experts: 13, 32, 54, 110, 164, 2,874
 - Equal weight combination (82 deaths; 18 to 400)
 - Performance weights (35 deaths; 16 to 54)

Kuwait did not base its claim on the results of the expert judgment, but instead provided these results to give the U.N. Claims Commission panel some perspective on the uncertainty inherent in estimating the mortality impacts of the fires. While the decision regarding whether to combine expert opinions and, if so, whether to rely on equal weights or performance-based weights is of academic interest (and leads to slight differences in the characterization of the number of deaths attributable to the fires), the results of the expert judgment provide support for the validity of Kuwait's claim.

Katherine D. Walker, Independent Consultant

Katherine Walker's talk, "Expert Judgment about Uncertainty in $PM_{2.5}$ -Mortality: What Have We Learned?" described her work with Henry Roman of Industrial Economics, Inc., Patrick Kinney of Columbia University, and Lisa Conner, Harvey Richmond, Robert Hetes, Mary Ross, and Bryan Hubbell, all of EPA. EPA initiated this project to develop a more comprehensive characterization of uncertainty in the concentration-response relationship between $PM_{2.5}$ and mortality for use in regulatory impact assessments for proposed regulations. This study is significant because it represents the direct involvement of a regulatory agency in using structured expert judgment to quantify uncertainty for a major regulatory program.

The impetus for this study was given as:

- premature deaths avoided by reduction of $PM_{2.5}$ constitute 85–95 percent of monetized benefits,
- \$93 billion in reduced mortality for one rule alone (U.S. EPA Clean Air Interstate Rule), and
- aforementioned calls by NAS and OMB to improve quantification of uncertainty.

The key uncertainties on which this study could shed some light include:

- How strong is the likelihood of a causal relationship?
- What is the true shape of the dose-response relationship? Does a threshold exist?
- What is the impact of confounders and effect modifiers?
- How do potential errors in measuring exposure influence results?
- What is the impact of relative toxicity of PM components or sources?

A pilot study has been completed and its results are available on EPA's website (<http://www.epa.gov/nonroad-diesel/2004fr/420r04007j.pdf>, www.epa.gov/cair/pdfs/finaltech08.pdf).

A more comprehensive study is now underway that focuses on the same issue but involves a larger number of experts. Final results are not yet available, but the authors raised several important methodological issues that were the focus of the panel discussion. These included: cost and other resource demands; the appropriate number and selection of experts; the design of questions/elicitation protocols; and the theoretical/experimental support for approaches to eliciting well-informed and, ultimately, well-calibrated judgments about uncertainty. A final issue raised was the need for advance planning within regulatory bodies on how to incorporate these characterizations of uncertainty into the decision making process.

Round Table on Practical Issues

The panel for the round table discussion included Drs. Aspinall, Cooke, Evans, Morgan, and Walker, and addressed the following questions:

- When is the use of expert judgment warranted?
- What is a typical budget?
- How do you select experts? How many?
- Do you pay experts?
- What is your policy regarding expert identities?
- Do you use group or individual assessments?
- Do you use experts for preference rankings?

Among the presenters, there was substantial consensus regarding the desirability of using structured expert judgment to quantify uncertainty, particularly where the stakes are high. There was also agreement that, at some point, the scientific basis for the issues of concern would become so thin as to render expert judgment useless. Questions of theology, for example, are not suitable for adjudication by structured expert judgment. No agreement was reached on where this boundary might lie or how we might discover it. Regarding budgets, there appears to be a range of estimates on the cost of conducting structured expert judgments studies among the panel. Panelists who work in the United States reported that studies (done in support of government regulation) cost \$100,000–300,000 or more; studies in Europe tend to cost between one and three “person” months, or

\$30,000–100,000, excluding experts' time. Of course, there are some differences in the scopes of these studies. In addition, the U.S. regulatory context imposes a high peer review and legitimization burden that may account for higher costs.

Paying experts for their time and travel has a strong impact on costs. In the U.S. Nuclear Regulatory Commission–European Union project estimating uncertainty of accident consequence codes for nuclear power plants, for example, experts were paid \$15,000 each for participation. They were required to attend a two-day workshop prior to the elicitation and to write up the rationales underlying their assessments in a form suitable for publication. This represents the high end of expert costs. At the low-cost end, experts are designated by their company or institution, do not convene for a common workshop to discuss the issues, and are interviewed in their offices by elicitors. Of course experts' time always costs money; the question is only on which budget ledger it appears. Most studies fall somewhere between these cases.

Panelists in discussion



Another key topic was how to select experts. There are a number of ways this is typically done. If a company or laboratory is contracting the study, it may supply its own experts. For example, when Aspinall conducted a study to assess uncertainty regarding safety margins in British Airways, the experts were British Airways pilots. In other cases, experts are drawn from the broader expert community. The choice of experts is time consuming and often subject to independent review. In some cases, literature reviews are used, perhaps in combination with iterated nomination rounds. Attempts may be made to balance opposing schools of thought and include minority viewpoints. In some cases—for example, in the evaluation of the toxicity of new chemical compounds—the number of experts may be very small and widely dispersed. In other areas, for example atmospheric dispersion, there may be a large number of experts from whom to choose.

According to the panelists, the number of experts for most studies they conducted was targeted to lie between 6 and 12, although constraints of a given study can lead to different numbers of experts. Generally, however, at least six experts should be included; otherwise there may be questions about the robustness of the results. The feeling of the practitioners is that beyond 12 experts, the benefit of including additional experts begins to drop off. Some participants noted that this is quite a small number if the goal is to survey the span of expert judgment.

Some studies use weighted combinations of expert judgments based on expert performance. The elicitation is seeded with variables from the experts' field whose true values are known post hoc. This enables experts to be treated as statistical hypotheses according to conventional statistical methods. In other words, we can analyze the experts' statistical accuracy and their informativeness.

Use of seed variables with experts in different disciplines drew some comments from workshop participants. In a large study, it may be necessary to form different panels of experts to assess different clusters of variables. In some cases, a panel's composition and focus does not line up perfectly with the background and expertise of all its members. Ultimately, it is up to the skill of the analyst conducting the expert elicitation to break down a large problem into pieces that map onto distinct fields of expertise. An example was given involving the failure frequency of underground gas pipelines. An initial panel of experts was broken into two sub-panels, dealing with corrosion and with third-party interference, respectively. Some experts served on both panels.

The use of experts' names is another recurring issue. Cooke stated that his policy on this had been formed very early on and has not deviated: expert names and affiliations are part of the published documentation but individual experts are not associated by name with their assessments. This association is preserved as part of the scientific record, and is open to competent peer review, but is not made public. The reasons for this are many and varied:

- experts working in a company may participate as a condition of a client and may have opinions at variance with the "company position;"
- experts should not be subject to "expert shopping" by litigants to a dispute who may seek experts whose views are sympathetic to their interests;
- experts do not want to be subpoenaed or otherwise cross examined as a result of their frank and honest participation in the study; and
- experts should be shielded from intrusive comments regarding their performance.

Another question concerned the advisability of eliciting experts individually versus eliciting a group. In a group assessment, experts are brought together under the guidance of a facilitator with the goal of producing one collective assessment representing the group. There was unanimous agreement on the preference for individual assessments, motivated principally by the desire to avoid effects of dominant or vocal participants.

The panel also discussed the advisability of combining expert assessments of an uncertain quantity. The practitioners voiced different points of view on this question but agreed on the following: the spread of expert assessments is valuable information that should always be part of the reporting. There is no reason to combine expert assessments if you do not need to. If performance assessment has been done, then the role of the analysts extends to identifying those combinations that are statistically defensible, and those that are not. Whether and how to combine judgments is clearly the client's call. Many practitioners noted that a successful expert elicitation requires the collaboration of a domain expert and a “normative” expert, that is, someone skilled in probabilistic assessments. These two together form the analysis team. Combining these roles in one person may work if this person's background happens to cover both areas.

The question of preference ranking was discussed at greater length on the second day of the workshop.

Challenges for Policymakers and Practitioners

The day's final session was focused on emerging issues for users and practitioners of structured expert judgment. Robert Hetes of EPA fixed his sights on “Uncertainty and the Regulatory Process,” and Alan Krupnick of RFF ended the day with a talk entitled, “Are Decision Makers at Home on the Range? Communicating Uncertainties in Regulatory Impact Analyses.”

Robert Hetes, Senior Advisor for Regulatory Support, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency

Hetes acknowledged that expert elicitation (EE) is a form of the broader category of expert judgment that has been used in one form or another since the inception of science. The distinguishing feature of EE is that it is a structured, traceable, data collection process (that some call *structured* expert judgment).

Increasing interest in using EE to support decisionmaking and emerging requirements for uncertainty analysis at EPA has been fueled by, for example, the National Academy of Science's 2002 report, *Estimating the Public Health Benefits of Proposed Air Pollution Regulations*, the Office of Management and Budget Circular A-4, and EPA Cancer Risk Assessment Guidelines (March 2005).

Hetes noted that EPA acknowledges the potential value of this method. However, he stressed the need for an interagency task force to explore the implications of promoting the conduct and use of EE in EPA because of:

- unfamiliarity with this method among most EPA analysts and decisionmakers;
- the lack of clear guidelines on how to conduct EE within EPA (or elsewhere);
- the desire to promote consistency;
- the potential impacts of precedents from near-term projects utilizing EE;

- the need to promote technically defensible assessments; and
- the broad range of statutory, regulatory, policy issues to be addressed

EPA pursued EE starting in the late 1970s through the mid-1990s in support of the National Ambient Air Quality Standards reviews. Based on these experiences, EPA feels that EE is an accepted methodology to characterize uncertainty. However, early efforts in the late 1970s were criticized due to lack of experience and formal methods. Subsequent collaborative efforts involving the EPA Science Advisory Board and outside experts were important to move the method along. Similar activities will likely be needed to promote the use of EE in other program offices within EPA to ensure that such efforts are relevant and of high quality.

Hetes noted that quantifying uncertainty is necessary and important in the regulatory world but there are genuine concerns about how to do it and how to use it. In particular, quantifying uncertainty might undermine confidence in regulatory decisions, open the regulation to legal challenges, and might provide fertile grounds for those wishing to delay a regulatory process. Moreover, formalizing a decision process might reduce the ability to respond in a flexible manner and account for case-specific considerations.

These concerns become more pronounced in the multi-stakeholder and adversarial context in which regulatory decisions are often taken. The nature of the regulatory process necessitates a high degree of scrutiny, which has implications on how EE may be conducted and how the results should be used. The rigor of any effort depends on the intended purpose and use. Given the relative inexperience with EE on the part of most stakeholders, some may perceive the tool as sensitive to bias and manipulation. Therefore, the acceptability of EEs conducted in support of regulatory decisions may depend on the degree to which they are transparent, credible, objective, sufficiently rigorous, and relevant to the issue at hand. As a result, properly conducted EEs for such purposes tend to be time and resource intensive and may not be appropriate in all cases. EE is one of many tools to characterize uncertainty and many factors (technical, administrative, political, and procedural) should be considered in deciding when and how to conduct an EE, and how to use its results.

Alan J. Krupnick, Senior Fellow, Resources for the Future

Krupnick prefaced his talk with a quote from former EPA Administrator Christine Todd Whitman, which captures the frustration of a regulator in the face of uncertainty. “A big part of my frustration was that scientists would give me a range,” she said. “And I would ask, ‘please just tell me at which point you are safe, and we can do that.’ But they would give a range, say from 5 to 25 parts per billion (ppb). And that was often frustrating.”

While sharing the motivations given by previous speakers for introducing uncertainty quantification into the regulatory process, Krupnick also voiced some concerns. He noted that better, more complete (and more complex) information can confound and paralyze rather than improve decisions. Improvements in capturing uncertainty analytically must be matched with improvements in communication.

Krupnick and his RFF collaborators, Richard Morgenstern, Peter Nelson, Dallas Burtraw, Jih-Shyang Shih, Michael Batz, and Michael McWilliams, performed a study to push the frontiers of Regulatory Impact Assessments (RIAs) by including new as well as standard areas of statistical uncertainties in an analysis of future reductions in the nitrogen dioxide cap (such as uncertainties in costs). Then, they modeled decision alternatives for regulation—tight cap, intermediate cap, and “do nothing”—that were presented to seven former EPA decisionmakers. The results were that:

- each of the three decision alternatives was preferred by at least one decisionmaker;
- two of the seven favored the tight option, one favored doing nothing, three favored the intermediate option, one ruled out the do-nothing option and said to decide he would need more information; and
- all seven said that an uncertainty presentation was useful and helped them get an idea of the confidence they should have in their decisions and better prepare them to defend them against their critics.

Principal findings from this exercise were:

- Different decisionmakers have different learning styles.
- All wanted more information for a “real” decision.
- It is important that technical information be presented in context, such as the record of previous rulemakings and estimates of their success.
- Tables are a preferred communication approach—generally at least as preferable as the best graphics.
- Probability density functions are a preferred graphic which
 - may push decisionmakers toward an intermediate option, but
 - nonetheless encourage thoughtful discussion.
- Some felt in real life they got analyses from their staff with pre-ordained outcomes.
- They often “don’t know what they don’t know.”

The presentation ended with a call for more research devoted to communicating with decisionmakers and decisionmaker training in understanding uncertainty. Standardized formats would be helpful in this regard and research should be devoted to selecting preferred options.

Key Issues

Participants raised a number of issues regarding the use of structured expert judgment. These issues and the flavor of the discussion surrounding them are presented below.

1. How should we deal with motivational bias?

Motivational bias refers to a situation in which an expert is motivated to give assessments at variance with his or her true beliefs. This can arise in various ways:

- An expert feels beholden to express views compliant with those of their employers or sponsors.
- An expert is rewarded for under- or over-estimating a quantity (as when salespeople are rewarded for exceeding their self-assessed sales targets).
- An expert “tilts” his or her assessments to compensate for suspected leanings of other experts.
- An expert is “gaming” the system by tailoring his or her assessments to gain influence in the process.

The views of practitioners varied on the extent to which motivational bias occurs. Some said that they never encountered this problem, while others reported that they are quite sure that they have, including one who recounted the story of an expert who used his reputation for mobilizing media attention to influence the elicitation process.

A clear protocol for dealing with motivational bias is not presently available. It was noted that structuring the elicitation and requiring experts to document the rationale underlying their assessments discourages motivational bias, as does the use of external validation with seed variables. Absolute guarantees, of course, are not available.

2. What are the costs of an expert judgment study?

The cost of an expert judgment study was a frequent concern raised throughout the day. In a regulatory context, the legitimation and peer-review process can be intensive and can drive costs up. Costs are influenced by a number of factors, including:

- holding a plenary workshop prior to and/or after the elicitation for the purpose of elicitation training or discussion of data,
- subjecting the expert identification process to peer review,
- subjecting the elicitation protocol to peer review,
- preparing the experts (determined by the cognitive depth of the subject area),
- paying the experts,
- having the experts write their own rationales, as publishable documents, and/or
- having a client on the analysis team, helping to draft the elicitation protocol.

The expert judgment studies on accident consequences of nuclear power plants, conducted jointly by the U.S. Nuclear Regulatory Commission and the European Union, with experts being paid \$15,000 each for their participation, including a two-day workshop prior to elicitation, and an extensive write-up.

3. Should we elicit experts individually or try to extract a consensus from a group?

Methods such as Delphi or Nominal Group Techniques that aim at delivering a group opinion were developed and extensively reviewed in the 1970s and 1980s. Practitioners at the workshop generally preferred individual elicitations without trying to reach group consensus. Attempting to extract or generate a group opinion is fraught with difficulty. Real agreement is created by, and only by, the operation of the scientific method itself: it cannot be created by an expert judgment methodology. Group elicitations are sensitive to the skill of the facilitator and to the presence of personalities with a penchant for self-validation. Moreover, in many cases, experts already know each other well and are familiar with each other's arguments from scientific conferences.

4. Should expert judgments be combined?

Is it better to synthesize experts' judgments into one single judgment, or simply to present the diversity of the experts' views? All the speakers agreed that the diversity of expert views itself carries information and should be part of the open reporting of the study results. If there is no compelling reason to combine the judgments, then they need not be combined, obviously. In some cases, however, the judgments are fed into models, or are concatenated with other expert panels. In such cases it may not be feasible to propagate all individual expert assessments through the calculations

5. If experts' judgments are to be combined, how should it be done?

This was a major topic of the workshop. The simplest answer is "give all experts equal weight." This approach has intuitive appeal, but few decisions in science are taken by a show of hands. Much discussion focused on the value of performance-based combinations of expert judgment and on the measures one should use to determine performance. If the number of experts is moderately large, an equal-weight combination can be rather uninformative. If statistical performance has been measured and verified for various combination schemes, then the choice of combination scheme is likely to be driven by other factors. The client may elect to sacrifice informativeness for greater inclusiveness.

6. Is it better to elicit the input or the output of a model?

The answer to this question depends on the problem. In cases where a well-specified and agreed-upon mathematical model for predicting a particular quantity does not exist and/or the parameters of that model are unobservable and difficult to estimate, it may be better to elicit the outcomes. Where an agreed-upon model exists, but the parameters are difficult to estimate, it may be desirable to find observable phenomena that are predictable with the model and with which the experts are familiar. The experts can be elicited on these variables and the experts' distributions "pulled back" onto parameters of the model. Finally, where a model exists for which the parameters are observable but the outcome is not, it may be preferable to elicit experts on the parameters. Finding the right

elicitation variables to enable the quantification of an abstract and sophisticated model requires skill on the part of the analyst.

7. If an expert's performance is validated on a series of seed variables, does that mean the same expert will do well on the variables of interest?

This is a recurring question that has not been fully resolved. Performance-based elicitation methods assume that an expert's views can be treated as *one* statistical hypothesis and that we can gauge his or her performance on the variables of interest by performance on other variables within the expert's field of expertise. Studies have indicated that performance on "almanac questions" (for example, "what is the population of Damascus?") does not predict performance in an expert's field. Indeed, for almanac questions there is no professional pride involved in giving informative and accurate answers. The elicitation methodology is caught on the horns of a dilemma here. If variables of interest could be observed, an expert's performance could be verified, but then expert judgment elicitation would not be necessary in the first place. By the nature of the case, we can never directly verify the predictive power of performance on the seed variables. In fact, the burden of proof should be reversed. If two experts gave very different assessments, and one of them performed very well on the seed variables whereas the other performed very badly, would you (the decisionmaker) be inclined to regard them equally for the variables of interest? If the answer to this question is "yes," then the seed variables fail to enhance the credibility of the final result.

8. Have there been cases where you can directly verify experts' performance post hoc?

In some studies, experts assess their uncertainty on economic variables that become known in the future; two examples are the opening price of a stock index, as assessed at closing time on the previous day, and office rent at various times in the future. In these cases, the prognosis based on seed variables was reasonably confirmed. In other cases, probabilistic weather forecasters have been followed over a course of years and were seen to perform consistently, after an initial learning period. In some of the volcano assessment examples cited by Aspinall, experts predicted measurable conditions a few days or weeks in the future, and these were reasonably well born out. Walker's published work did not use seed variables but assessed the calibration of exposure experts' judgments about uncertainty regarding the benzene concentrations in ambient, indoor, and personal air. The experts were asked essentially to predict the outcome of a large exposure study under way in EPA Region V. Their judgments were subsequently calibrated with the measured values. Most of the experts performed reasonably well, particularly on the predictions of mean values; those who did not were typically anchored to older data and failed to adjust for more recent changes in the body of knowledge.

9. Are there any demographic aspects that can determine experts' performance (like gender, age, or professional position)?

Although there has been psychometric research on related questions, generally using graduate students in psychology, the results do not seem transferable to this area. There is

some recent and ongoing work using the existing expert database from the Delft University of Technology (described in the supporting documents). In general, there is little research that proves that age, gender, geography, background, or any other such variable predicts performance. Some evidence from medical applications suggests the predictions of more experienced physicians are better calibrated than those of medical residents. Fields with a great deal of empirical measurement data (as in weather forecasting) seem to produce experts who perform these probabilistic assessment tasks better than fields without much measurement data. In one volcano study, the best expert was not a volcanologist, but an electronic engineer with extensive experience with field measurements.

10. Given that peer review is a well-recognized instrument for reaching agreement on some issues, what is the connection between EE and peer review?

Peer reviewers should not render a judgment on the experts' assessments. Peer review should concentrate on reviewing protocols and process, not the results. As such, it should be done at the beginning of a study, not at the end.

11. When is a problem an expert judgment problem?

It seems clear that as questions become increasingly speculative and removed from established scientific fields, the application of structured expert judgment becomes increasingly problematic. At some point it would no longer make sense. At what point that would be, and how we would know we had reached that point are questions that currently await answers. Cooke suggested that if it is not possible to measure expert performance with seed variables, then it is not an expert judgment problem.

12. Where could more research effort in EE be spent?

Many participants regretted the lack of structural funding for research into EE. Too often, funding is provided for EE studies without any budget or resources for advancing the state of the art. Researchers are therefore condemned to do their research "on the fly." There are many areas that could benefit from more focused evaluation, including:

- performance variables and their measurement,
- factors affecting performance,
- optimal training,
- mathematical methods of aggregation,
- dependence modeling and dependence elicitation,
- training decisionmakers in dealing with uncertainty, and
- developing tools for communication about uncertainty.

Technical Workshop, March 14, 2006

Day two of the workshop was targeted for people who might actually perform a structured expert-judgment study. The goals were to expose participants to the elicitation process, give them feedback on their performance as probabilistic assessors, and to familiarize them with one type of software used in analyzing expert judgments, EXCALIBUR.

Refresher: Foundations of Decision Theory

Cooke's opening presentation reviewed the decision-theory background behind the representation of uncertainty. The key, according to Cooke, is: "Uncertainty is that which disappears when we become certain." Unproblematic as that may seem, an abundance of confusion exists regarding concepts such as:

- subjective versus objective probability,
- the interpretation of the probability axioms,
- non-probabilistic representations of uncertainty, and
- the role of experts and stakeholders in a structured decision process

Decision theory has its point of departure in the preferences of a "rational individual" and shows that such an individual's preferences can be uniquely represented as an expected utility. A rational agent structures decision problems by assessing their subjective probability of possible states of the world and assessing his or her values or utilities over possible outcomes of his actions. This explains why uncertainty or "degree of belief" is represented by subjective probability. In a complex group-decision context, the roles of assessing uncertainty and assessing values and priorities should be allocated to different players. Uncertainty should be assessed by domain experts, and values should be assessed by stakeholders, voters and/or their elected representatives.

According to Cooke, the "frequentist" or objectivist interpretation is another, equally viable, operationalization of probability axioms. Other programs of interpretation, such as classical and logical, are generally abandoned for lack of suitable operational definitions.

Cooke also focused on putative representations of uncertainty that do not give any operational definitions but rather leave this for the user and his or her interlocutors to sort out. These include the so-called fuzzy and possibilistic paradigms. The fuzzy representation was discredited on a disarmingly simple counter example involving an unknown emailer, Quincy. If the uncertainty that Quincy were a man were equal to the uncertainty that Quincy were a woman, then based on the usual fuzzy interpretation, this value would be equal to the uncertainty that Quincy was both a man *and* a woman.

The program EXCALIBUR used in the mock-up elicitation groups was developed by Cooke and his colleagues at the Delft University of Technology. It enables performance-based combinations of expert judgments, based on two performance variables:

- calibration—that is, the degree of statistical support for the expert’s assessments, considered as a statistical hypothesis; and
- informativeness—the degree to which the expert concentrates his or her probability mass on a small region.

The software also supports simple pair-wise comparisons used in determining group valuations of outcomes.

After broaching these issues, the techniques for preference elicitation were briefly reviewed.

Mock-Up Elicitation Exercises

The EXCALIBUR software was distributed free-of-charge to the participants, and mock-up elicitation exercises were conducted to allow the participants to undergo an elicitation and to familiarize themselves with the software.

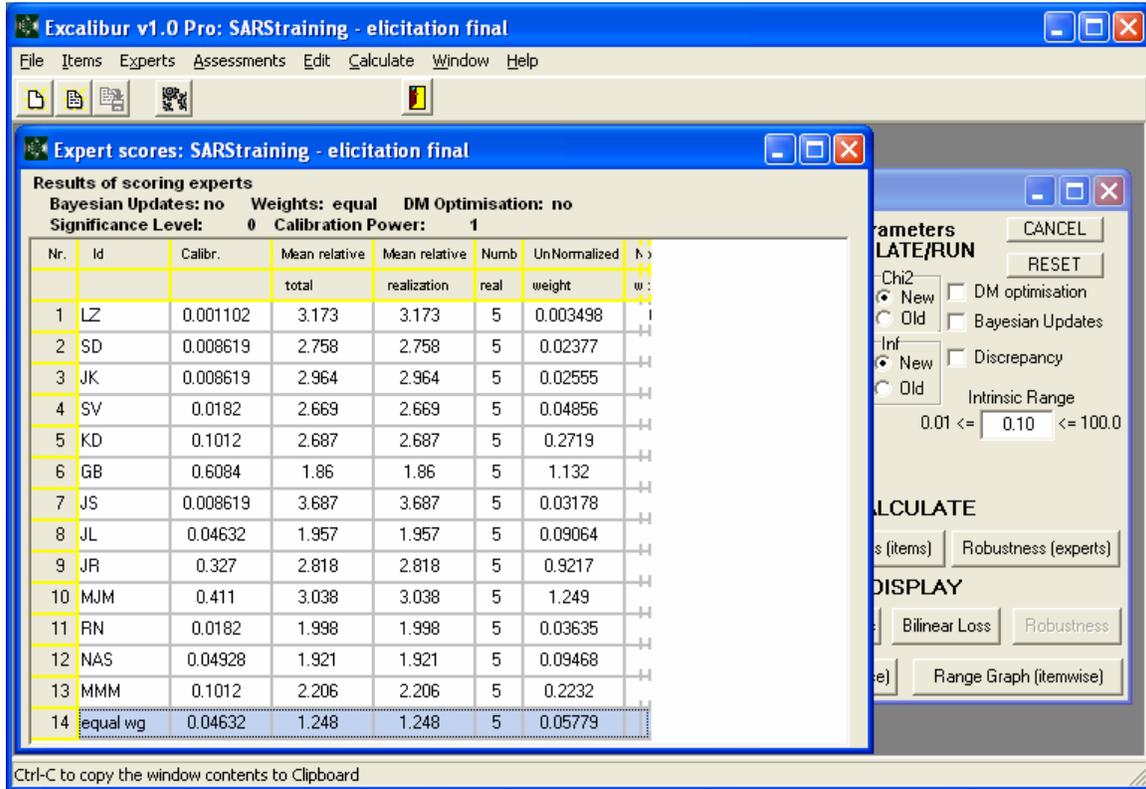
Cooke, supported by Walker and Evans, conducted one mock-up session on air quality, and Aspinall, supported by Daniel Lewandowski, conducted another mock-up on SARS worker health and safety.

Aspinall leads a mock-up session on SARS worker health and safety.



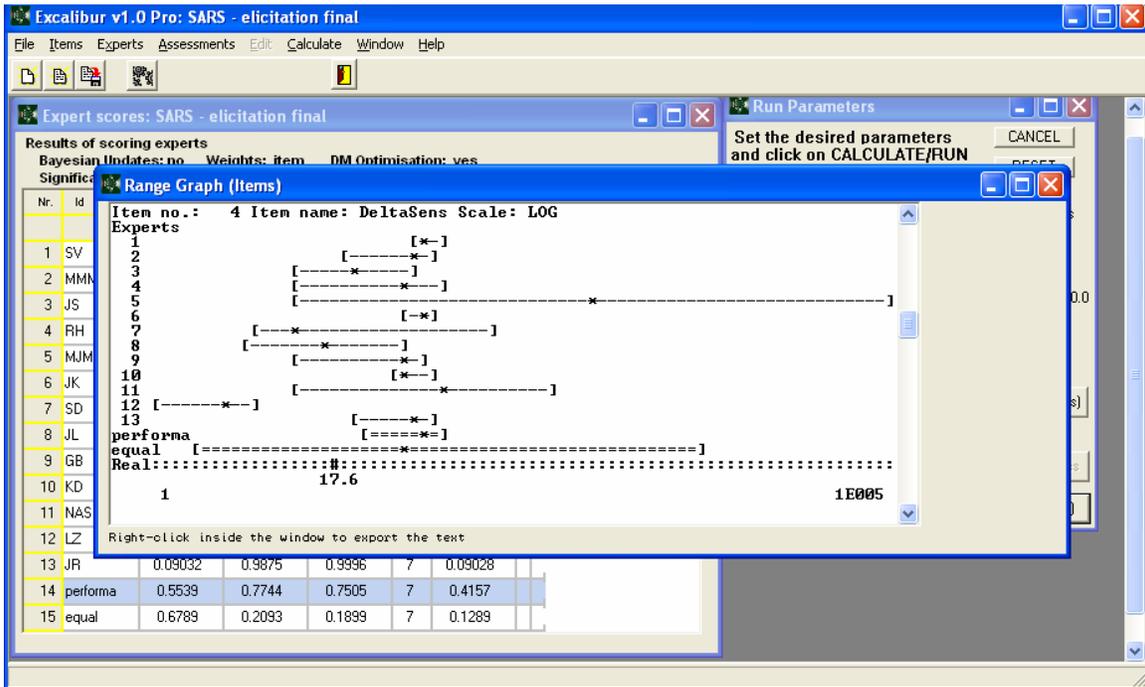
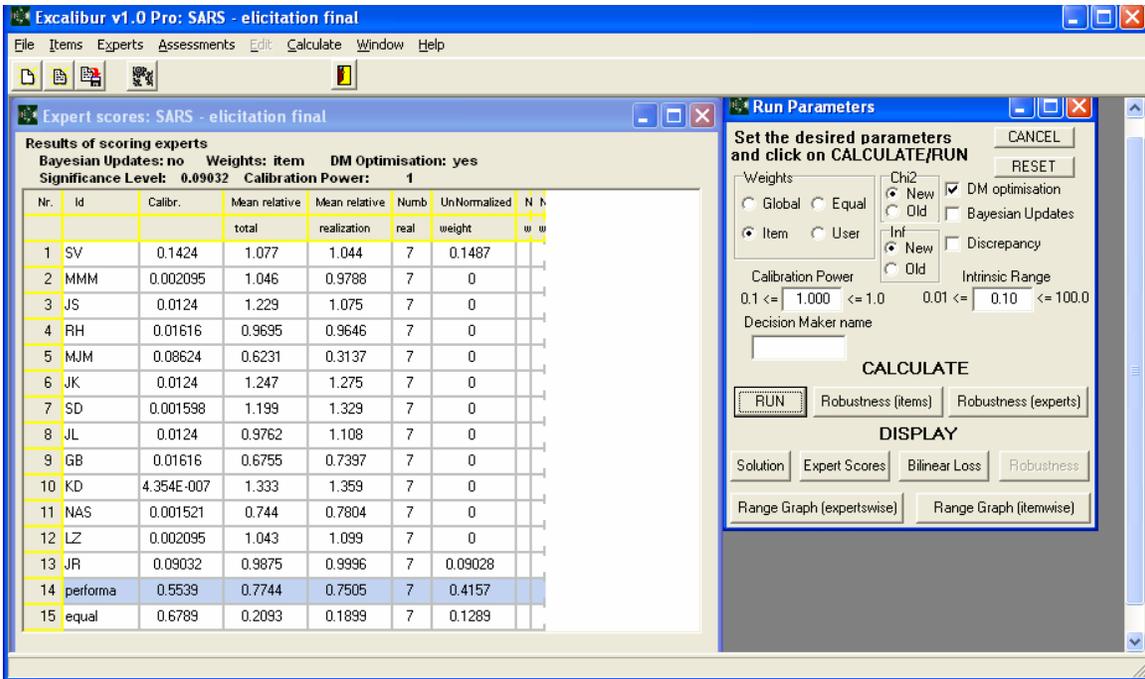
The participants underwent a training exercise as their first exposure to probabilistic assessment. The results from the SARS group (below) show that some participants have a steep learning gradient. In the first exercise, the calibration scores (the p-value of the hypothesis that an expert’s probabilistic statements are correct) ranged from 0.0011 to 0.411. These scores are shown in the far left column marked “calibration” in the first

table below. The “equal weight” decisionmaker also has rather lackluster statistical performance (i.e., below the statistical performance criterion of $p=0.05$).



After the training, the results in the second table and figure below were obtained, showing improvement in some individuals.⁵ The table shows the performance-based decisionmaker as well as the equal-weight decisionmaker. Both show acceptable statistical performance, as indicated by a p-value (calibration score) well above the common rejection level of 5 percent. The performance-based decisionmaker is noticeably more informative as shown in the columns “Mean relative (information).” This is due to the fact that the DM optimization chooses an optimal significance level for the hypothesis testing, so as to maximize the product of the decisionmaker’s calibration and information scores. For the performance-weighted decisionmaker, weight is proportionally allocated to two experts, SV and JR, who performed best. The greater informativeness of the performance-based decisionmaker is also illustrated in the range graph following the scores screenshot.

⁵ The scores cannot be directly compared because the number of seed variables differed in the two exercises.



Stakeholder Preference

After participants worked with the results of their own elicitations, the group reconvened for a stakeholder-preference elicitation. The choice was between six iconic fuel policies.

1. Tax at the pump: \$1-per-gallon gasoline surcharge, to be used for research in renewables.

PRO: encourages people to use less gasoline, reduces greenhouse emissions, encourages carpooling to reduce congestion, and accelerates renewables time-to-market.

CON: costs consumer; research better left to the private sector.

2. Tax break: (a) No sales tax for purchase of new hybrid or electric car; (b) first-time owners can deduct purchase cost from their income tax; (c) no sales tax on bio-diesel or ethanol; (d) tax credits for energy-efficient home improvements (e.g., insulation, double glass windows, solar panels).

PRO: provides incentives to achieve greater fuel efficiency, reduce greenhouse emissions.

CON: may not have significant impact; favors foreign auto imports; increases federal debt.

3. Vehicle tax: Annual road tax of 1\$-per-pound of weight for all light-duty vehicles, no tax rebate for driving to work or parking, to be used for research in fuel-efficient vehicles and bio fuels.

PRO: encourages lighter vehicles and mass transit.

CON: costs consumer; unfairly punishes multiple vehicle owners who drive little.

4. CO₂ cap: CO₂ emissions cap on electricity generation.

PRO: cuts greenhouse gases.

CON: will raise electricity price.

5. Subsidies for clean coal: Subsidies for clean coal with carbon sequestration to make coal competitive with natural gas.

PRO: cuts greenhouse gases; accelerates clean coal technology; reduces dependence on oil.

CON: subsidies discourage research, costs are uncertain, increases federal deficit.

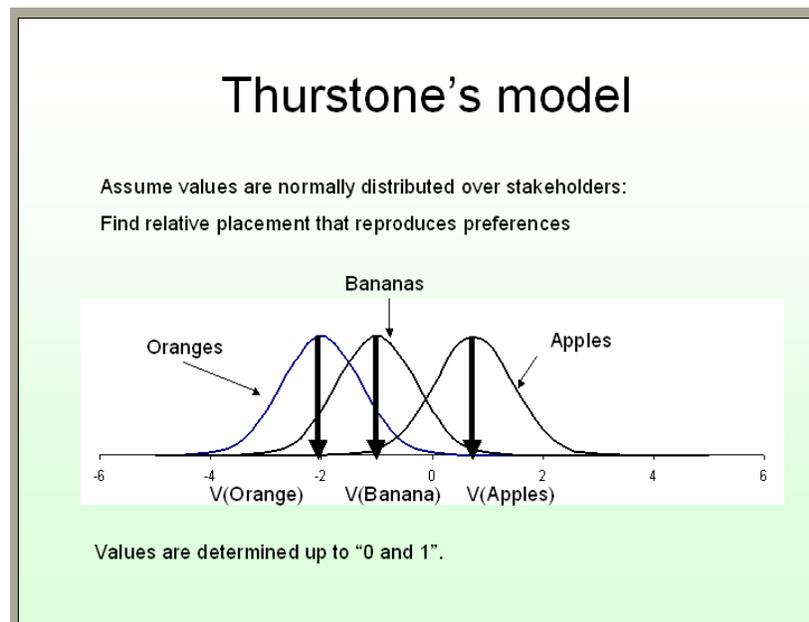
6. Do nothing.

PRO: lets the market solve the problem.

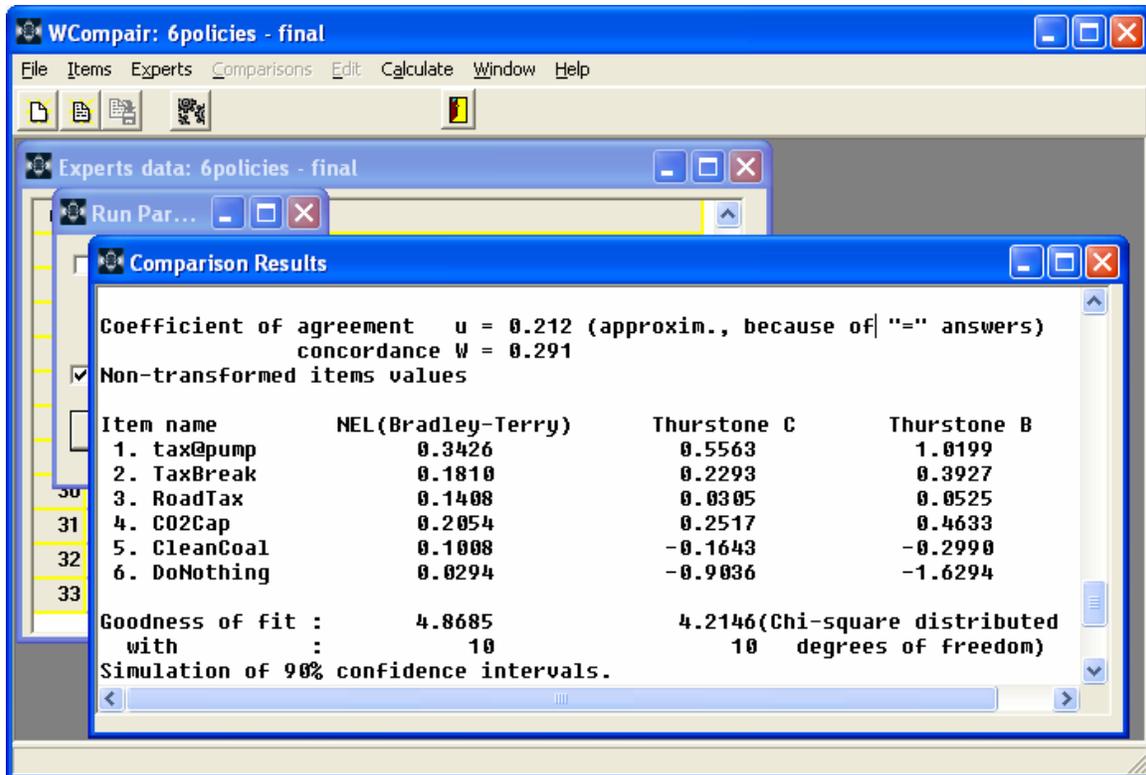
CON: the depletion of oil and global warming are problems the market can't solve.

Participants expressed their pair-wise preferences, or lack of preference for these policy options, and the data was fed into the EXCALIBUR program, WCOMPAIR. This program performs simple statistical analysis of paired comparison data and fits the Thurstone and Bradley-Terry models, briefly described below.

The “Thurstone C” model can be explained briefly as follows. We assume that true utility values for the six options exist for each stakeholder, and that these are independently distributed over the stakeholder population as normal variables with constant variance. The program then finds relative placement of the means of these distributions, such that independently sampling from these distributions would optimally reproduce the paired comparison data. The idea is illustrated graphically below for the alternatives “apples, oranges, or bananas.” According to the picture, the percentage of stakeholders preferring bananas to oranges is less than the percentage preferring apples to bananas, and both are less than the percentage preferring apples to oranges.



The “Thurstone B” model allows for constant correlation between the normal distributions, and extracts relative placement of the means. The Thurstone values are unique, up to a choice of zero and unit.



The Bradley-Terry model is based on a different statistical assumption regarding the generation of pairwise preferences and are scaled to sum to one. The conditional probability of choosing apples above oranges, given that a random stakeholder must choose one of the two, is modeled as the ratio:

Value (apples) / Value (apples) + Value (oranges).

Whichever model we choose, tax at the pump emerges as the strong favorite.

More Information Online

More information on the workshop and supporting documentation can be found at <http://www.rff.org/expertjudgmentworkshop>.