U.S. Environmental Protection Agency

Expert Elicitation Task Force White Paper

August 2011

Prepared for the
Science and Technology Policy Council
U.S. Environmental Protection Agency
Washington, DC 20460

DISCLAIMER

This Expert Elicitation White Paper was reviewed in accordance with U.S. Environmental Protection Agency (EPA) peer review policy. A draft version of this document[1] was reviewed by an independent panel of experts established by EPA's Science Advisory Board.[2]

This document represents the opinions of the Task Force on the topic of expert elicitation. It does not create or confer any legal rights for or on any person or operate to bind the public. The use of any mandatory language in this document is not intended to, and does not impose any legally enforceable rights or obligations. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

---

[1] The January 6, 2009, external review draft is available at: http://yosemite.epa.gov/sab/sabproduct.nsf/fedrgstr_activites/F4ACE05D0975F8C68525719200598BC7/$File/Expert_Elicitation_White_Paper-January_06_2009.pdf

[2] The SAB's January 21, 2010, advisory report is available at: http://yosemite.epa.gov/sab/sabproduct.nsf/fedrgstr_activites/061D0E0A700B6D4F852576B20074C2E2/$File/EPA-SAB-10-003-unsigned.pdf

Expert Elicitation Task Force

Co-Chairs

Bob Hetes	Harvey Richmond	Zachary Pekar
ORD	OAR (retired)	OAR

Chapter Leads

Chapter 1: Kathryn Gallagher, OSA (now with OCSPP)
Chapter 2: Bob Hetes, ORD
Chapter 3: Cynthia Stahl, Region 3; Lester Yuan, ORD
Chapter 4: Bob Hetes, ORD; Mark Corrales, OP
Chapter 5: Harvey Richmond, OAR (retired)
Chapter 6: Neil Stiber, OSA (now with FDA)
Chapter 7: Bob Hetes, ORD; Harvey Richmond, OAR (retired)

Task Force Members

Gary Bangs, OSA (retired)
David Chen, OCHP
Lisa Conner, OAR
Ila Cote, ORD
Joseph Greenblott, OCFO
Bryan Hubbell, OAR
Jane Leggett, OAR
Michael Messner, OW
Steve Nako, OCSPP
Barry Nussbaum, OEI
Marian Olsen, Region 2
Nicole Owens, OP
Resha Putzrath, OSA (now at Navy)
Zubair Saleem, OSWER
Brad Schultz, Region 5
Nathalie Simon, OP
David Simpson, OP
Holly Stallworth, SABSO
Brad Venner, OECA
Lester Yuan, ORD
Kathryn Gallagher, SPC Staff
Neil Stiber, SPC Staff

SPC[3] Leads

Rob Brenner (OAR, retired), William Farland (ORD, retired) and Kevin Teichman (ORD), Pai-Yei Whung (OSA on detail to World Bank)

[3] This Expert Elicitation White Paper was developed under the auspices of the former Science Policy Council. Section 1.1.1 contains additional details about the document's history.

Table of Contents

EXECUTIVE SUMMARY **1**

1.0 INTRODUCTION **3**

1.1 CONTEXT 3
1.2 PURPOSE OF THIS WHITE PAPER 4
1.3 ORGANIZATION OF THIS WHITE PAPER 4

2.0 BACKGROUND **5**

2.1 WHEN DID EXPERT ELICITATION ORIGINATE? 7
2.2 WHY IS THERE INCREASED INTEREST IN EXPERT ELICITATION? 7
2.3 WHY IS EPA EXPLORING THE USE OF EXPERT ELICITATION? 8
2.4 IS EXPERT ELICITATION THE SAME AS EXPERT JUDGMENT? 11
2.5 WHAT IS THE EXPERIENCE WITH EXPERT ELICITATION AT EPA? 14
2.6 WHAT IS THE EXPERIENCE WITH EXPERT ELICITATION AT OTHER FEDERAL GOVERNMENT AGENCIES? 21
2.7 WHAT IS THE EXPERIENCE WITH EXPERT ELICITATION OUTSIDE THE U.S. FEDERAL GOVERNMENT? 22
2.8 WHAT EXAMPLES OF EXPERT ELICITATION ARE RELEVANT TO EPA? 23
2.9 SUMMARY 23

3.0 WHAT IS EXPERT ELICITATION? **24**

3.1 WHAT IS EXPERT ELICITATION? 24
3.2 WHY IS EXPERT ELICITATION NECESSARY? 25
3.3 WHAT DO EXPERT ELICITATION RESULTS REPRESENT? 26
3.4 WHAT ARE SOME APPROACHES FOR EXPERT ELICITATION? 28
3.5 WHAT ARE GOOD PRACTICES FOR ELICITING EXPERT JUDGMENT? 33
3.6 SUMMARY 45

4.0 WHAT THOUGHTS ABOUT APPLICABILITY AND UTILITY SHOULD INFORM THE USE OF EXPERT ELICITATION? **46**

4.1 HOW IMPORTANT IS IT TO CONSIDER UNCERTAINTY? 46
4.2 WHAT IS THE NATURE OF THE UNCERTAINTIES TO BE ADDRESSED? 52
4.3 WHAT ARE OTHER METHODS TO CHARACTERIZE UNCERTAINTY? 53
4.4 WHAT ROLE MAY CONTEXT PLAY FOR AN EE? 61
4.5 WHAT ARE THE RESOURCE IMPLICATIONS WHEN CONDUCTING AN EXPERT ELICITATION? 64
4.6 SUMMARY 68

5.0 HOW IS AN EXPERT ELICITATION CONDUCTED? **69**

5.1 WHAT ARE THE STEPS IN AN EXPERT ELICITATION? 69
5.2 WHAT ARE THE PRE-ELICITATION ACTIVITIES? 72
5.3 WHAT APPROACHES ARE USED TO CONDUCT EXPERT ELICITATIONS? 81
5.4 WHAT POST-ELICITATION ACTIVITIES SHOULD BE PERFORMED? 86
5.5 WHEN AND WHAT TYPE OF PEER REVIEW IS NEEDED FOR REVIEW OF EXPERT ELICITATION? 92

5.6. SUMMARY 92

6.0 HOW SHOULD RESULTS BE PRESENTED AND USED? 93

6.1 DOES THE PRESENTATION OF RESULTS MATTER? 93
6.2 WHAT IS THE STAKEHOLDER AND PARTNER COMMUNICATION PROCESS? 93
6.3 HOW CAN COMMUNICATIONS BE STAKEHOLDER-SPECIFIC? 94
6.4 WHAT IS IN A TECHNICAL SUPPORT DOCUMENT? 95
6.5 WHAT ARE EXAMPLES OF EFFECTIVE EXPERT ELICITATION COMMUNICATIONS? 100
6.6 HOW CAN EEs BE TRANSPARENT, DEFENSIBLE, AND REPRODUCIBLE? 113
6.7 SHOULD EXPERT JUDGMENTS BE AGGREGATED FOR POLICY DECISIONS? 114
6.8 HOW CAN EXPERT ELICITATION RESULTS AND OTHER PROBABILITY DISTRIBUTIONS BE INTEGRATED? 114
6.9 HOW CAN AN EXPERT ELICITATION BE EVALUATED *POST HOC*? 115
6.10 SUMMARY 115

7.0 FINDINGS AND RECOMMENDATIONS 116

7.1 FINDINGS 116
7.2 RECOMMENDATIONS 118

REFERENCES 122

APPENDIX A: FACTORS TO CONSIDER WHEN MAKING PROBABILITY JUDGMENTS 138

APPENDIX B: GLOSSARY 142

Acronyms

ACS	American Cancer Study
BBN	Bayesian Belief Network
CASAC	Clean Air Scientific Advisory Committee
CDF	Cumulative Density Functions
CFR	Code of Federal Regulations
EE	Expert Elicitation
EPA	Environmental Protection Agency
FACA	Federal Advisory Committee Act
IEC	Industrial Economics, Inc.
IPCC	Intergovernmental Panel on Climate Change
IQ	Intelligence Quotient
NAAQS	National Ambient Air Quality Standards
NAS	National Academy of Sciences
NCEA	National Center for Environmental Assessment
NO_x	Nitrogen Oxides
NRC	National Research Council
OAQPS	Office of Air Quality, Policy, and Standards
OAR	Office of Air and Radiation
OCFO	Office of the Chief Financial Officer
OCHP	Office of Children's Health Protection
OCSPP	Office of Chemical Safety and Pollution Prevention
OECA	Office of Enforcement and Compliance Assurance
OEI	Office of Environmental Information
OGC	Office of General Counsel
OMB	Office of Management and Budget
OP	Office of Policy
ORD	Office of Research and Development
OSA	Office of the Science Advisor
OSWER	Office of Solid Waste and Emergency Response
OW	Office of Water
Pb	Lead
PDF	Probability Density Functions
PM	Particulate Matter

PRA	Paperwork Reduction Act
RAF	Risk Assessment Forum
RIA	Regulatory Impact Analysis
RIVM	Dutch National Institute for Public Health and the Environment
SAB	Science Advisory Board
SABSO	Science Advisory Board Staff Office
SOT	Society of Toxicology
SPC	Science Policy Council
SRI	Southern Research Institute
STPC	Science and Technology Policy Council
TSD	Technical Support Document
USDOT	United States Department of Transportation
USEPA	United States Environmental Protection Agency
USNRC	United States Nuclear Regulatory Commission
USOMB	United States Office of Management and Budget

EXECUTIVE SUMMARY

This White Paper was developed by the Expert Elicitation Task Force under the auspices of the former Science Policy Council (SPC). In 2005, the SPC established this Task Force and asked it to initiate dialogue within the Agency about the use of expert elicitation (EE). EE is a systematic process of formalizing and quantifying, typically in probabilistic terms, expert judgments about uncertain quantities. EE can inform decisions by characterizing uncertainty and filling data gaps where traditional scientific research is not feasible or data are not yet available. This White Paper presents what the EE Task Force believes can constitute "good practices" for conducting an EE to obtain quantitative judgments on scientific questions, such as an uncertain quantity, or the probability of different events, relationships, or parameters. Because input from a range of internal and external of stakeholders was not formally solicited, this White Paper does not present official EPA guidelines or policy.

In the 1950s, EE emerged from the growing field of decision theory as a technique for quantifying uncertainty and estimating unobtainable values to support decision-making. Since the late 1970s, EPA has performed and interpreted EEs primarily as part of its regulatory analyses for the air pollutant program. EEs have been conducted and used by at least five other federal agencies and international organizations.

Many factors influence whether to conduct an EE. In most cases, EE is but one of several methods that could be used to characterize or address critical uncertainties or data gaps. Like other analytical tools, EE has theoretical and methodological cautions for its use. Expert judgment does not exist in a vacuum, apart from sociological and personality influences. Hence, users of expert judgments should attempt to prevent bias, consider the impact of bias on results, and document how the expert judgments were obtained and used.

There are many important factors and "good" practices to consider when conducting an EE. In general, the degree to which practices are "good" or "acceptable" depends substantively on the following: (1) clear problem definition; (2) appropriate structuring of the problem; (3) appropriate staffing to conduct EE and select experts; (4) protocol development and training, including the consideration of group processes and methods to combine judgment, if appropriate; (5) procedures to verify expert judgments; (6) clear and transparent documentation, and (7) appropriate peer review for the situation. This White Paper presents the factors that the EE Task Force believes can constitute "good" practice for EPA EEs; however, a range of approaches currently are used among EE practitioners.

How results are presented is important and may influence how decision-makers form opinions and make decisions. Pertinent findings must be abstracted differently for effective presentation to different users. The presentation of results should be part of a communication strategy that focuses on users' needs without distorting findings and being faithful to the strengths and limitations of the EE.

1.0 INTRODUCTION

1.1 CONTEXT

EE is a systematic process of formalizing and quantifying, typically in probabilistic terms, expert judgments about uncertain quantities. EE can inform decisions by characterizing uncertainty and filling data gaps where traditional scientific research is not feasible or data are not yet available. The EE process is multidisciplinary and may involve integrating empirical data with scientific judgment and identifying a range of possible outcomes and likelihoods. An important part of the EE process includes documentation of the underlying thought processes of the experts. If performed using appropriate methods and quality standards, including peer review and transparency, EE can be a reliable component of sound science.

EE has been used by federal agencies, the private sector, academia, and other groups. For example, in the 1980s, the U.S. Environmental Protection Agency's (EPA) Office of Air Quality, Planning and Standards (OAQPS) used EE to assess exposure-response relationships for lead and ozone. More recently, OAQPS used EE to analyze uncertainty in the relationship between exposures to fine particles and the annual incidence of mortality (Industrial Economics, Inc. [IEC], 2004). The Department of Energy used EE to evaluate nuclear waste and other related issues. Other uses of EE by the government and academia include cost-benefit analysis, risks associated with climate change, technology development, and food safety.

1.1.1 Role of the Former Science Policy Council

In March 2005, the Expert Elicitation Task Force (hereafter cited as "Task Force") was formed under the auspices of the SPC. The EE Task Force was charged to initiate a dialogue within the Agency about EE and facilitate future development and appropriate use of EE methods. The SPC was established in 1993 by EPA's Administrator as a council of senior managers with primary responsibility for addressing and resolving cross-program, cross-media, and interdisciplinary science policy issues. In July 2010, to reflect Administrator Lisa Jackson's strong interest in promoting technology innovation to achieve environmental and public health goals, the SPC was reconstituted as the Science and Technology Policy Council (STPC).

In 2005, the former SPC members from program and regional offices nominated staff with appropriate expertise to serve on the Task Force. In fulfillment of its charge, the Task Force developed this White Paper to provide a framework for determining the appropriate conduct and use of EE.

This White Paper was peer reviewed internally by members of the former SPC Steering Committee (senior science scientists from EPA's program and regional offices) and by additional

EPA staff with expertise in uncertainty analysis and group methods. The White Paper was peer reviewed externally by an Advisory Panel of the EPA's Science Advisory Board.[4]

1.2 PURPOSE OF THIS WHITE PAPER

The purpose of the Task Force was to initiate an intra-Agency dialogue about the conduct and use of EE and then to facilitate future development and appropriate use of EE methods. To that end, the Task Force facilitated a series of discussions to familiarize Agency staff with EE and to evaluate and address issues that may arise from its use. This White Paper reflects those discussions and presents pertinent issues, including: What is EE? When should the use of EE be considered? How is an EE conducted? And, how should results be presented and used?

Because input from a range of internal and external of stakeholders was not formally solicited, this White Paper does not present official EPA guidelines or policy.

1.3 ORGANIZATION OF THIS WHITE PAPER

This White Paper reflects discussions about EE that were coordinated by the Task Force. Chapter 2 provides background for EE at EPA and summarizes the context of increasing interest in this approach. In addition, it reviews experiences with EE at EPA, throughout the federal government, and with international groups. It also shares applications that are relevant to EPA issues. Chapter 3 provides the definition of EE for this White Paper and considers its advantages and disadvantages. Chapter 4 recognizes that EE is one of many tools to characterize uncertainty and examines what factors may help determine when EE is appropriate. Chapter 5 summarizes what is needed to conduct a credible and acceptable EE. Chapter 6 offers considerations for presenting and using EE results in EPA decisions. Finally, Chapter 7 presents some significant issues regarding EE. Where consensus was reached by the Task Force, the White Paper provides recommendations for further development and use of EE within EPA or by parties submitting EE assessments to EPA for consideration.

[4] The SAB's January 21, 2010, advisory report is available at: http://yosemite.epa.gov/sab/sabproduct.nsf/fedrgstr_activites/061D0E0A700B6D4F852576B20074C2E2/$File/EPA-SAB-10-003-unsigned.pdf

2.0 BACKGROUND

EPA often makes decisions on complex environmental issues that require analyses from a broad range of disciplines. Among the many sources of uncertainty and variability in these analyses are estimates of parameter values and choices of models. Furthermore, in some cases critical data may be unavailable or inadequate. Although the presence of uncertainty complicates environmental decision-making, EPA still must make timely decisions. Accordingly, EPA routinely considers and addresses uncertainty when appropriate in its decisions. EE is one of many methods for characterizing uncertainty and providing estimates in data-poor situations. EPA guidance has recognized that a frank and open discussion of uncertainty is critical to the full characterization of risk (i.e., *Risk Characterization Handbook*, USEPA, 2000a). Even though this handbook was developed with risk assessment in mind, it is applicable and useful for many kinds of EPA assessments and methods, including EE.

In 1983, the National Academy of Sciences (NAS) published *Risk Assessment in the Federal Government: Managing the Process* (NAS, 1983; commonly referred to as the "Red Book"), which formalized the risk assessment process. EPA integrated the "Red Book" principles of risk assessment into its practices and, the following year, published *Risk Assessment and Management: Framework for Decision Making* (USEPA, 1984), which emphasizes making the risk assessment process transparent by fully describing the assessment's strengths and weaknesses and addressing plausible alternatives. Then, starting in 1986, EPA began issuing a series of guidelines for conducting risk assessments (e.g., exposure, carcinogen, chemical mixtures, mutagenicity, and suspect developmental toxicants, USEPA, 1986a-e). Although EPA's initial efforts focused on human health risk assessment, in the 1990s the basic approach was adapted to ecological risk assessment to address a broad array of environmental risk assessments in which human health impacts are not a direct issue. EPA continues to make a substantial investment in advancing the science and application of risk assessment through updates to these guidelines and the development of additional guidelines, as needed.

Over the next several years, the NAS expanded on its "Red Book" risk assessment principles in a series of subsequent reports, including *Pesticides in the Diets of Infants and Children* (NAS, 1993), *Science and Judgment in Risk Assessment* (NAS, 1994), and *Understanding Risk: Informing Decisions in a Democratic Society* (NAS, 1996). The purpose of the risk assessment process, as characterized by the NAS, is to ensure that assessments meet their intended objectives and are understandable. Over time, EPA risk assessment practices advanced along with NAS's progression of thought.

In 1992, EPA provided the first risk characterization guidance to highlight the two necessary elements for full risk characterization: (1) address qualitative and quantitative features of the assessment and (2) identify any important uncertainties and their influence as part of a discussion on confidence in the assessments. Three years later, EPA updated and issued the current Agency-wide *Guidance for Risk Characterization* (USEPA, 1995a). To ensure that the risk assessment process is transparent, this guidance calls for risk characterization for all EPA risk assessments. In addition, this guidance emphasizes that risk assessments be clear, reasonable, and consistent with other risk assessments of similar scope across the Agency. Effective risk characterization should feature transparency in the risk assessment process and clarity, consistency, and reasonableness of the risk assessment product. EPA's *Risk Characterization Handbook* (USEPA, 2000a) was developed to implement the *Guidance for Risk Characterization*. The importance of characterizing uncertainty was re-affirmed in the recent EPA staff paper, *An Examination of EPA Risk Assessment Principles and Practices* (USEPA, 2004a). This staff paper identified the use of probabilistic analyses as an area in need of major improvement.

Risk assessments often are used as the basis for calculating the benefits associated with Agency regulations. Such benefits-costs analyses can be important tools for decision-makers, where statutorily permitted, both in the context of regulatory reviews required under Executive Order (EO) 12866 and Section 812 of the Clean Air Act Amendments of 1990, which require EPA to assess the costs and benefits of the Clean Air Act. In its 2002 report, *Estimating the Public Health Benefits of Proposed Air Pollution Regulations*, the NAS emphasized the importance of fully characterizing uncertainty for decision-makers and encouraged EPA to use EE in the context of expressing uncertainty associated with estimated benefits.

Guidance for conducting regulatory analyses as required under EO 12866 is provided in the Office of Management and Budget's (OMB) *Circular A-4* (USOMB, 2003b). This guidance emphasizes that the important uncertainties connected with regulatory decisions need to be analyzed and presented as part of the overall regulatory analysis. Whenever possible, appropriate statistical techniques should be used to determine the probability distribution of relevant outcomes. For major rules involving annual economic effects of $1 billion or more, Circular A-4 calls for a formal quantitative analysis of uncertainty. OMB guidelines outline analytical approaches, of varying levels of complexity, which could be used for uncertainty analysis, such as qualitative disclosure, numerical sensitivity analysis, and formal probabilistic analysis (required for rules with impacts greater than $1 billion). EE is one of the approaches specifically cited in these guidelines for generating quantitative estimates (e.g., cost-benefit analysis) when specific data are unavailable or inadequate.

2.1 WHEN DID EXPERT ELICITATION ORIGINATE?

The origins of EE can be traced to the advent of decision theory and decision analysis in the early 1950s. In 1954, Savage established the "probabilities of orderly opinions," which states that the choice behavior of a rational individual can be represented as an expected utility with a unique probability and utility measure. EE's development also drew on the operational definitions of probability that arose out of the semantic analysis discussions of Mach, Hertz, Einstein, and Bohr (Cooke, 1991). Since the early 1970s, decision analysts in the private and public sectors have used formal EE processes to obtain expert judgments for their assessments. Section 2 presents a variety or examples from EPA, other federal agencies, and beyond.

2.2 WHY IS THERE INCREASED INTEREST IN EXPERT ELICITATION?

There are numerous quantitative methods for characterizing uncertainty. The available types of methods are described briefly Section 4.2. Although there is no consensus on a preferred method to characterize uncertainty, there is general agreement that practitioners should describe uncertainty to the extent possible with available data and well-established physical and statistical theory. Limitations in data and/or understanding (i.e., lack of a theory relevant to the problem at hand), however, may preclude the use of conventional statistical approaches to produce probabilistic estimates of some parameters. In such cases, one option is to ask experts for their best professional judgment (Morgan and Henrion, 1990). EE (which is defined and discussed in greater detail in Chapter 3) is a formal process by which expert judgment is obtained to quantify or probabilistically encode uncertainty about some uncertain quantity, relationship, parameter, or event of decision relevance.

EE is recognized as a powerful and legitimate quantitative method for characterizing uncertainty and providing probabilistic distributions to fill data gaps where additional research is not feasible. The academic and research community, as well as numerous review bodies, have recognized the limitation of empirical data for characterization of uncertainty and have acknowledged the potential for using EE for this purpose. In *Science and Judgment in Risk Assessment* (NAS, 1994), the NAS recognized that for "parameter uncertainty, enough objective probability data are available in some cases to permit estimation of the probability distribution. In other cases, subjective probabilities might be needed." In its report, the NAS further recognized the "difficulties of using subjective probabilities in regulation" and identified perceived bias as one major impediment, but noted that "in most problems real or perceived bias pervades EPA's current point-estimate approach." In addition, the NAS stated that "there can be no rule that objective probability estimates are always preferred to subjective estimates, or vice versa."

The utility of EE has been discussed by NAS, OMB, and EPA. In the following examples, they provide advice for the appropriate and beneficial use of EE. With respect to benefits analyses, NAS (2002) recommends:

> "EPA should begin to move the assessment of uncertainties from its ancillary analyses into its primary analyses by conducting probabilistic, multiple-source uncertainty analyses. This shift will require specifications of probability distributions for major sources of uncertainty. These distributions should be based on available data and *expert judgment*."

In its *Circular A-4* (USOMB, 2003b), OMB suggests using EE to address requirements for probabilistic uncertainty analysis:

> In formal probabilistic assessments, *expert solicitation*[5] is a useful way to fill key gaps in your ability to assess uncertainty. In general, experts can be used to quantify the probability distributions of key parameters and relationships. These solicitations, combined with other sources of data, can be combined in Monte Carlo simulations to derive a probability distribution of benefits and costs.

In addition, the EPA *Guidelines for Carcinogen Risk Assessment* (USEPA, 2005a) provide for the use of EE in such assessments:

> In many of these scientific and engineering disciplines, researchers have used rigorous expert elicitation methods to overcome the lack of peer-reviewed methods and data. Although expert elicitation has not been widely used in environmental risk assessment, several studies have applied this methodology as a tool for understanding quantitative risk. … These cancer guidelines are flexible enough to accommodate the use of expert elicitation to characterize cancer risks, as a complement to the methods presented in the cancer guidelines. According to NAS (NAS, 2002), the rigorous use of expert elicitation for the analyses of risks is considered to be quality science.

2.3 WHY IS EPA EXPLORING THE USE OF EXPERT ELICITATION?

EPA recognizes the value of EE as a powerful tool that may enhance the characterization of uncertainty in risk and other types of assessments. EPA's experience with EE (described in Section 2.5) highlights its benefits of enhancing the scientific and technical credibility of EPA assessments and acceptability of these assessments within the scientific and technical community (NAS, 2002). Concerns, however, have been raised about using EE within the context of EPA decision-making. These include transparency in the use of empirically versus judgment-derived estimates, potential for delays in rulemaking while EE is conducted, and the lack of EPA guidelines on and limited experience with EE. The American Bar Association (2003), in its

[5] OMB used the phrase "expert solicitation" rather than "expert elicitation," but text and references are similar to that associated with expert elicitation. For the purposes of this White paper, EPA assumes that they are equivalent.

comments on the requirement for formal probabilistic analyses in OMB's *Draft 2003 Report to Congress on the Costs and Benefits of Federal Regulation* (OMB, 2003a), stated:

> …formal probabilistic analysis will be impossible to meet rigorously in cases where the underlying science is so uncertain as to preclude well-founded estimates of the underlying probability distribution…. In such situations, the effort to generate probability distributions in the face of fundamental uncertainty through guesses derived from so-called 'expert elicitation' or 'Delphi' methods runs the risk of creating that 'false sense of precision' which OMB elsewhere cautions agencies to avoid. Accordingly, we believe such methods should be used sparingly, and we strongly endorse the recent recommendation of the National Research Council that agencies disclose all cases in which expert elicitation methods have been used.

EPA's limited experience conducting EEs has primarily been within the Office of Air and Radiation (OAR), mainly focused on improving risk and benefits assessments but not directly used to support a regulatory decision. EPA has no clear guidelines to assist in the conduct and use of such techniques for regulatory analyses or other purposes. Given these limitations, the Task Force believed that EPA would benefit from a thoughtful discussion about the conduct and use of EE to support regulatory and nonregulatory analyses decision-making. Future efforts to apply EE should include a dialogue on their regulatory, legal, and statutory implications as well as their technical aspects.

Given this background, the Task Force initiated a dialogue within the Agency to consider a broad range of technical, statutory, regulatory, and policy issues including:

- When is EE an appropriate (well-suited) methodology to characterize uncertainty?
- What are good practices based on a review of the literature and actual experience within EPA and other federal agencies in conducting an EE, considering the design objectives and intended use of the results (e.g., prioritizing research needs, input to risk assessment, input to regulatory impact analysis [RIA])?
- When, and under what circumstances, is it appropriate to aggregate/combine expert judgments? How should such aggregation/combination be conducted?
- When in the EE process is peer review beneficial?
- What type of peer review is needed to review EE methods and their use in specific regulatory actions?
- What are the implications of EPA's Quality System and Information Quality Guidelines on EE?

2.3.1 How Does this Expert Elicitation Activity Relate to Efforts to Develop and Promote Probabilistic Risk Assessment at EPA?

A major recommendation in the 2004 EPA staff paper on *Risk Assessment Principles and Practices* (USEPA, 2004a) was more *appropriate and timely use* of probabilistic assessments. As a result, EPA initiated several efforts to promote the appropriate use of probabilistic analyses in support of regulatory analyses, including activities sponsored by the Risk Assessment Forum (RAF). These major activities included a RAF Colloquium on Probabilistic Risk Assessment, follow-up workgroup activities, and EPA's co-sponsorship with the Society of Toxicology (SOT) of a Workshop on Probabilistic Risk Assessment.

In April 2004, EPA's RAF held a Colloquium on Probabilistic Risk Assessment to address the following topics: identifying probabilistic techniques that can better describe variability and uncertainty, communicating probabilistic methods for the purpose of risk management and risk communication, supplementing the *Guiding Principles for Monte Carlo Analysis* (USEPA, 1997), and deciding the next steps for advancing probabilistic methods in risk assessments or improving their implementation. As a follow-up to the 2004 Colloquium, the RAF formed a workgroup to address how to improve support for EPA decision-making through the use of probabilistic methods. One of this workgroup's first activities addressed a major and multidisciplinary recommendation from the 2004 RAF Colloquium: the need for a dialogue between risk assessors and decision-makers (risk managers). These discussions are essential to identify specific issues of concern and determine needs for promoting the useful application of these methods. Without these upfront understandings, probabilistic methods might be applied at high resource cost to EPA but provide information that is irrelevant and/or has little or no impact on decisions.

As a follow up to the staff paper on *An Examination of Risk Assessment Principles and Practices* (USEPA, 2004a), EPA co-sponsored, with the SOT and several other organizations, a Workshop on Probabilistic Risk Assessment as part of the SOT's Contemporary Concepts in Toxicology workshop series. The workshop provided an opportunity for in-depth discussion of four critical topic areas: (1) exposure assessment, (2) ecological risk assessment, (3) human health risk assessment and medical decision analysis, and (4) decision analysis/multicriteria decision analysis/cost-benefit analysis. Draft white papers for each topic area that were prepared for and discussed at the workshop were submitted for journal publication. EE was discussed as a critical method for advancing probabilistic analyses in the Agency.

It also should be noted that individual program and regional offices have, in the past and may in the future, developed their own guidance on the conduct and use of probabilistic methods. For example, EPA's Office of Solid Waste and Emergency Response developed

guidance for this purpose, *Risk Assessment Guidance for Superfund: Volume 3—Part A: Process for Conducting Probabilistic Risk Assessment* (USEPA, 2001a).

The Task Force intends to complement the above efforts and developed this White Paper to serve as a discussion guide.

2.4 IS EXPERT ELICITATION THE SAME AS EXPERT JUDGMENT?

EPA often needs to make decisions that address complex problems, many of which lack direct empirical evidence. Such decisions require judgment to assess the impact and significance of existing data or theory. As a result, judgment is an inherent and unavoidable part of most EPA assessments and decisions. Judgment also is inherent in empirical data and highly targeted technical activities. For example, the analyst's expert judgment is critical and unavoidable when developing and selecting a study design, a sampling strategy, specific statistical tests, goodness-of-fit measures, and rules for excluding outliers. EPA relies on various forms of professional, scientific, and expert judgment throughout the scoping, design and implementation of its assessments. EPA's *Risk Characterization Handbook* (USEPA, 2000a) recognizes the role and importance of professional and scientific judgment in the risk assessment process and provides guiding principles for fully and transparently describing all information, including judgments, used to assess and characterize risks. As illustrated in Figures 2-1 and 2-2, decision-making at EPA concerns complex problems that are influenced by multiple factors. Even within a particular discipline or activity, inherent expert judgment may include values or preferences (e.g., default to a health-protective conservative, or less conservative, estimate of risk). In addition to risk assessment results, social, economic, political, statutory/legal, public health, and technological factors may influence a final decision. Therefore, scientific (state of knowledge) and value and preferences are both included in any particular assessment.

Although values and preferences, including risk perception, play a major role in how to weight or balance information across various disciplines, expert judgment can help to integrate this information. Tools that rely on formal integration of values and preferences, however, are not within the Task Force's definition of EE.

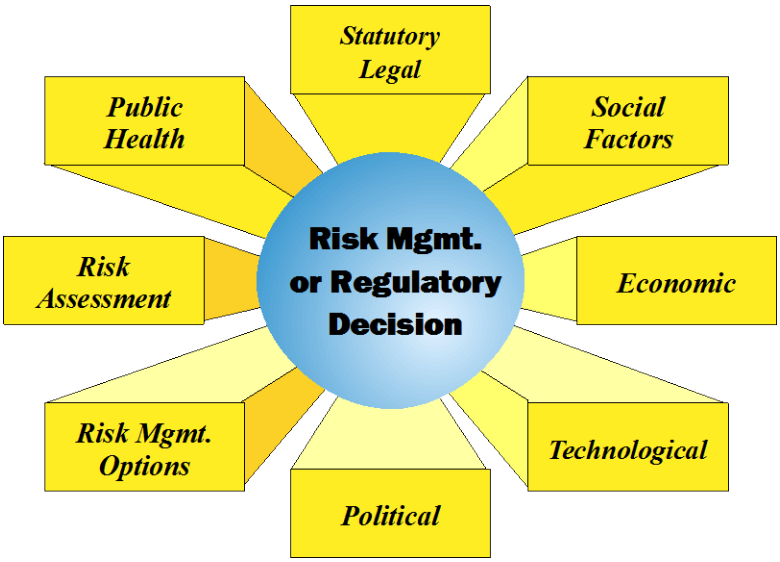


Figure 2-1. Factors That Influence Risk Management Planning at EPA

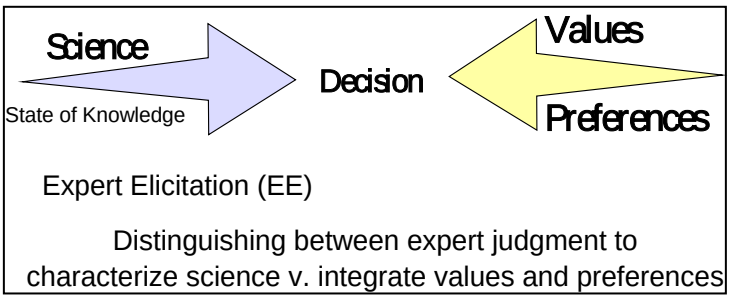


Figure 2-2. The Role of Science and Values in Environmental Decision Making

For the purposes of this White Paper, the Task Force has chosen to define EE more narrowly than "expert judgment"— as a method limited to characterizing the science (state of knowledge)—rather than a method focused on characterizing values and preferences *per se*. This definition is consistent with the field of EE, and practitioners in the field generally advocate a separation of the analyses of uncertainty associated with quantitative physical parameters or empirical models from the phase involving preferences and social value judgments leading up to decisions. Experts are best suited for answering science questions, whereas appointed or elected decision-makers (with input from risk managers and appropriate stakeholders) are best suited for providing values and preferences.

EPA also relies on various forms of expert input in formulating its judgments through external review and evaluation of the quality of its assessments. These include internal and

external peer review and peer involvement such as through the EPA Science Advisory Board (SAB). In these cases, reviewers typically bring technical expertise as well as experience. Such experts rely on peer-reviewed published literature, as well as their own knowledge (e.g., unpublished studies) and perception. The distinguishing feature of EE relative to these other forms of expert input is the use of a systematic process to formalize judgments. Expert input via these other mechanisms provides expert opinions, and although some may take procedural steps to minimize the effects of heuristics or other biases in expert judgments, they generally do not apply formal scientific protocol to address these effects.

In considering the range of approaches to include in EE, the Task Force relied on the volume of decision analysis literature and history that are derived from or consistent with Bayesian theory.[6] For an effort to be considered as EE within the scope of this White Paper, at a minimum, all of the following elements[7] (as described in detail later) must be present:

- Problem definition—unambiguous, which meets Clairvoyance Test.[8,9]

---

[6] Bayesian theory, discussed in greater detail in Chapter 3 as it relates to EE, proposes that a person's belief in a proposition can be described according to probability theory.

[7] As described in later chapters, there are other elements that are described as components of EE. Many of the elements are not unique to EE but may be associated in some form with others modes of expert input. The ones listed here are suggested as the Task Force's minimum distinguishing operational features for EE. This does not imply that an exercise containing these minimum elements would represent a good EE. How well these elements are carried will determine the overall quality of any particular effort.

[8] Such that a hypothetical omniscient being with complete knowledge of the past, present, and future could answer the question definitively and unambiguously. The importance of the Clairvoyance Test can be illustrated by a sample dialogue between a clairvoyant and a homeowner who is concerned about environmental contamination that may expose children to lead from the soil in their yard.

> *Homeowner: What is the average concentration of lead (Pb) in the surface soil at my home?*
> Clairvoyant: I know where your home is, but you need to better specify your question so that I can answer it. By "at my home," do you mean "within my legal property boundaries"?
> *Homeowner: Yes. What is the average concentration of lead within the legal property boundaries of my home?*
> Clairvoyant: That is more precise, but I still need more information. By the way, "average concentration" is somewhat redundant. I prefer "concentration." Tell me, what do you mean by surface soil?
> *Homeowner: The top 2 inches, all around my home.*
> Clairvoyant: Including the paved areas?
> *Homeowner: No. I'm interested in my kids' exposure. They would not dig below the pavement.*
> Clairvoyant: I assume you want the concentration in units of mg/kg, but you should know that this varies from day to day because of moisture content.
> *Homeowner: I want to know the current concentration (today at this time), whatever the current moisture content. Tell me the answer in mg/kg soil, where the soil mass includes the current moisture content.*
> Clairvoyant: Thank you. Now we have a well-specified question:
>> What is the current concentration (mg/kg) of lead in the top 2 inches of surface soil within the legal property boundaries of my home, excluding paved areas?

[9] Additionally the problem should be clearly defined such that the quantity in question is measurable (at least in principle, if not necessarily in practice).

- Formal protocol—required to ensure consistency in elicitation and control for heuristics and biases.
- Identification, summary, and sharing of the relevant body of evidence with experts.
- Formal elicitation—encoding of judgments as probabilistic values or distributions—typically via interaction with objective independent party.
- Output—judgment or degree of belief is expressed quantitatively (typically probabilistically).

The Task Force's selection of these factors was influenced by the available literature and expertise on the Task Force. The Task Force did not fully consider other non-Bayesian, nonquantitative, semiquantitative, or nonprobabilistic social science encoding methods that also control for heuristics and biases (e.g., Grounded Theory, Nominal Group Technique, and variants of the Delphi method). Nevertheless, such methodologies may be appropriate for formal elicitation of expert judgments where quantitative (i.e., probabilistic) characterization of judgment is not the desired output, if they fulfill the analytical requirements of the charge and are implemented with sufficient rigor, transparency, and review.

2.5 WHAT IS THE EXPERIENCE WITH EXPERT ELICITATION AT EPA?

The majority of EPA's experience with EE is in its OAR OAQPS. OAQPS first explored the use of EE in the late 1970s. In fact, NAS (2002) notes that "OAQPS has been a pioneer in the application of (use of expert judgment) approaches to estimating the health risks due to exposure to air pollutants (Richmond, 1991; Feagans and Biller, 1981; Whitfield et al., 1991; Rosenbaum et al., 1995)." As summarized below, EPA's experience includes OAQPS studies to evaluate the health effects of various criteria pollutants and other EPA offices efforts to forecast sea level rise, store radioactive waste, and support ecological model development.

2.5.1 Criteria Pollutants

2.5.1.1 1977–1978 Ozone National Ambient Air Quality Standards (NAAQS) Review

Motivated by the statutory requirement to protect public health with an "adequate margin of safety," OAQPS pursued development of probabilistic estimates. As part of the 1977–1978 ozone NAAQS review (Feagans and Biller, 1981), OAQPS drew on techniques developed within the field of decision analysis (probability encoding) to derive probabilistic concentration-response relationships from experts for several health endpoints. These early efforts were based on scant experience and lacked formal protocols as well as pre- and post-elicitation workshops.

To review this OAQPS approach and EPA's efforts to develop probabilistic risk assessment methods, the SAB Executive Committee created the SAB Subcommittee on Health

Risk Assessment in 1979. This SAB Subcommittee held several meetings from 1979 to 1981 to review reports prepared by six teams of nationally recognized experts, additional developmental efforts by OAQPS, a literature review of probability encoding (Wallsten and Budescu, 1983), and two illustrative applications involving EEs for health effects associated with carbon monoxide. In spring of 1981, the SAB encouraged EPA to take the best elements of the illustrative applications and the original OAQPS-proposed approach in carrying out an EE addressing health effects of lead (Pb) for EPA's Pb NAAQS risk assessment.

2.5.1.2 Lead NAAQS Risk Assessment (1983-1986)

Following the advice of the SAB, OAQPS sponsored a full EE on the health effects of Pb to support the Pb NAAQS review. OAQPS developed and pilot tested a formal protocol. The actual elicitation focused on two health endpoints (IQ decrement and hemoglobin decrement), and its results (Figure 2-3) received a favorable review from the Clean Air Scientific Advisory Committee (CASAC) Pb Subcommittee. The elicitation was crucial in deriving and characterizing the uncertainty in policy-relevant concentration response functions beyond those available in the empirical literature. Although the SAB Health Risk Assessment Subcommittee was dissolved following its review of the elicitation project, three of its members were added to CASAC Pb Subcommittee.

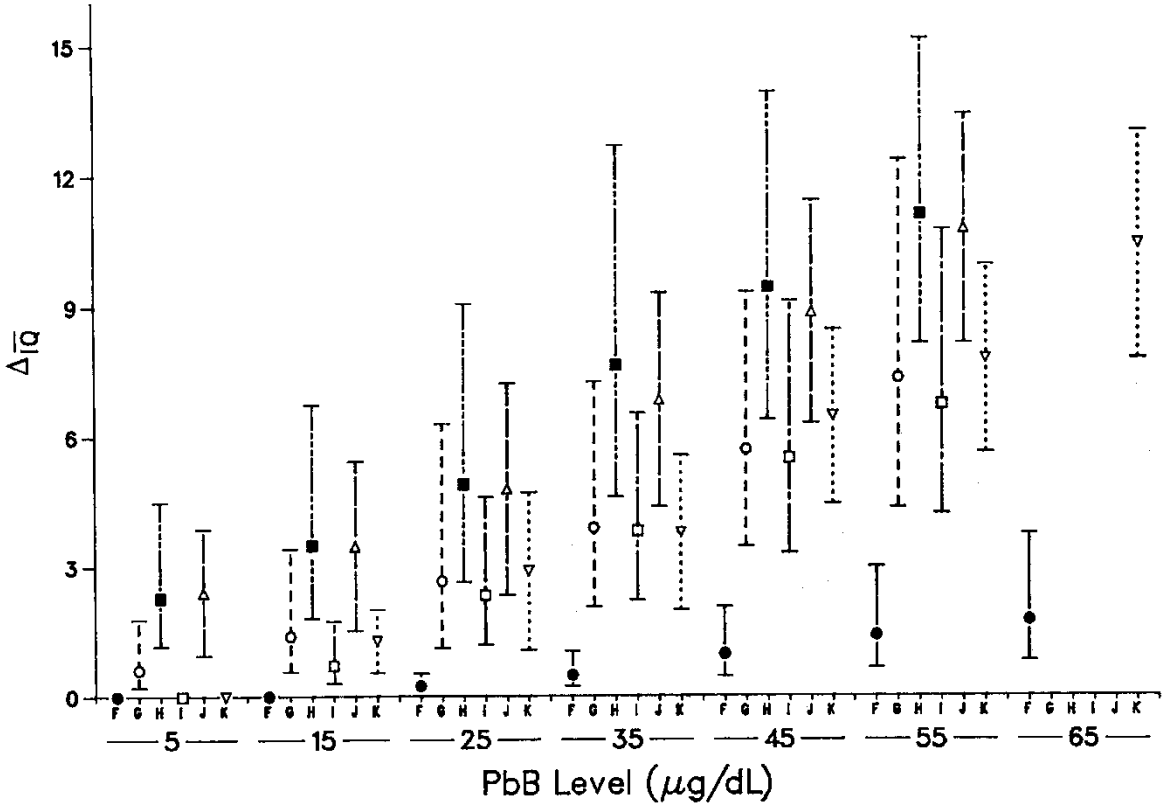


Figure 2-3. Median and 90% Credible Interval Judgments of All Experts About Lead-Induced IQ Decrements Among Low-SES Children Aged 7 Years (Whitfield and Wallsten, 1989)

2.5.1.3 Ozone Chronic Lung Injury Assessment

Drawing on its experience with the Pb NAAQS assessment, OAQPS pursued an additional study to develop concentration-response functions for chronic lung injury associated with long-term exposures to ambient ozone concentrations (Winkler et al., 1995). The specific objective was to characterize scientific judgments regarding the risk of chronic lung injury to children aged 8 to 16 years and adult outdoor workers as a result of long-term ozone exposure in areas with exposure patterns similar to Southern California and the Northeast. Again, a formal protocol was developed and pilot-tested prior to the elicitation exercise. Experts were provided with air quality information, exposure model estimates, and dosimetry model estimates. The measure of injury was the incidence of mild or moderate lesions in the centriacinar region of the lung. Probabilities of population response rates were elicited. After a postelicitation workshop to encourage information exchange among experts, a second round of encoding was conducted (Figure 2-4).

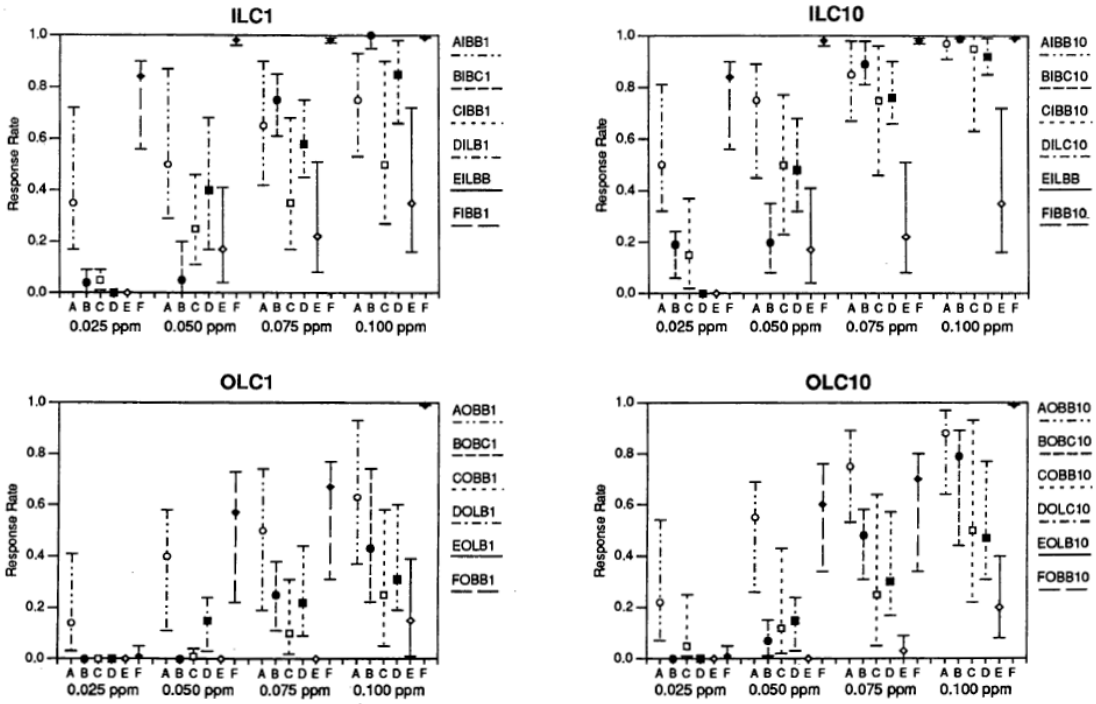


Figure 2-4. Medians and 90% Credible Intervals for Judgments About Mild and Moderate Lesions, Children, Los Angeles, and 1 and 10 Ozone Seasons (Whitfield et al., 1994)

2.5.1.4 Lessons Learned from Early OAQPS Efforts

The use of EE in risk assessment is not necessarily straight-forward. When EE was fairly new, much of the scientific community was skeptical. Early efforts were controversial and questioned, in part, because of the lack of both experience and formal procedures. Through a lengthy collaborative effort with the SAB, OAQPS was able to improve the quality of the assessments and increase their credibility within the scientific community. Although OAQPS EEs have gained acceptance, it is likely that similar collaborative efforts will be needed for EE method development and application sponsored by other EPA offices.

The results of these initial elicitations informed policy-makers and the public of possible health implications as a result of short- and long-term exposures to specific criteria pollutants. These results were not used as the basis for any EPA regulatory decisions, however.[10]

2.5.1.5 Particulate Matter (PM) Concentration-Response (C-R) for Mortality

[10] No formal regulatory decision was made following the completion of the Pb risk assessment for reasons other than EE, and the EE was used to develop the chronic ozone lung injury assessment but was not intended to support the ozone NAAQS review decision.

As mentioned in Section 2.2, the NAS (2002) recommended that EPA improve its characterization of uncertainty in its benefits analyses by using both available data and expert judgment. It recommended that EPA build on the prior OAQPS experiences in formally elicited expert judgments but noted that a number of issues must be addressed. The NAS stressed that EPA should distinguish clearly between data-derived components of an uncertainty assessment and those based on expert opinions. As a first step in addressing these NAS recommendations, EPA collaborated with OMB to conduct a pilot EE to characterize uncertainties in the relationship between ambient fine particles ($PM_{2.5}$) and premature mortality. This pilot EE provided EPA with an opportunity to improve its understanding of the design and application of EE methods to economic benefits analysis. The results of the pilot EE were presented in RIAs for both the Nonroad Diesel and Clean Air Interstate Rules (USEPA, 2004b, 2005b).

The collaboration with OMB was linked to the RIA for the final Nonroad Diesel Rule and thus required completion within 1 year. The scope of the pilot was limited to focus on the concentration-response function of PM mass rather than on individual issues surrounding an estimate of the change in mortality as a result of PM exposure. A consequence of the limited time for completion of the pilot was that certain aspects of a more comprehensive EE process were eliminated (e.g., neither pre- nor postelicitation workshops were held), and some aspects of the uncertainty surrounding the $PM_{2.5}$-mortality relationship could not be characterized. In addition, to meet time constraints for the pilot EE, experts were selected from two previously established NAS expert panels.

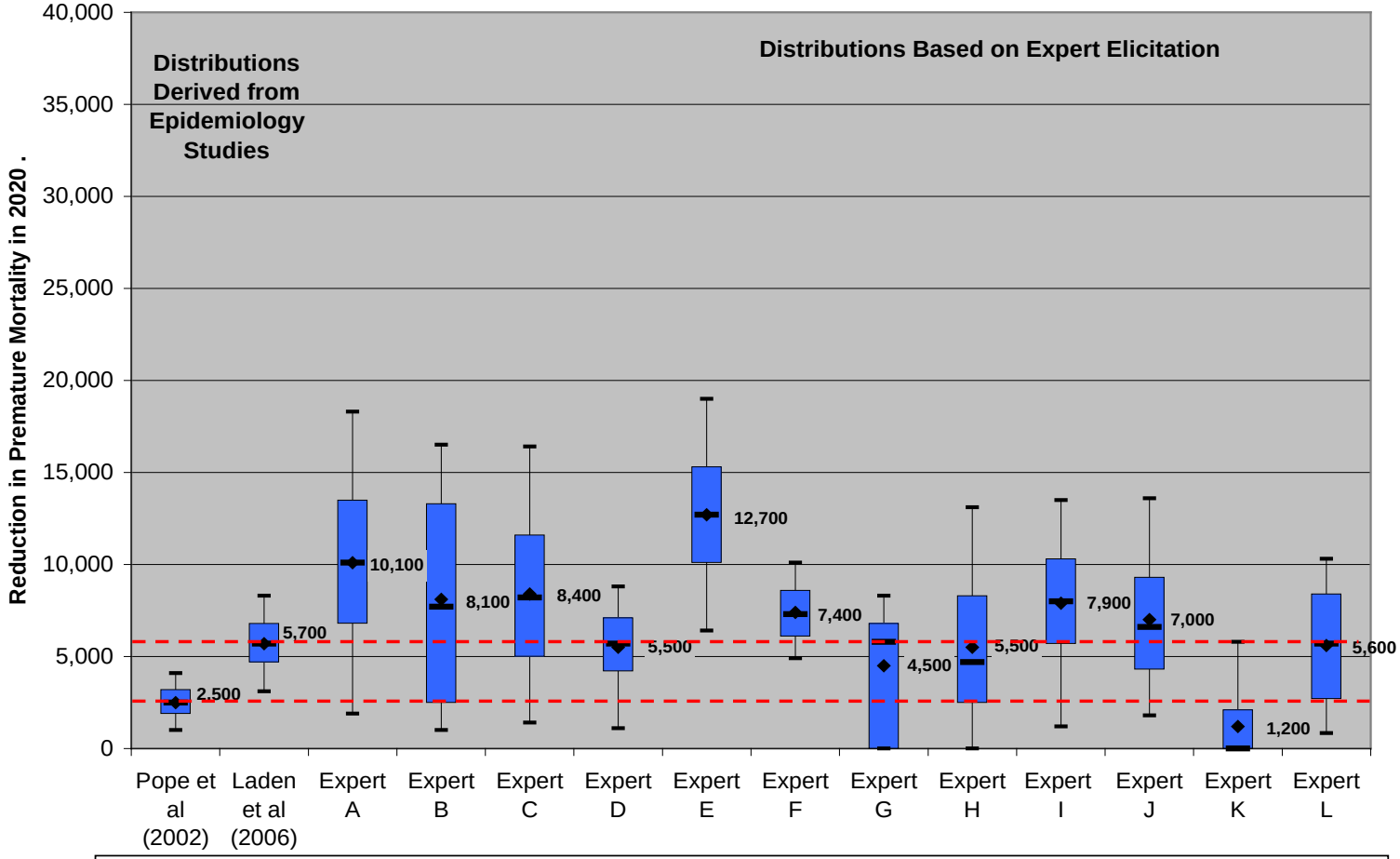
The plan for assessment and the draft protocol was initially reviewed by the Health Effect Subcommittee of the Council on Clean Air Compliance Analysis. The protocol was pilot-tested with EPA and non-EPA PM health scientists who were not involved in the final elicitation process. The project team that carried out the assessment consisted of individuals with experience in EE and individuals with expertise in PM health effects and health benefits.

EPA and OMB conducted an external peer review of the methods used in this pilot EE. In accordance with EPA's peer review guidelines (USEPA, 2005a), this peer review also considered the approaches to presenting the results (particularly with respect to combining results across experts).

Based on the experience gained from the pilot EE, EPA completed a full-scale EE that incorporated peer-review comments on the pilot application. This provided a more robust characterization of the uncertainty in the premature mortality function. The full-scale PM-mortality elicitation included an in-depth review of the protocol design, drew from a larger pool of experts using a peer-nomination process, and allowed for increased communication among

experts and the project team via pre- and postelicitation workshops. The PM-mortality elicitation was designed to evaluate uncertainty in the underlying causal relationship, the form of the mortality impact function (e.g., threshold versus linear models), and the fit of a specific model to the data (e.g., confidence bounds for specific percentiles of the mortality effect estimates). Additional issues, such as the ability of long-term cohort studies to capture premature mortality resulting from short-term peak PM exposures, were also addressed in the EE. As with the pilot EE, the full-scale PM-mortality elicitation underwent extensive review with internationally renowned PM experts, EPA management, and OMB. The pilot and full-scale EEs can be found at: http://epa.gov/ttn/ecas/benefits.html.

The findings from the PM-mortality elicitation were presented in the *Regulatory Impact Analysis of the Final PM National Ambient Air Quality Standards* (USEPA, 2006b) and in the Proposed Ozone NAAQS (USEPA, 2007). Figure 2-5 presents a depiction of results from the PM NAAQS benefit analysis, showing box plots of the distributions of the reduction in $PM_{2.5}$-related premature mortality based on the C-R distributions provided by each expert, as well as that from the data-derived health impact functions, based on the statistical error associated with Pope et al. (2002) and Laden et al. (2006). Each distribution is depicted as a box plot with the diamond symbol (♦) showing the mean, the dash (–) showing the median (50th percentile), the box defining the interquartile range (bounded by the 25th and 75th percentiles), and the whiskers defining the 90 percent confidence interval (bounded by the 5th and 95th percentiles of the distribution). The RIA also utilizes a variety of other formats for presenting the results of the elicitation, including: tables, bar graphs, and cumulative distribution functions.



Note: Distributions labeled Expert A - Expert L are based on individual expert responses. The distributions labeled Pope et al. (2002) and Laden et al (2006) are based on the means and standard errors of the C-R functions from the studies. The red dotted lines enclose a range bounded by the means of the two data-derived distributions.

Key: Closed circle = median; Open circle = mean; Box = interquartile range; Solid line = 90% credible interval

Figure 2-5. Results of Application of Expert Elicitation: Annual Reductions in the Incidence of PM-Related Mortality in 2020 Associated With the Final National Ambient Air Quality Standards for Particulate Matter

In presenting these results, EPA was sensitive to the NAS's advice to clearly label estimates based on the results of the EE (as opposed to empirical data). In addition, EPA addressed NAS's concerns that EE results be presented within the context of describing the uncertainty inherent in the concentration-response function. Recent RIAs have described the C-R functions based on the EE and considered whether these EE results should replace the primary estimated response based on analysis of the American Cancer Study (ACS) cohort (Pope et al.,

2002). This ACS cohort study was recommended by the Agency's SAB specifically for use in the 812A analyses of benefits associated with the Clean Air Act.

EPA has used or recommended EEs for other applications beyond the criteria air pollutant program, including:

- Assessment of the magnitude of sea level rise associated with climate change (USEPA, 1995b; Titus and Narayanan, 1996).
- *Criteria for the Certification and Re-Certification of the Waste Isolation Pilot Plant's Compliance With the 40 CFR Part 191 Disposal Regulations* (*Federal Register*: February 9, 1996 [Volume 61, Number 28])—expert judgment can be used to elicit two types of information: (1) numerical values for parameters (variables) that are measurable only by experiments that cannot be conducted because of limitations of time, money, and physical situation; and (2) unknown information, such as which features should be incorporated into passive institutional controls that will deter human intrusion into the repository.
- Ecological model development—EPA's National Center for Environmental Assessment (NCEA) recently evaluated the utility of Bayesian belief networks (BBN) for modeling complex ecological phenomena. Because BBNs often have been cited as a promising method for modeling complex, uncertain phenomena, NCEA attempted to better understand the strengths and weaknesses of the approach. Beliefs were elicited from the panel of ecologists regarding the mechanisms by which sediment affects stream biota. Overall, the method showed promise despite the fact that a complete model was not developed within the allotted timeframe.

2.6 WHAT IS THE EXPERIENCE WITH EXPERT ELICITATION AT OTHER FEDERAL GOVERNMENT AGENCIES?

EE has been used or recommended for use by other agencies of the federal government across a broad range of applications including:

- U. S. Nuclear Regulatory Commission (USNRC)—established acceptable procedures for formal elicitation of expert judgment to support probabilistic risk assessments associated with the high-level radioactive waste program—*Branch Technical Position on the Use of Expert Elicitation in the High-Level Radioactive Waste Program* (NUREG-1563) (USNRC, 1996).
- U.S. Army Corps of Engineers—uses expert opinion to support risk studies on the assessment of unsatisfactory performance probabilities and consequences of engineered systems—*A Practical Guide on Conducting Expert-Opinion Elicitation of*

Probabilities and Consequences for Corps Facilities, IWR Report 01-R-01 (Ayuub, 2001).

- National Aeronautics and Space Administration—considers EE as an additional formal qualitative tool in evaluating upgrades to the Space Shuttle in terms of "cost, technological readiness, contribution to meeting program goals, risks, and ability to satisfy other NASA or federal government requirements."—*Upgrading the Space Shuttle* (NAS, 1999).
- U.S. Department of Transportation (USDOT)/Federal Railroad Administration—uses experienced domain experts as the basis for estimating human error probabilities—*Human Reliability Analysis in Support of Risk Assessment for Positive Train Control*, Chapter 2, Approach to Estimation of Human Reliability in Train Control System Studies, DOT/FRA/ORD-03/15 (USDOT, 2003).
- U.S. Department of Agriculture—used EEs to support the development of a risk-based inspection program under the Food Safety and Inspection Service—*Report to Congress*, March 2, 2001. (Hoffman et al., 2007).

2.7 WHAT IS THE EXPERIENCE WITH EXPERT ELICITATION OUTSIDE THE U.S. FEDERAL GOVERNMENT?

EE has been used in governmental organizations outside of the United States. The most notable of these is the Intergovernmental Panel on Climate Change (IPCC). The IPCC has used EE for many years to address specific components of the climate change issue (e.g., biomass, temperature gradient, thermohaline circulation, aerosol forcing) and generate overall estimates and predictions. Its reports characterize the science and generate estimates or projections that are used by various governmental and nongovernmental entities in support of climate-related decisions. The IPCC has developed formal procedures for EE in various aspects of the program (e.g., IPCC, 2001) and developed guidance on issues such as how to address qualitative expressions of uncertainty (IPCC, 2005).

Other international examples of EE include:

- Dutch National Institute for Public Health and the Environment (RIVM)—*Uncertainty Analysis of NO_x Emissions from Dutch Passenger Cars in 1998: Applying a Structured Expert Elicitation and Distinguishing Different Types of Uncertainty* (RIVM, 2003).
- European Commission, Nuclear Science and Technology—*Procedures Guide for Structured Expert Judgment* (2000).

2.8 WHAT EXAMPLES OF EXPERT ELICITATION ARE RELEVANT TO EPA?

Academia and industry have conducted numerous EEs in scientific areas that are relevant to EPA's work, including:

- **Environmental Transport—**Amaral et al. (1983) and Morgan et al. (1984) used expert judgment in the evaluation of the atmospheric transport and health impacts of sulfur air pollution.
- **Dose-response—**Hawkins and Graham (1988) and Evans et al. (1994) used EE for formaldehyde and Evans et al. (1994) for risk of exposure to chloroform in drinking water. Other researchers used EE to estimate dose-response relationships for microbial hazards (Martin et al., 1995).
- **Exposure—**Hawkins and Evans (1989) asked industrial hygienists to predict toluene exposures to workers involved in a batch chemical process. In a more recent use of EE in exposure analysis, Walker et al. (2001, 2003) asked experts to estimate ambient, indoor, and personal air concentrations of benzene.

2.9 SUMMARY

In the 1950s, EE emerged from the growing field of decision theory as a technique for quantifying uncertainty and estimating unobtainable values to support decision-making. A wide range of activities may fall under the term "expert judgment," but this White Paper restricts the term EE to a formal systematic process to obtain quantitative judgments on scientific questions (to the exclusion of personal or social values and preferences). This process includes steps to minimize the effects of heuristics or other biases in expert judgments. Since the late 1970s, EPA has performed and interpreted EEs as part of its regulatory analyses for the air program. EEs have been conducted and used by at least five other federal agencies and international organizations. In recent years, interest in using EE has increased because of encouragement from the NAS and OMB.

3.0 WHAT IS EXPERT ELICITATION?

This chapter provides a brief review of EE research and defines terms that will be used throughout this White Paper. This chapter includes discussions about the origins and foundations of EE, the general reasons why EE is conducted or should be conducted (Chapter 4 provides a detailed EPA-centric discussion of this topic), the components of an EE, and some cautions and criticisms of EE.

3.1 WHAT IS EXPERT ELICITATION?

EE is a multidisciplinary process that can inform decisions by characterizing uncertainty and filling data gaps where traditional scientific research is not feasible or data are not yet available. Although there are informal and nonprobabilistic EE methods for obtaining expert judgment, for the purposes of this White Paper, EE is defined as a formal systematic process to obtain quantitative judgments on scientific questions, such as an uncertain quantity, or the probability of different events, relationships, or parameters (Southern Research Institute [SRI], 1978; Morgan and Henrion, 1990).

Classical (frequentist) statistical techniques are based on observed data without incorporating explicitly subjective judgment. In many situations, however, especially in environmental statistics, complete or adequate data for statistical analysis do not exist and judgment must be used to analyze existing data. Probabilistic statements of belief that can inform (or add utility to) decision-making are accommodated within a Bayesian framework, even though they generally are excluded under a frequentist approach.

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. The EE process uses expert knowledge, synthesized with experiences and judgments, to produce probabilities about their confidence in that knowledge. Experts derive judgments from the available body of evidence, including a wide range of data and information ranging from direct empirical evidence to theoretical insights. Even if direct empirical data were available on the item of interest, such measurements would not necessarily capture the full range of uncertainty. EE allows experts to use their scientific judgment transparently to interpret available empirical data and theory. It also should be noted that the results of an EE are not limited to the quantitative estimates. These results also include the rationale of the experts regarding the available evidence that they used to support their judgments and how they weighed different pieces of evidence.

3.2 WHY IS EXPERT ELICITATION NECESSARY?

EPA and other federal regulatory agencies often need to make decisions in situations with significant data gaps and that lack scientific consensus. When relevant data are unavailable to characterize the uncertainty of the problem at hand, decision-makers often rely on expert judgment through informal or formal processes. The discipline of decision analysis has developed to assist decision-makers who must make decisions in the face of this uncertainty. Quantitative uncertainty analysis can be useful when important decisions depend on uncertain assumptions, estimates, or model choices. EE provides a formal process to obtain this expert judgment. The various reasons that EE might be used rather than or to compliment other methods of addressing uncertainty are discussed in Chapter 4.

Among federal government agencies, the USNRC has the longest and most extensive experience in conducting EE. The USNRC's Branch Technical Position (NUREG-1563) states that EE should be considered if any of the following conditions exist:

- Empirical data are not reasonably obtainable or the analyses are not practical to perform.
- Uncertainties are large and significant.
- More than one conceptual model can explain, and be consistent with, the available data.
- Technical judgments are required to assess whether bounding assumptions or calculations are appropriately conservative (USNRC, 1996).

These and similar situations have motivated many EEs during the past several decades. Additional reasons for conducting an EE are to:

- Obtain prior distributions for Bayesian statistical models and help interpret observed data.
- Provide quantitative bounds on subjective judgments. Interpretations of qualitative terms (e.g., "likely" and "rare") vary widely. EE can provide numbers with real uncertainty bounds that are more useful for subsequent analyses.
- Promote consensus among experts regarding a contentious decision (heretofore *a rare application*) (Cooke and Goossens, 2000).
- Provide input on the prioritization of potential research options or potential decision options.

EE can produce distributions to characterize uncertainty about parameters that lack empirical data or cannot actually be measured in practice. This is particularly helpful when better scientific information is too costly or impractical to obtain, will not be available within the timeframe of the decision, or is unobservable (van der Fels-Klerx et al., 2002; Merkhofer and Keeney, 1987). For example, the goal of a recent study, which involved the elicitation of several continuous variables related to animal health safety (van der Fels-Klerx et al., 2002), was to obtain one aggregated probability density function for each continuous variable based on the combined (weighted) distributions obtained from the collection of individual experts. In addition, many of the analyses to assess the accident probabilities for nuclear power plants and radiation leakage from nuclear waste disposal options have relied on EE to characterize distributions for parameters or events that lack empirically frequency data (USNRC, 1996; Merkhofer and Keeney, 1987). In many cases, even quantities that are measurable in principle often are difficult to measure in practice because of the time constraints or limitations in current technology to conduct such measurements. Sometimes the needed data are derived from studies that require many years to complete (e.g., a cancer study in a target population for which observations require 20 years or more). In these cases, EE may be more expedient. In cases in which direct observations are impossible (e.g., safety of nuclear facilities or the risks of terrorist attacks), EE may provide the only available information to address a particular question (O'Hagan, 2005).

3.3 WHAT DO EXPERT ELICITATION RESULTS REPRESENT?

Most scientists are comfortable with empirical observations, view such results as objective, and understand what statistical analyses represent. However, EE results and their dependence on subjective judgment are unfamiliar to most people, including many scientists. They lack an understanding of how subjective judgment can be combined with empirical data to obtain a different type of information—one that focuses on the likelihood of the nature of an unknown quantity, event, or relationship. A useful comparison of objective and subjective probabilities is (NAS, 1994):

> ...Objective probabilities might seem inherently more accurate than subjective probabilities, but this is not always true. Formal methods (Bayesian statistics) exist to incorporate objective information into a subjective probability distribution that reflects other matters that might be relevant but difficult to quantify, such as knowledge about chemical structure, expectations of the effects of concurrent exposure (synergy), or the scope of plausible variations in exposure. The chief advantage of an objective probability distribution is, of course, its objectivity; right or wrong, it is less likely to be susceptible to major and perhaps undetectable bias on the part of the analyst; this has palpable benefits in defending a risk assessment and the decisions that follow. A second advantage is that objective probability distributions are usually far easier to determine.

> However, there can be no rule that objective probability estimates are always preferred to subjective estimates, or vice versa...

Subjectivity is inherent to scientific methodologies, collection and interpretation of data, and developing conclusions. For example, in traditional scientific research, the choice of theories, models, and methods may influence data and conclusions. EE is no different in this respect. Because EE findings contain knowledge from data combined with probability judgments about that knowledge, however, the subjectivity is more transparent and therefore accessible to critique. The remainder of Section 3.3 describes in more detail what the results of an EE represent. Where appropriate, this section also highlights some areas in which practitioners should pay particular attention, so that EE results are described accurately and represented fairly.

3.3.1 Are Expert Elicitation Results Arbitrary?

Because EE results are based on subjective judgment, there is a concern that they may be considered arbitrary or biased. However, EE results are based on the experts' knowledge and understanding of the underlying science. To obtain EE results, experts are asked to extrapolate from their synthesis of the empirical and theoretical literature using judgments that conform to the axioms of probability. As stated above, the EE results include both quantitative estimates and the underlying thought process or rationale. By reviewing the qualitative discussion that summarizes an expert's reasoning, one can assess whether the expert's rationale is reasonable and consistent with available evidence and theory.

3.3.2 Do Expert Elicitation Results Represent New Data or Knowledge?

Science can be thought of as two ideas: (1) a description of our state of knowledge of the world—what we know and do not know (epistemic evaluation of knowledge and uncertainty) and (2) the process by which we obtain better information (primarily to reduce uncertainty). The former improves our understanding of existing observations; the latter involves the creation of new data or knowledge. Consistent with NAS (2002) recommendations, this White Paper asserts that EE results encompass only the first aspect of science (characterization of existing knowledge) because no new experimentation, measurement, or observations are conducted. Furthermore, the purpose of EE is to characterize or quantify uncertainty and not to remove uncertainty, which can only be done through investigative research. An improved characterization of the knowledge, however, can better inform a decision.

This distinction is particularly important because a common reason for conducting an EE is to compensate for inadequate empirical data (Keeney and von Winterfeldt, 1991; Meyer and Booker, 2001; O'Hagan, 2005). In contrast, it has been suggested that EE judgments themselves

be treated similar to data (Meyer and Booker, 2001). Although the results of EE can be used in ways similar to data (e.g., model inputs), however, one should ensure that the distinction between experimental data and EE results is maintained, and the pedigree of data is clear. Users of EE results are cautioned to understand the differences between EE results and experimental data and to be aware of the role of expert judgments in EE results. For these reasons, NAS recommended that EPA identify clearly which analyses are based on experimental data and which are based on expert judgment. This distinction should be maintained and communicated to the decision-maker (NAS, 2002).

EE reflects a snapshot of the experts' knowledge at the time of their responses to the technical question. Because of this, users of EE should expect that the experts' judgments will change as the experts receive new information and are exposed to new interpretations of existing data. An alternative approach to EE is to use experts to develop principles or rules that generalize the data so that very sparse data can be used in a broader manner (i.e., provide additional certainty for sparse data). In one study, decision rules for making health hazard identifications were elicited from national experts (Jelovsek et al., 1990). The authors concluded that: (1) many experts must be consulted before determining the rules of thumb for evaluating hazards, (2) much human effort is needed in evaluating the certainty of the scientific evidence before combining the information for problem solving, (3) it still is unknown how experts view uncertainty in their areas of expertise, and (4) the knowledge elicited from experts is limited but workable for medical decision making.

3.3.3 Are Expert Elicitation Results Equivalent to a Random Statistical Sample?

EE results should not be treated as a random statistical sample of the population being studied. In contrast to a "scientifically valid survey" that randomly samples the study population to obtain a representative sample, an EE seeks to reflect the range of credible scientific judgments. If experts are selected from multiple legitimate perspectives and relevant expertise, EE can indicate the range of plausible opinions, regardless of the prevalence of those opinions in the population. Consequently, the selection of experts is critical to the success of an EE.

3.4 WHAT ARE SOME APPROACHES FOR EXPERT ELICITATION?

This section describes the advantages and disadvantages of the two general approaches to the EE process: individual and group techniques. Chapter 5 presents how expert judgments are formally captured and lays out the process and specific steps needed to control rigorously for potential biases that may arise during elicitations.

3.4.1 Individual Elicitation

EPA's early EE efforts have primarily utilized individual elicitation techniques. This is the EE approach recommended by the NRC (1996). In it, individual experts are elicited separately using a standardized protocol. Often, the intent of these individual elicitations is to characterize uncertainty rather than defining a "best" estimate or consensus position. One advantage of individual elicitations is that the broadest range of factors contributing to overall uncertainty can be identified explicitly by each expert. In an EE involving elicitation of individuals, one can assess which parameters (quantitative or qualitative [e.g., model choice]) have the greatest impact on uncertainty. Furthermore, using individuals' elicitations eliminate the potential biases that arise from group dynamics.

Although relying on a collection of individual elicitations provides the most robust picture of uncertainty, it does not necessarily promote harmony or represent consensus. By encouraging a diverse spectrum of responses, some may think individual elicitations obfuscate rather than illuminate decisions. There are decision analytical techniques to evaluate the impact of this diversity on decisions, however (Clemen, 1996). Chapter 5 presents more detail on how individual elicitations are conducted.

3.4.2 Group Elicitation

A second EE approach is a group process in which experts evaluate data interactively and determine their collective judgment (Ehrmann and Stinson, 1999). By sharing data and judgments, group interactions can identify a "best" estimate or consensus opinion given the current state of knowledge. Group processes typically generate data interpretations that are different from those obtained by individual experts. These group processes include the Delphi method, nominal group techniques, group nomination, team building, and decision conferencing.[11]

Although group processes have the advantage that they can often obtain consensus, they are potentially limited by the influence of group dynamics (e.g., strong and controlling personalities and other forms of circumstantial coercion). Therefore, if group techniques are used, the effect of group dynamics needs to be considered in technique selection and application and in interpreting the results, in addition to considering and controlling general heuristic biases (Section 3.5.5). In addition, group processes that promote consensus may not characterize the full range or extent of the uncertainties. Chapter 5 includes a more detailed discussion about conducting group elicitations.

[11] Using group processes to obtain a collective judgment may trigger the Federal Advisory Committee Act and its requirements.

3.4.3 Combining Individual Experts' Judgments

EE results from multiple experts often produce insights without combining the experts' judgments. Because EE results are often used as model inputs or information for decision makers, however, there are many circumstances in which it is desirable to aggregate or combine multiple expert judgments into a single metric. There are a number of methodologies that aggregate individually elicited expert judgments, which is different than obtaining collective judgments via a group process (Section 3.4.2). Section 5.4.4 provides a discussion on the advantages and cautions of combining expert judgments. Section 3.4.3.1 presents methodologies for aggregation of individual expert judgments to produce a combined result. Section 3.4.3.2 discusses the aggregation of expert judgments by consensus processes.

3.4.3.1 Mathematical and Behavioral Approaches for Combining Individual Judgments

A number of approaches have been proposed and used to combine individual expert judgments. Mathematical aggregation methods involve processes or analytical models that operate on the individual probability distributions to obtain a single combined probability distribution. Mathematical approaches range from simple averaging using equal weights (Keeney and von Winterfeldt, 1991) to a variety of more complex Bayesian aggregation models. Although the Bayesian aggregation methods are theoretically appealing, difficult issues remain concerning how to characterize the degree of dependence among the experts and how to determine the quality of the expert judgments (e.g., how to adjust for such factors as overconfidence). Clemen and Winkler (1999) reviewed both mathematical and behavioral approaches for combining individual judgments along with empirical evidence on the performance of these methods. Using mathematical methods to combine expert opinions relies on an assumption that the individual expert opinions are independent (O'Hagan, 1998). Behavioral aggregation approaches "attempt to generate agreement among the experts by having them interact in some way" (Clemen and Winkler, 1999). Chapter 5 provides additional discussion of the use of group processes for EE.

Based on their review of the empirical evidence evaluating both mathematical and behavioral aggregation methods, Clemen and Winkler (1999) found both approaches tended to be similar in performance and that "simple combination rules (e.g., simple averaging) tend to perform quite well." They also indicated the need for further work in the development and evaluation of combination methods and suggest that the best approaches might involve aspects of both the mathematical and behavioral methods. In the meantime, they express the view that simple mathematical averaging will always play an important role given its ease of use, robust performance, and defensibility in public policy settings in which it may be difficult to make distinctions about the respective quality of different expert judgments.

Cooke (1991) recognized that all individuals do not possess equal skill in generating or thinking in probabilistic terms. Given similar technical knowledge, some experts are more adept at providing higher quality estimates (see Section 3.5.4 about what makes a good judgment). Therefore, Cooke advocated assessing an individuals' ability to provide "statistically" robust probability estimates. To assess individual probabilistic abilities, he used seed questions that are similar in nature to the EE's questions of interest but for which answers are known. Their performance was characterized by their statistical calibration (i.e., their ability to capture the correct proportion of answers within stated bounds) and their informativeness (i.e., the degree to which probability mass is distributed relative to background, the narrowness of the bounds). An expert's performance was gauged relative to other experts in the EE exercise and weighted accordingly. Cooke showed that such weighted combinations are superior to equal weighting and citation-based weighting. The success of Cooke's approach, however, hinged on the quality of the seed questions in terms of their clarity (i.e., ability of the experts to correctly understand and respond to them) and their relevance to the specific problem area and question of interest. To overcome these obstacles, significant time and effort may be needed to develop, review, and evaluate these seed questions.

3.4.3.2 Consensus Processes for Combining Individual Judgments

Individual judgments also can be combined through a consensus process. This approach differs from group elicitation (Section 3.4.2) in which the entire elicitation was conducted as a group. Here, the experts are elicited individually, and then as a second step, their judgments are combined via a group process. In this iterative approach, experts are allowed to discuss their original opinions and arrive together at a collective opinion (i.e., group EE) (Meyer and Booker, 2001; Gokhale, 2001).

Under this approach, the aggregation of individual expert judgments requires the experts to adjust their judgments and move toward consensus. By defining the quantitative issues of interest and removing ambiguous judgments, this process can help experts to refine their understanding of the problem and potentially narrow their differences. Thus, when used interactively, EE can aid in moving experts toward greater consensus on science-relevant problems that cannot be directly measured (Cooke and Goossens, 2000; Meyer and Booker, 2001). This approach to EE has particular utility when the goal of the assessment is to obtain consensus views.

3.4.4 Problems Combining Expert Judgments

In individual elicitation, each expert supplies judgments on the same set of questions. Any decision to combine multiple judgments is left to the analyst and decision-maker. A typical

EE obtains judgments from at least three experts because diversity is more likely to reflect all relevant knowledge. In addition to their knowledge, however, each expert brings different experiences, perspectives, and biases to the question(s) of interest. Combining expert judgments may present several pitfalls, including the potential for misrepresenting expert judgments, drawing misleading conclusions about the scientific information, and adding biases to the conclusions (Hora, 2004; Keith, 1996; O'Hagan, 2005). Therefore, EE practitioners must be cautious about aggregating expert judgments and presenting combined conclusions about EE results. As discussed above, although individual expert judgments can be combined to provide an aggregated single average value or can be aggregated through subsequent discussion and consensus, the chosen approach should be part of the EE study methodology and agreed-on procedures.

According to Keith (1996), combining judgments could be problematic because an underlying methodological assumption is that the experts chosen for the elicitation represent the entire continuum of "truth" with respect to the technical question. He cautions that the "fraction of experts who hold a given view is not proportional to the probability of that view being correct" (Keith, 1996). In part, this is caused by how the experts are selected and the extent to which any of the selected experts possess knowledge that approximates the "truth." As mentioned in Section 3.1, expert opinions are not necessarily evenly distributed across the entire spectrum of potential opinions. Furthermore, prior to interviewing experts, it may not be possible to determine the range of expert opinion on a particular question. Consequently, depending on which experts are selected (and agree) to participate in the EE, the fraction of experts used for the elicitation cannot be assumed to be proportional to the probability of that view or opinion being correct. In addition, if all else is equal and because a true value cannot be known, there is no objective basis to value the opinion of any one expert over any other.

Despite these cautions, combining judgments may nevertheless be desirable (e.g., to provide a single input for a model or to reduce complexity for a decision-maker). In such cases, resolving differing expert views can be done by combining individual judgments with a mathematical method or via consensus building. Combining expert judgments requires the relative weighting of individual expert judgments to each other. They may be weighted equally or in some differential manner—for example, by social persuasion (as might occur in Delphi consensus building methods), by expert credentials, or by some form of calibration or performance assessment (Cooke, 1991). Keith (1996) argues that equal weighting of expert judgments is generally inappropriate because it is not possible to obtain a sufficiently large sample of in-depth EEs to ensure that all possible expert views are represented. Others argue

that equal weighting is often as effective as more sophisticated differential weighting approaches (Clemen and Winkler, 1999).

It is also possible to combine expert judgments via consensus building. Unlike the combination of individual expert judgments, which can be performed without the participation of the experts themselves, consensus building is a component of a group EE process. Group elicitation provides a forum for experts to interact and exchange information, with the intention of promoting a shared set of prior knowledge that is used by all experts in making their judgments and which they collectively update through the deliberative process. The combination of judgments often is accomplished implicitly by eliciting a single judgment from the entire group. Through group elicitation, experts often can produce a common definition of the problem and a common judgment. Allowing experts to interact can help mitigate the problem of expert selection in a particular elicitation by providing an opportunity for a wider range of opinions to be articulated and explored among the expert group than they may have expressed individually on their own. A disadvantage of this type of group elicitation is that the social dynamics and interaction may lead to an overly narrow uncertainty characterization, especially if the retention of minority views that express a broader range of uncertainty is displaced by the goal of reaching a consensus judgment. It is therefore important that minority opinions and their rationale also be preserved and presented to decision makers. Another consideration is that using group processes to obtain a collective judgment may trigger the Federal Advisory Committee Act (FACA) and its requirements.

Other EE practitioners also urge caution about combining the individual judgments (Wallsten et al., 1997). In a methodology similar to Jevolsek et al. (1990), Wallsten et al. (1997) proposed a model that considers both "the structure of the information base supporting the estimates and the cognitive processes of the judges who are providing them." Wallsten et al. determined areas in which experts agree and derived rules that satisfy those points of agreement. The resulting model avoids some of the criticisms of combining expert judgments by specifically considering subjective inputs for data and the processes used by the experts in the elicitation.

Other formal methods have also been devised that combine individual and group elicitation (e.g., Delphi). In all of these cases, one can expect that the final elicited judgments will vary with the methods that are selected. Therefore, care should be exercised to use elicitation methods that are most appropriate for a particular problem.

3.5 WHAT ARE GOOD PRACTICES FOR ELICITING EXPERT JUDGMENT?

Based on a literature review and actual experience within EPA and other federal agencies, this White Paper now will concentrate on good practices for eliciting expert judgment.

Many of the issues discussed in this section are relevant to EE as well as other quantitative methods that express judgment through probability.

3.5.1 Why is a Probabilistic Approach Often Used to Elicit Judgments?

In support of decision-making, EEs provide insights into parameter values, quantities, events, or relationships and their associated uncertainty. A common mathematical language is sometimes helpful in providing a rigorous, credible, and transparent assessment to support decisions. For EE assessments, the common language that is most effective at ensuring usability of results and comparability across experts is probability. Although subjective terminology (e.g., "likely" or "unlikely") can convey probabilities, numerous studies have shown that a natural language approach often is inadequate because:

- The same words can mean very different things to different people.
- The same words can mean very different things to the same person in different contexts.
- Important differences in expert judgments about mechanisms (functional relationships) and about how well key coefficients are known can be easily masked in qualitative discussions.

Wallsten (1986) documented that individual interpretations of words can differ dramatically if they are presented without context. In this study, they evaluated 10 qualitative descriptions of likelihood (almost certain, probable, likely, good chance, tossup, unlikely, improbable, doubtful, and almost impossible). For each description, the participants expressed an associated probability. The range varied considerably between participants, including overlap across words such that some were indistinguishable.

Similarly, Morgan (1998) presented the results of an exercise in which he queried the members of the SAB Executive Committee at a time when EPA was considering moving toward a more qualitative description of cancer hazard. He asked the committee members about their interpretations of the terms "likely" and "not likely." This exercise found that the minimum probability associated with the word "likely" spanned four orders of magnitude, the maximum probability associated with the word "not likely" spanned more than five orders of magnitude, and most importantly, there was an overlap of the probability associated with the word "likely" and that associated with the word "unlikely." Because interpretations of qualitative descriptions have such high inter-individual variability, a quantitative framework is helpful for experts to provide comparable and tractable expressions of belief. Probability can provide this framework; and, in particular, a subjectivist approach to probability is ideally suited for this application. One

form of subjective probability is a formal expression of the degree of belief about some unknown quantity, which is therefore ideal for quantifying uncertainty in expert beliefs.

3.5.2 What are Problems With Probability? Are There Alternatives?

Shackle (1972a) states that "probability is…a word with two quite opposed uses. It is at the same time the name of a kind of knowledge, and an admission of a lack of knowledge." When subjective probabilities sum to one, it implies omniscience and that all alternative hypotheses have been considered. It is often the case, however, that statisticians using subjective probability do not know all the hypotheses and cannot set up a statistical problem that satisfies this major premise (Shackle, 1972b). Consequently, Shackle asserts that when statisticians use statistics to draw conclusions about data, they need to be mindful that the statistics may be derived from an incomplete set of potential hypotheses.

Although a useful tool, the field of statistics involves simplification of real world problems, and subjectivity is intrinsic in its methodologies. Pretending that statistical approaches are objective may result in misplaced confidence in data and conclusions. In EE, the expression of expert judgment as probabilities assumes that experts understand all alternatives so that their judgments can be compared. This may be true for a binary outcome in which the expert is asked for the probability of occurrence (or nonoccurrence). In most situations, however, the expert is asked to make judgments about the probability of one event occurring compared with the occurrence of multiple other events, some of which may be unknown. This challenge is further complicated by "self -reinforcing" or inherently evolutionary systems (Shackle, 1972b). Evolutionary systems have elements of unpredictability (some events are completely unrelated to previous events) that make it unacceptable for using probability to describe the system because probability contains an inherent assumption of stability within a given system.

Although expert judgment commonly is expressed solely in probabilistic terms, there are other feasible approaches. Meyer and Booker (2001) define expert judgment as "data given by an expert in response to a technical problem." Using such a definition, it is possible to obtain expert judgments in a variety of nonprobabilistic forms. Expert judgment often is used where data cannot be collected practically or are too expensive to assemble. Quantitative but nonprobabilistic methods for expressing expert judgment have been commonly used in decision analysis. Such approaches tend to use pair-wise comparisons and stated preferences among the pairs. Doing so does not require the expert to formally give probability estimates for an event, parameter, or relationship. This method has been particularly useful for supporting decisions in which values and preferences (not just scientific evidence) are considered. It, however, also can be used to elicit expert judgment about a specific technical problem or scientific interpretation as

well. Analysts who are considering the use of EE but are reluctant because of concerns about probabilistic approaches may find these alternative methods for expert judgment more suitable.

3.5.3 What Makes a Good Expert?

The intent of an EE is to characterize the state of knowledge by integrating available evidence with scientific judgment to provide as complete a picture as possible of the relevant knowledge regarding a particular question. Elicitation of expert judgment provides a useful vehicle for combining empirical data and observations, as reflected in the published literature, with expert judgment. Hence, there are two criteria for defining experts as "good" (in terms of their competence and skill, though not necessarily their ethical conduct). The first is an understanding of the body of literature for the problem of interest. Technical knowledge alone, however, does not constitute a good expert. Experience and judgment, including intuition and the ability to integrate information and theory, also are critical.

3.5.4 What Constitutes Good Expert Judgment?

A well-conducted EE should reflect accurately the selected experts' judgments and capture the "truth" within the range of expert judgments. EE goes beyond empirical observation, which in general cannot capture the true estimate of uncertainty. Therefore, good expert judgments should consider more than just the statistical confidence limits from empirical studies.

A good judgment properly captures the range of uncertainty, but it still should be reasonable. Some individuals are more capable of formulating and expressing their judgments probabilistically than others. The skills for good probabilistic judgment can be demonstrated by eliciting known, or soon to be known, quantities. Cooke (1991) identified two characteristics of good probability judgment:

- ***Being calibrated or statistically accurate***—a good probability judgment is one that mimics the underlying probability of predicting the "truth" if it were known. In other words, the credible intervals presented by the experts should capture the "true" value within the expressed credible intervals (i.e., 90% confidence intervals should include 90% of the true values). Furthermore, the estimates should be balanced. For example, 50 percent of any "true" values should be above, and 50 percent should be below an expert's estimated median values.
- ***Informativeness***—a good probability judgment is one in which the probability mass is concentrated in a small region (preferably near the true value) relative to the background rate.

As illustrated in Figures 3-1 and 3-2, ideally one prefers experts whose judgments are unbiased[12] and precise. Building on that premise when it comes to expert judgments, one would like to be as unbiased as possible where the central mass is closest to the true value. In addition, the credible limits should be sufficiently broad so as to include the proper proportion of "true" values.

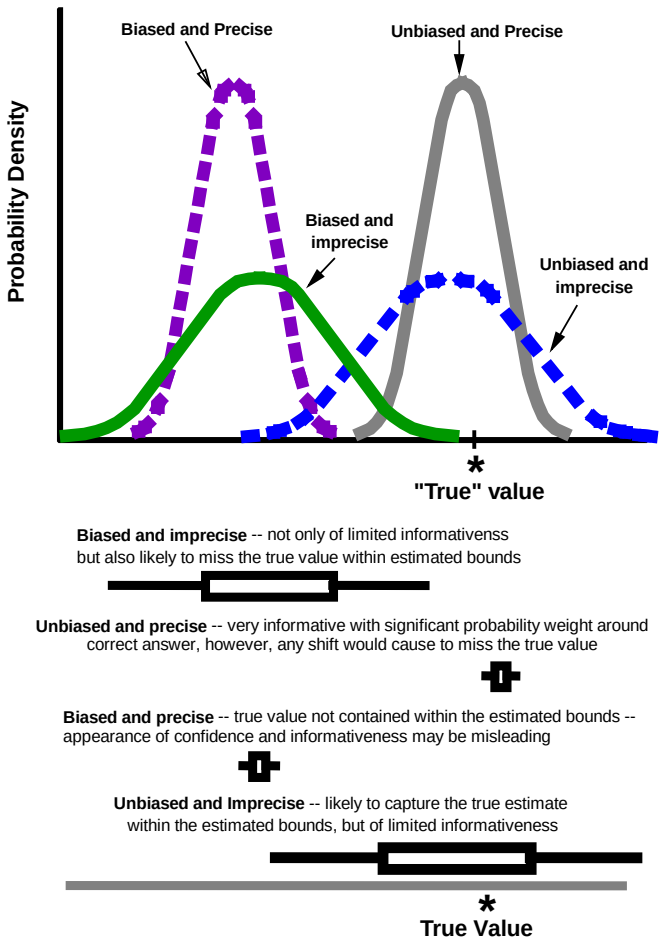


Figure 3-1. Distributions To Illustrate Bias and Precision

A good expert, however, should not have bounds that are too broad so as to reduce the mass or confidence around the true value. In addition to expressing probabilities that are statistically robust, it also is important that experts describe clearly the information they used to support their opinions. Experts who can express the basis for their judgments are strongly preferred.

[12] The term "bias" is used here in its statistical sense and does not refer to ethical issues (e.g., loss of impartiality under 5 CFR 2635 Subpart E) that may occur when there is a nexus between a person's private interests and public duties.

3.5.5 In Which Situations Can the Elicitation Process Go Awry? Why?

Most lay people, and even experts, do not have well-formed probability distributions for quantities of interest *a priori*. For EE, therefore, a process is necessary to conceptualize and develop these probability values. During this process, experts use existing experience and knowledge to infer the probabilities that are elicited. To accomplish this task, most people (including subject matter experts) make use of simple rules of thumb called "cognitive heuristics." In some instances, these heuristics can lead to biases in judgment. The psychometric literature (e.g., Kahneman et al., 1982) describes the heuristic biases that most impact EE judgments:

- Overconfidence.
- Availability.
- Anchoring and adjustment.
- Representativeness bias.
- Motivational bias.

3.5.5.1 Overconfidence

One consistent finding across all elicitation techniques is a strong tendency toward overconfidence (Morgan and Henrion, 1990). Figure 3-2 provides a histogram summary of the results of 21 different studies on questions with known answers and the observed surprise indices from a wide range of studies of continuous distributions. The surprise index is the percentage of time the actual value would fall out of the elicited credible range (e.g., 98% credible interval). Ideally, the actual value would occur outside of the elicited credible range 2 percent of the time. As Figure 3-2 illustrates, however, the surprise index is almost always far too large, ranging from 5 to 55 percent instead of 2 percent.

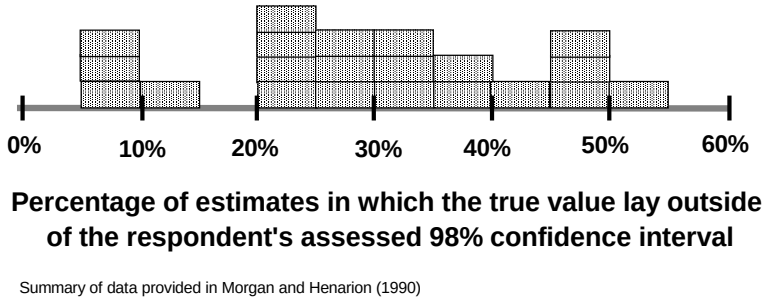


Figure 3-2. Summary of "Surprise Index" From 21 Studies With Known Answers (Morgan and Henrion, 1990)

Given this tendency toward overconfidence, there has been significant interest in whether training potential experts, using trial tasks such as encyclopedia questions, can improve

performance (Morgan and Henrion, 1990). Table 3-1 summarizes the impacts of several training experiments and their impact on overconfidence. Some studies attempted to reduce the overconfidence by explaining prior performance and exhorting the experts to increase the spread of their estimates. Results showed modest decreases in overconfidence. Other studies that provided comprehensive feedback showed significant improvement on discrete elicitations; however, improvement was marginal for continuous distributions. The nature of the feedback is critical. In particular, it may be important to include personal discussion in feedback.

Table 3-1. Impact of Training Experiments on Reducing Overconfidence

	Number of assessed distributions N	Interquartile index (ideal = 50)		Surprise index (ideal = 2)	
		before	after	before	after
Alpert & Raiffa (1969)					
Groups 2 & 3	1,670	33	44	39	23
Group 4	600	36	43	21	9
Schaefer & Borcherding (1973)					
Fractiles	396	23	38	39	12
HFS	396	16	48	50	6
Pickardt & Wallace (1974)					
Group 1 (5 sessions)	?	39	49	32	20
Group 2 (6 sessions)	?	30	45	46	24
Schaefer (1976)					
(5 sessions)	660	27	34	25	14
Lichtenstein & Fischhoff (1980)					
(Training on discrete tasks)	924	33	37	41	40

3.5.5.2 Availability

Probability judgment also is driven by the ease with which people can think of previous occurrences of the event or can imagine such occurrences. In Figure 3-3, the results of elicited judgments about annual deaths rates from various causes are plotted against actual death rates. The judgments tend to overestimate the occurrence of rare events and underestimate the rate of more common causes of death. The explanation may be that rare events receive relatively more publicity and are more readily recalled. Therefore, they are believed to occur more frequently than they do. By comparison, deaths from common causes receive minimal publicity, and, therefore, elicited judgments may underestimate their true rates.

3.5.5.3 Anchoring and Adjustment

Anchoring occurs when experts are asked a series of related probability questions. Then, when forming probabilities, they may respond by adjusting values from previous questions. For example, if an expert is first asked to state the median probability of an event, this stated probability for the median may become an “anchor.” Then, when responding to subsequent

questions about other quantiles (e.g., 10th percentile) of the distribution, the responses may be influenced by the anchor.

Probability judgment frequently is driven by the starting point, which becomes an "anchor." For the example shown in Figure 3-4, people were asked to estimate annual death rates from different causes. They were provided with either the true annual deaths rate for autos (50,000 per year) or the true annual deaths for electrocutions (1,000 per year). The given death rate provided an anchor and influenced the results. As can be seen from the graph, estimated death rates were shifted by which reference value was given. This given value becomes an anchor and subsequent estimates are made relative to it by adjustment.

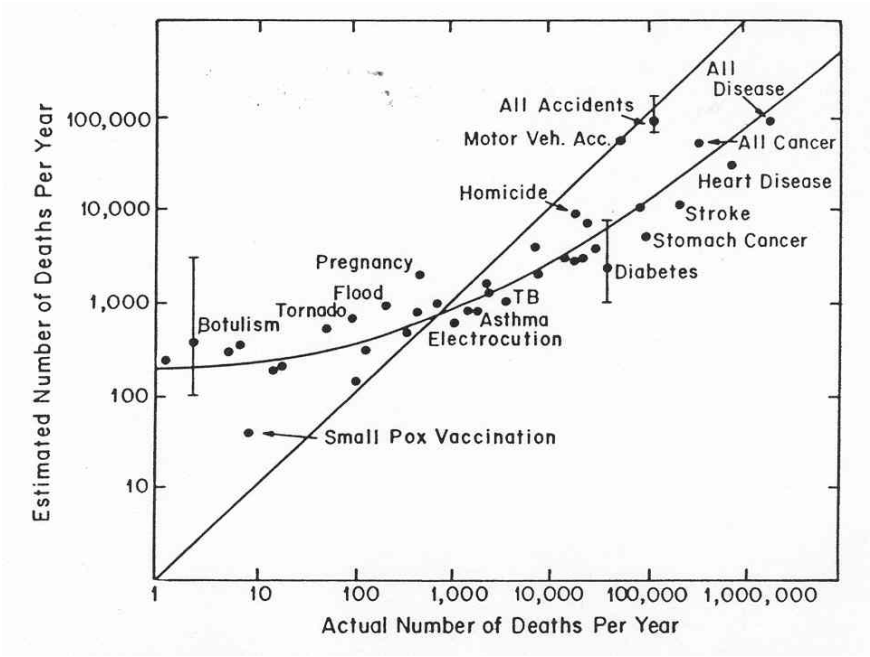


Figure 3-3. Availability Bias (Lichtenstein et al., 1978)

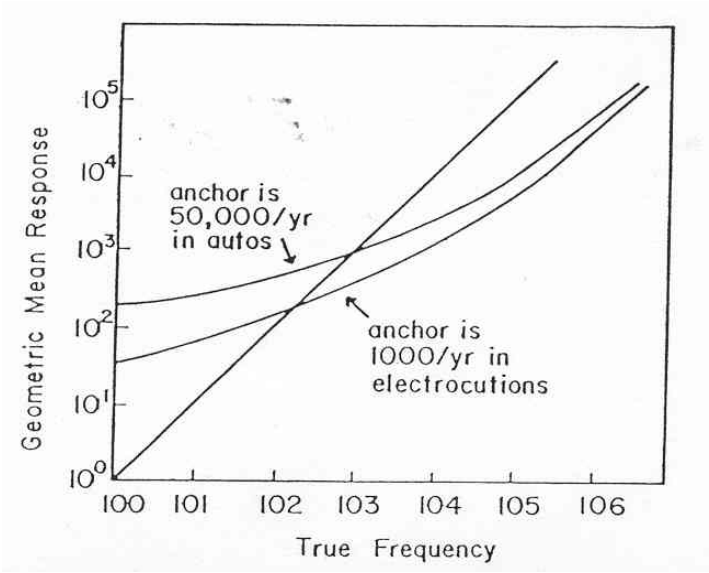


Figure 3-4. Anchoring for Estimating of Death Rates (Lichtenstein et al., 1978)

In an observation of flawed probabilistic assessments, Bruine de Bruin et al. (2006, 2007) found that respondents to some probability questions tend to have an "elevated frequency" of 50 percent responses. This disproportionately high number of 50 percent responses does not reflect true probabilistic beliefs. Rather, it is "caused by intrusion of the phrase '50-50,' which represents epistemic uncertainty, rather than a true numeric probability of 50 percent" These nonnumeric 50 percents may be an artifact of the question format and the elicitation methodology that are compounded by the elicitee's understanding of probability. Because treating these nonnumeric 50 percents as true 50 percents could lead to erroneous conclusions, Bruine de Bruin et al. (2000, 2002) presents two redistribution techniques to mitigate this difficulty.

3.5.5.4 Representativeness Bias

People also judge the likelihood that an object belongs to a particular class based on how much it resembles that class. This phenomenon can be illustrated considering the following example. Suppose one flips a fair coin 10 times. Which of the following two outcomes is more likely?

Outcome 1: T, T, T, T, T, H, H, H, H, H
Outcome 2: T, T, H, T, H, T, T, T, H, H

Both sequences are equally likely, but the second may appear more likely because it seems to better represent the underlying random process. By contrast, the first sequence gives the

appearance of a nonrandom pattern. In general, people tend to underestimate the occurrence of patterns or sequences that appear to be nonrandom.

3.5.5.5 Motivational Bias

Frequently, experts may have direct or indirect interests in the outcome to the question at hand. Hence, whether consciously or not, their judgments may be influenced by motivational bias. In some cases, the stakes may be clear (e.g., when the outcome of a question may impact employment or investments). For other cases, motivational bias may be subtler. For example, the professional reputation of a particular expert may be associated with a particular point of view or theory, making it difficult to express an alternative perspective (Morgan and Henrion, 1990).

The regulatory process is characterized by complex multifactor problems in which decisions are influenced by technical analyses as well as social, economic, and political considerations. Furthermore, this process involves multiple stakeholders, each with their own positions, frames, and agendas. As a result, motivational biases are among the more elusive and yet critical biases to consider for EEs that support regulatory decisions. The existence of motivational bias may be difficult to demonstrate, but the adversarial nature of regulatory decisions suggests that motivational biases exist. In part, this is caused by the fact that people tend to trust that their judgments are less prone to bias than those of others. One explanation of this misperception may be that people tend to rely on introspection for evidence of bias in themselves but rely on lay theories when assessing bias in others. As a result, people are more inclined to think that they are guilty of bias in the abstract than in specific instances. Also, people tend to believe that a personal connection to a given issue is for them a source of accuracy and enlightenment but that for others it is a source of bias (Ehrlinger et al., 2005). Because of the importance of transparency and appearance in the regulatory process, motivational bias may be an important consideration in perceptions of the legitimacy of an EE, whether or not it actually influences an assessment.

3.5.5.6 Cognitive Limitations of Experts

Human limits on cognition restrict the complexity of the relationships that can be attempted with EE. Hence, eliciting conditional probabilities relating three or more variables can be very difficult (O'Hagan, 2005; O'Hagan et al., 2006). Poor performance when the number of variables increases can be caused by many different factors. Among the major factors are the degree of correlation among the variables (a higher degree of correlation produces more confusion), human information processing capacity, and barriers to learning (Fischhoff, 2003). Proper accounting of dependence is critical, both to establish the values of other quantities on which that quantity is being conditioned and also—if being asked about multiple quantities—to

elicit their judgments about dependencies among these quantities. As the number of parameters increases, one's ability to either properly define the exercise or to maintain or establish the dependencies between quantities is diminished. Experts who do not work with a specific model in which a parameter is defined may have little knowledge about the value of the parameter. Moreover, the relationship between the parameter value and outcomes that are potentially measurable may depend on the choice among several alternative models, some or all of which the expert may be unfamiliar or even reject. Although these may all be "knowable" to a "clairvoyant," it is not feasible to assume that an expert could make these connections.

In addition, Hamm (1991) argues that expert performance in EE can be compromised when the expert does not make numerical judgments carefully because he does not understand the model or is unfamiliar with elicitation language that is different from how he expresses himself within his field (Hamm, 1991). Even when they are involved in model construction, experts tend to think about a few familiar cases rather than consider all applicable cases. This well-documented phenomenon has motivated the development of many decision-support tools (Chechile, 1991; Saaty, 1990).

Another limitation is that experts may not update their beliefs when new information becomes available (Meyer and Booker, 2001). Judgments are contingent on the consideration of all possibilities and the assessment of a relative belief in one possibility compared with the other possibilities. Shackle (1972b) argues that when experts lack knowledge about all alternatives, the premise of subjective probabilities is violated.

3.5.5.7 Experts and Controversial Issues

Mazur (1973) found that when issues are controversial (e.g., nuclear power and water fluoridation) experts performed as nonexperts. He noted that many conflicts were not disagreements among experts but rather arguments about different points (i.e., a different understanding/definition of the problem). The conflicts resulted from divergent premises, usually resulting from poor communication between adversaries. When scientific controversies contain subtle perceptions and nuances, each side's proponents must simplify the conclusions in a manner that results in one side choosing to accept as a conclusion what the other side regards as an unproven hypothesis. Then, each side seeks to gain acceptance of its view through nonscientific persuasion. The real uncertainty of the science is "removed" by proponents of each side as they state their own case with increasing certainty, not necessarily supported by the scientific information. In addition, the desire to "win" the argument may heighten conflicts.

Experts can take different positions along the uncertainty/ambiguity continuum and are subject to the same biases as lay people. For both experts and lay people, controversy heightens emotion and subjugates ambiguities. Therefore, when the issue is controversial and the

science/scientific analysis is ambiguous or uncertain, the value of that science/scientific analysis may be more questionable. The process of EE may reduce the emotion and ambiguity in an assessment by breaking down a problem into distinct questions and carefully and explicitly defining the uncertain quantities, events, relationships, or parameters of interest.

3.5.5.8 Additional Limitations of Eliciting Experts

Three additional cautions are offered for the decision of whether or how to use EE. The first pertains to an expert's carefulness when translating qualitative thinking to quantitative judgment. Hamm (1991) states that if experts are careless, they may unintentionally respond without adequately understanding the implications of their responses. Sometimes, the carelessness is an artifact of experts failing to understand how their responses will be used. Furthermore, when the decision context is either not shared with or understood by the expert, factors that might have been considered important when responding may be forgotten or ignored.

The second caution pertains to whether or to what degree models correspond with how experts think of the problems presented to them. By design and necessity, models are simplifications. They provide approximations of reality but in the process abstraction, may alter the problem being analyzed.

The third caution pertains to the need to balance analysis and experience (Hammond et al., 1987). Hammond states that judgment, even by experts, includes subjectivity. This does not mean that the judgment is arbitrary or irrational; however, the use of scientific information requires interpretation of its meaning and significance. Because the facts do not speak for themselves, there is a need for experts to provide judgments about the information. When experts from the same discipline disagree, it is important to evaluate where and why the disagreements arise. The analysis of disagreement can be the source of important insights.

3.5.6 How Can the Quality of Expert Judgments Be Improved?

These limitations do not preclude the use of EE methods; rather, they provide targets for their improvement. Heuristic biases are well-recognized by EE practitioners, and methods have been developed to control those biases. For example, anchoring biases can be mitigated by first eliciting extreme values of distributions (e.g., the highest and lowest possible values) and then the median. Also, the same information can be elicited redundantly to help validate values. In any case, EE practitioners are urged to be mindful of these biases when designing elicitations and interpreting results. An understanding of these biases highlights the need for rigor in the design and implementation of an EE protocol. This rigor can produce a more credible analysis. Chapter 5 describes many of the approaches to reducing the impact of such biases and heuristics. In addition, to controlling for these biases the quality of the expert judgment can be maintained if

the problem is formulated in such a way that is unambiguous and within the confines of the expert's expertise.

3.5.7 How Can the Quality of An Expert Elicitation Be Assessed?

In most circumstances, the true accuracy of an EE cannot be quantified. Uncertainty exists in the experts' understanding of the process and in the process itself. Disentangling these two sources of uncertainty is difficult. Furthermore, experts are not value-free (Shrader-Frechette, 1991; Slovic et al., 1988) and bring their own biases to the elicitation (Renn, 1999; Renn, 2001; Slovic et al., 1988). Although experts' training and experience add credence to their judgments, it still is a judgment and incorporates values. Expert judgments are a fusion of science and values. Garthwaite et al. (2005) summarizes by stating that a successful elicitation "faithfully represents the opinion of the person being elicited" and is "not necessarily 'true' in some objectivistic sense, and cannot be judged that way." In the case of EE results being used as missing data, it is important for EE practitioners to understand that they are not obtaining traditional experimental data. If expert judgments are being combined (as discussed in Section 3.5.3), Garthwaite and colleagues' caution become even more important, as the combination of judgments presumes that the "truth" lies within the spectrum of those judgments.

3.6 SUMMARY

For the purposes of this paper, EE is defined as a formal process for developing quantitative estimates of the probability of different events, relationships, or parameters using expert judgment. The use of EE may be of value where there is missing data that either cannot be obtained through experimental research, where there is a lack of scientific consensus, and/or where here is a need to characterize uncertainty. As with many other analytical tools, however, there also are significant theoretical and methodological cautions for its use. As with experimental data, users of the elicited expert judgments must be aware of the biases incurred as a result of the choice of methods and to biases that are caused by the fact that experts are human too. Expert judgment does not exist in a vacuum, apart from sociological and personality influences. Therefore, the best uses of expert judgments obtained through probabilistic EE are those that consider the impacts of these biases on the final results and provide good documentation for how the expert judgments were obtained and how they will be used.

4.0 WHAT THOUGHTS ABOUT APPLICABILITY AND UTILITY SHOULD INFORM THE USE OF EXPERT ELICITATION?

This Task Force recognizes that EE is one of many tools to characterize uncertainty and/or address data gaps. Many factors influence whether an EE would be helpful and should be used by EPA, including the: (1) purpose and scope of the EE (e.g., to estimate missing data or characterize uncertainty), (2) nature of available evidence and the critical uncertainties to be addressed, (3) nature of the overarching project or decision that the EE will support, (4) potential time and resource commitment required for conducting EE, and (5) impact on the decision in the absence of the perspective provided by the EE.

The advantages and disadvantages of EE should be evaluated in light of the particular application for which it is being considered. An EE may be advantageous because it uses experts to help characterize uncertainties and/or to fill gaps when additional data are unattainable within the decision time-frame, allowing analysts to go beyond the limits of available empirical evidence. This is especially important when additional data are unavailable or unattainable. An EE may be disadvantageous where the perceived value of its findings is low and/or the resource requirements to properly conduct an EE are too great. It may be necessary to balance the time, money, and effort needed to conduct a defensible EE against the requirements of other forms of expert input, such as external peer review. This chapter reviews some questions and issues that influence the decision to conduct an EE. The process for considering formal EE to characterize uncertainty and/or address data gaps can be divided into three steps:

1. How important is it to quantitatively characterize major sources of uncertainty or address a critical data gap in a particular case?
2. Is EE well-suited for characterizing the uncertainty or for providing estimates to address a particular data gap?
3. Is EE compatible with the overall project needs, resources, and timeframe?

Each of these is discussed in turn, below.

4.1 HOW IMPORTANT IS IT TO CONSIDER UNCERTAINTY?

To support its decisions, EPA often conducts complex assessments that draw on diverse expertise. In many cases, empirical data are unavailable on the outcome of interest. Therefore, EPA may rely on extrapolations, assumptions, and models of the real world.

These abstractions of reality all are sources of uncertainty in the assessment. When risk or cost information is presented to decision-makers and the public, the findings often have been

reduced to a single numerical value or range. This approach may provide insufficient information to the decision-maker and has the hazard of conveying undue precision and confidence.

In their text *Uncertainty*, Morgan and Henrion (1990) present criteria that define when considering uncertainty is important:

- When people's attitudes toward uncertainty are likely to be important (e.g., if uncertainty itself is likely to be an argument for avoiding a policy option).[13]
- When various sources of information need to be reconciled or combined, and some are more certain than others (i.e., where more weight should be given to the more certain information).
- When deciding whether to expend resources to collect more information (e.g., prioritizing possible areas of research or data collection or deciding whether to seek additional data).
- When the "expected value of including uncertainty" (EVIU)[14] is high, such as when the consequences of underestimating would be much worse than for overestimating (e.g., underestimating the time needed to get to the airport may have severe consequences, so uncertainty about traffic has to be taken into account).

4.1.1 What is EPA's Position on Characterizing Uncertainty?

EPA has long recognized the importance of characterizing uncertainty. To that end, it has established Agency-wide policy and guidance that provide attention to how uncertainty is characterized and addressed: EPA's risk characterization policy (USEPA, 1992; USEPA, 1995a), *Principles of Monte Carlo Analysis* (USEPA, 1997), *Risk Characterization Handbook* (USEPA, 2000a), *Risk Assessment Guidance for Superfund (RAGS) Volume 3* (USEPA, 2001a), and the *Risk Assessment Staff Paper* (USEPA, 2004a).

EPA's *Risk Characterization Handbook* (USEPA, 2000a) provides examples of ways to characterize uncertainty. The Handbook states:

"While it is generally preferred that quantitative uncertainty analyses are used in each risk characterization, there is no single recognized guidance that currently exists on how to

[13] Numerous studies have shown that many people will choose a course of action that has a more certain outcome over one with a relatively uncertain outcome, even if the expected net benefits are somewhat higher in the uncertain case. They are willing to give up the additional expected benefit just to avoid the uncertainty. This behavior is called "risk aversion," and the "risk premium" is the amount of benefit they are willing to give up in exchange for avoiding the risk of loss (i.e., avoiding the uncertainty).

[14] EVIU, as defined by Morgan and Henrion (1990), refers to the quantitative impact that uncertainty analysis can have on a decision. From a decision-analytic perspective, the EVIU is a measure of how much the expected value outcome of a decision will increase if uncertainty is included in the analysis. The EVIU is high if considering uncertainty can lead to a decision with a higher expected value outcome.

conduct an uncertainty analysis. Nonetheless, risk assessors should perform an uncertainty analysis. Even if the results are arrived at subjectively, they will still be of great value to a risk manager. The uncertainty analysis should, in theory, address all aspects of human health and ecological risk assessments, including hazard identification, dose-response assessment, and exposure assessment. Uncertainty analysis should not be restricted to discussions of precision and accuracy, but should include such issues as data gaps and models."

EPA's *Guidelines for Preparing Economic Analyses* (USEPA, 2000b) presents a tiered, practical approach:

> "In assessing and presenting uncertainty the analyst should, if feasible: present outcomes or conclusions based on expected or most plausible values; …[and as an initial assessment] perform sensitivity analysis on key assumptions…. If, however, the implications of uncertainty are not adequately captured in the initial assessment then a more sophisticated analysis should be undertaken …. Probabilistic methods, including Monte Carlo analysis, can be particularly useful because they explicitly characterize analytical uncertainty and variability. However, these methods can be difficult to implement, often requiring more data than are available to the analyst."

EPA practice often involves a "tiered approach" to conducting uncertainty analysis; beginning as simply as possible (e.g., with qualitative description) and sequentially employing more sophisticated analyses (e.g., sensitivity analysis to full probabilistic). These additional analyses are only added as warranted by the value added to the decision process (USEPA, 2004a).[15] This approach focuses on the need to balance limited resources, time constraints, and analytical limitations against the potential for quantitative uncertainty analysis to improve the analysis and regulatory decision.

4.1.2 How Can Expert Elicitation Characterize Uncertainty and Address Data Gaps?

As previously discussed, EPA generally has used EE to address data gaps and/or characterize uncertainty surrounding estimates of important quantities, such as the health impacts of a specified change in air quality. Because EE can provide subjective probability distributions that quantify uncertainty estimates, it often is suggested as an analytic method worth considering. In general, EE can be useful when:

[15] Although it may be important to *consider* each source of uncertainty in an analysis, it may not make sense to *quantify* every uncertainty. This is an important distinction because quantifying uncertainty is sometimes very difficult and not always very useful. Some sources of uncertainty can be adequately addressed with a qualitative discussion and a judgment that quantitative analysis is not merited.

- Acceptable quantitative estimates of uncertainty cannot be made adequately with additional data collection (e.g., cannot be observed, such as future oil prices), cannot be observed *directly* (e.g., effects of a new substance on human health), or the events are so rare that data are very limited (e.g., risk of nuclear plant accident). Statistical methods cannot address this type of data limitation. When empirical data are essentially impossible to obtain, EE is a viable approach to quantification.
- Uncertainty estimates using other techniques will not be quantified adequately because of the timeframe for a decision or decisions about available resources. Situations may arise in which data collection would need more time than analyses based on expert judgment, data collection is not technically feasible, or the benefits of additional data collection in terms of improved confidence may not justify the cost and/or time.

As defined in this document, EE is a formal process for developing quantitative estimates for the probability of unknown events, relationships, or parameters using expert judgment (SRI, 1978; Morgan and Henrion, 1990). EE goes beyond empirical data and allows experts to integrate across various lines of evidence. When data are unavailable or unattainable[16], EE may be used to either fill data gaps and/or to characterize uncertainty associated with available empirical data.

4.1.3 What Are the Alternative Methods for Expert Judgment?

As described in Chapter 2, EE is one of many expert judgment methods, including activities that range from informal to formal. Other expert judgment methods include public comment and peer review. These methods vary in their level of rigor and the degree to which they control for heuristics and biases. They also differ in the range of questions that they can address, the level of effort and resources required, and the degree of public acceptability. Table 4-1 presents basic descriptors for expert judgment methods that should be considered when determining whether an EE should be conducted. One should consider whether a particular activity is compatible with the timeline, available resources, and the overall nature of the issue and decision-making process. Table 4-1 compares various methods of expert judgment in terms of resource needs. It should be noted that these forms of expert judgment are not necessarily comparable in terms of purpose and their ability to provide equivalent information to decision makers and stakeholders. If detailed quantitative characterization of uncertainty is necessary,

[16] Although data may be unattainable, the quantity of interest should be defined unambiguously as a quantity that is measurable (at least in principle, if not in practice).

then EE could be compared to a range of uncertainty methods. The estimates provided in Table 4-1 focus on EE, which should be compared to other methods that characterize uncertainty.

Table 4-1. Illustrative Comparison of EE and Other Methods for Expert Judgment

	Public Comments	**Limited (Letter) Peer Review**	**Formal FACA (or Panel) Peer Review**[17]	**Expert Elicitation**
Problem addressed	Broad, no limit, defined by commenter	Broad, but defined by charge	Broad, but defined by the charge	Narrow, specific, and well-defined
Timeframe	Typically 30–90 days	1–4 months	4–12 months[18]	8 months–2 years
Resource needs	Limited	~$25K	~$250K	~$250K–$2M
Role of public/ stakeholders	Open to all to provide comments	Formal selection process	Public nomination, selection process, open public process	Nominations by peers and limited involvement of public/stakeholders
Evidence considered	No limit	No limit	No limit	No limit but must be formally shared with all experts to evaluate
Acceptance	Publicly acceptable	Familiar though not transparent to public	Generally accepted, recognized	Some wary of method (i.e., concerns about perceived bias)
Selection of experts	None	Formal selection process	Formal and public nomination process	Systematic process usually involving nomination by technical experts

[17] Note: Review by the NAS could also be included in this category. If so, time and resource requirements may be substantially increased compared to Panel or FACA review such as with the SAB.
[18] This timeframe includes only the meeting time and not the time to set up the committee, which takes about 8 months.

4.2 WHAT IS THE NATURE OF THE UNCERTAINTIES TO BE ADDRESSED?

Many different sources of uncertainty may arise when models and analyses are used to support policy analysis and decision-making. This section presents different sources of uncertainty and discusses how EE could be used to address them. As discussed below, EE has the potential to be helpful in characterizing uncertainty regardless of its source.

4.2.1 What are the Categories of Uncertainty?

Based on the literature (Cullen and Frey, 1999; Finkel, 1990; Hattis and Burmaster, 1994), sources of uncertainty can be classified into four categories:

- ***Input (or parameter) uncertainty*:** Models and assessments utilize a wide range of parameters and other inputs to generate estimates. Typically some values for these inputs are not known with confidence. Among the factors that can introduce uncertainty into model inputs are random error (including lack of precision in measurement), systematic errors (i.e., bias), lack of empirical data, and lack of representativeness of empirical data.
- ***Model uncertainty*:** All models include uncertainty about the appropriate modeling approach (e.g., which model best represents reality, including how inputs and constants should be combined in equations based on an understanding of the real world). Because models are simplified representations of the real world, uncertainty can result from imperfect knowledge about the appropriate conceptual framework, specific model structure, mathematical implementation, detail (precision/resolution), boundary conditions, extrapolations, and choices among multiple competing models.
- ***Scenario uncertainty*:** Decisions related to the overall design of the scenarios modeled in the analysis (e.g., selection of receptor populations, chemicals, exposure sources, and study area delineations) can be a source of uncertainty for analyses.
- ***Decision rule uncertainty*:** Decisions on the types of questions asked to support policy-making and the theoretical framework used to make those decisions can introduce uncertainty. These areas of uncertainty include: (1) the design of the decision framework used in guiding policy-making (e.g., acceptable risk levels and the types of risk metrics used such as individual- versus population-level risk) and (2) global protocols used in the analysis (e.g., use of standard risk reference doses and associated hazard quotient values as the basis for noncancer risk assessment versus the use of epidemiologically based disease incidence estimates).

4.2.2 Which Types of Uncertainty Are Well-Suited for Expert Elicitation?

If a problem statement can be formulated clearly and consensually, then EE may be used to address any uncertainty that is measurable (at least in principle, if not in practice). When an adequate knowledge base exists and there are qualified experts, their judgments can form a credible basis for insight about uncertainty. The questions must be characterized adequately so that experts can understand how to apply their knowledge and experience. EE may not be suitable for problems that are intractable or those for which the complexity is so great that experts would not have sufficient targeted expertise in the area of interest.

Analyses that support EPA decisions typically involve numerous components, such as risk assessments that include toxicity assessments, emissions or discharge estimates, air or water quality modeling, exposure assessment, economic impact analyses, and so on. For each of these steps, EE may be valuable, but EE may be more useful or appropriate for some steps than others. When EE is considered for use in an analysis, EE may be more or less appropriate depending on what specific questions the EE would be used to address (Morgan and Henrion, 1990). Therefore, it is important to identify a very specific problem statement and questions when deciding whether to use EE in an analysis. Experts should not be used to define regulator-related questions but may be used to evaluate among options.

4.3 WHAT ARE OTHER METHODS TO CHARACTERIZE UNCERTAINTY?

In addition to EE, other methods are available to characterize uncertainty. Some methods can both characterize the uncertainty of particular parameters and propagate this uncertainty through the model. For example, EE can be used to develop subjective probability distributions; characterize the uncertainty of specific parameters, events, quantities, or relationships; and estimate the overall uncertainty of a modeled process. The context for characterizing uncertainty can vary greatly. For example, the focus could be on a single modeling parameter. An EE could be conducted to estimate the magnitude of a cancer slope factor (including that number's uncertainty). The carcinogenic process for that chemical, however, could be viewed as a complex multi-element process with multiple modeling steps. In this case, the estimates from an EE may be used to propagate uncertainty through an entire model (e.g., the cancer process).

Methods for probability-based uncertainty characterization can be divided into five broad categories: (1) statistical/frequentist (e.g., Monte Carlo and Latin Hypercube Simulation); (2) judgmental/subjectivist (e.g., EE, Bayesian); (3) scenario analysis; (4) other (e.g., interval, probability bounds, fuzzy logic, and meta analysis); and (5) sensitivity analysis techniques. These categories and methods are discussed briefly below:

- ***Statistical/frequentist*:** These uncertainty characterization methods are based on the frequentist paradigm and hence require empirical data to establish a probabilistic characterization of uncertainty. These approaches treat probability as an objective measure of likelihood based on frequencies observed in data that are subject to sampling error, measurement error, and other (assumed to be) random processes. For example, wind speed and its associated uncertainty could be described by reporting the range, mean, and 95th percentile of historical measured values. Common methods that are founded on the frequentist paradigm include numerical uncertainty propagation, bootstrap, and response surface methods. Some of these methods, such as bootstrap, can quantify uncertainty even with very small sample sizes. In general, they are less capable and used less often to characterize uncertainty about model choice or causality. In addition, they typically cannot address uncertainty arising from data that are not representative of the value to be estimated (e.g., an epidemiological study that focused on a population that is very different from the one to be analyzed in a risk assessment in which no data on the population differences are available).
- ***Judgmental/subjectivist*:** These methods are based on the concept that probability is an expression of the degree of confidence in some parameter, quantity, event, or relationship. In addition, they are based on the concept of logical inference—determining what degree of confidence an expert may have for various possible conclusions, based on the body of evidence available. Common subjectivist methods include Bayesian analysis, EE, and Generalized Uncertainty Likelihood Estimation.
- ***Scenario analysis*:** Uncertainty can be characterized through presentation of alternative scenarios that are thought to span the range of plausible outcomes. Scenario analysis is useful to: evaluate groups of variables or assumptions that are correlated and/or vary together (e.g., worst-case scenario), predict future conditions, and assess model uncertainty.
- ***Other methods*:** In addition, there is a group of diverse methods that do not depend heavily on subjective judgment (as does Bayesian analysis) and can be applied in contexts where uncertainty characterization is limited by inadequate empirical data. These methods occupy a middle ground between the frequentist and subjective methods. They include interval methods, fuzzy methods, and meta-analysis.
- ***Sensitivity analysis techniques*:** These methods assess the sensitivity of the results to choices of inputs, assumptions, or models. Sensitivity analysis does not necessarily quantify the probability of those alternatives choices, however. Methods for sensitivity analysis include local methods (these examine the impact of individual

inputs in relative isolation on model outputs); combinatorial methods (varying two or more inputs simultaneously while holding all other inputs constant and determining the impact on model output); and global methods (these generate output estimates by varying inputs across the entire parameter space and determine contribution of individual inputs to overall uncertainty).

4.3.1 How Does Expert Elicitation Relate To the Other Methods?

EE falls within the judgmental/subjectivist category of methods. These methods have the advantage that they can provide a robust characterization of uncertainty without requiring as much data as frequentist approaches.

For a particular analysis, the uncertainty characterization is not limited to a single method. In some situations, EE can be used as a complement or substitute for other methods. For example, EE may be employed to generate a probability distribution for an input parameter, whereas a frequentist approach is used for other input parameters. Alternatively, EE can be applied to assess the appropriateness of specific model choices, whereas a frequentist approach can be drawn on to address uncertainty in the inputs to those models.

Table 4-2 describes some of the strengths and weaknesses of EE. This can be used to compare EE with other methods for aggregating information and quantifying uncertainty from multiple sources. This list of strengths and weaknesses is presented in the context of environmental analysis. It is expected that the strengths and weaknesses would be similar for other application areas. Note that some of the weaknesses could also be described as caveats for the appropriate use of the EE-derived information (i.e., cautions about how EE output could be misused or misinterpreted).

Table 4-2. Strengths and Weaknesses of EE

Strengths:

- EE provides a means to bridge data gaps in modeling and analysis. The information obtained from EE, however, does not represent "new" information. Rather, it reflects human extrapolation and interpolation from existing data and experience. EE also can be used to characterize uncertainty associated with a particular analysis element.
- Because EE is a structured approach, the probabilistic distributions and characterizations of uncertainty that it provides can be transparent and rigorous. Specifically, the use of a highly structured approach (including for example, pre- and postelicitation workshops and carefully structured elicitation procedures) ensures that experts posses a common baseline understanding of the issue being addressed and that they have access to a similar set of relevant data.
- The use of a structured approach in eliciting input from the experts helps identify potential sources of bias in the information provided by the experts. One extension of this idea is the potential to evaluate the experts' responses to understand their thought processes.
- The structured approach helps to avoid mitigating factors such as the influence of personalities and group consensus pressure.
- EE provides a means for obtaining perspectives from a group of experts, which can (and perhaps should) reflect a broad range of informed views on the issue.
- The application of EE to the overall analysis can be highly flexible. It can be combined with probabilistic simulation (e.g., uncertainty distributions derived by EE can be used in probabilistic simulation).

Weaknesses:

- Information obtained from EE can be mischaracterized and used inappropriately (e.g., caveats on what it represents can erode such that the information begins to have the same standing as empirically obtained information).
- It is easy to aggregate the output of EE in inappropriate ways (e.g., averaging opinions to obtain a single best estimate).
- If the EE process lacks proper design and conduct, a range of heuristic biases can impact the credibility of the outcomes.
- There is a potential that the model structure will not correspond to the way experts conceptualize the issue (e.g., differences in the form of dose-response functions as at low exposure levels). This can have the unfortunate affect of constraining, and potentially biasing, the experts by imposing an "uncomfortable" conceptual framework on the issue. If the EE process does take the time to determine why experts disagree, including model uncertainty, the process may provide outputs that do not represent the experts' true perspectives.
- Time and cost can be significant. Done properly, EE can be relatively resource-intensive and time-consuming. Some shortcuts may jeopardize the integrity of the process.

4.3.2 How Relevant and Adequate Are Existing Data?

As mentioned above, EE may be useful when empirical data are severely limited or are contradictory. An assessment of the adequacy of any empirical data and theory should be part of a decision to use EE requires. Data may be limited in quantity, quality, or both; adequate data may be difficult or even impossible to obtain. Quality problems may arise from relevance of the data or problems with imprecision or bias in the data.

4.3.2.1 What Types of Evidence Are Available?

Rarely are there direct empirical data that are specific to the quantity of interest within the exact context of interest. In other words, EPA rarely has direct observations of the impacts of environmental pollutants at concentrations encountered in the environment within the specific exposed population. As a result, EPA often makes inferences based on lines of evidence (Crawford-Brown, 2001). These five types of evidence can be ordered by relevance as shown in Figure 4-1.

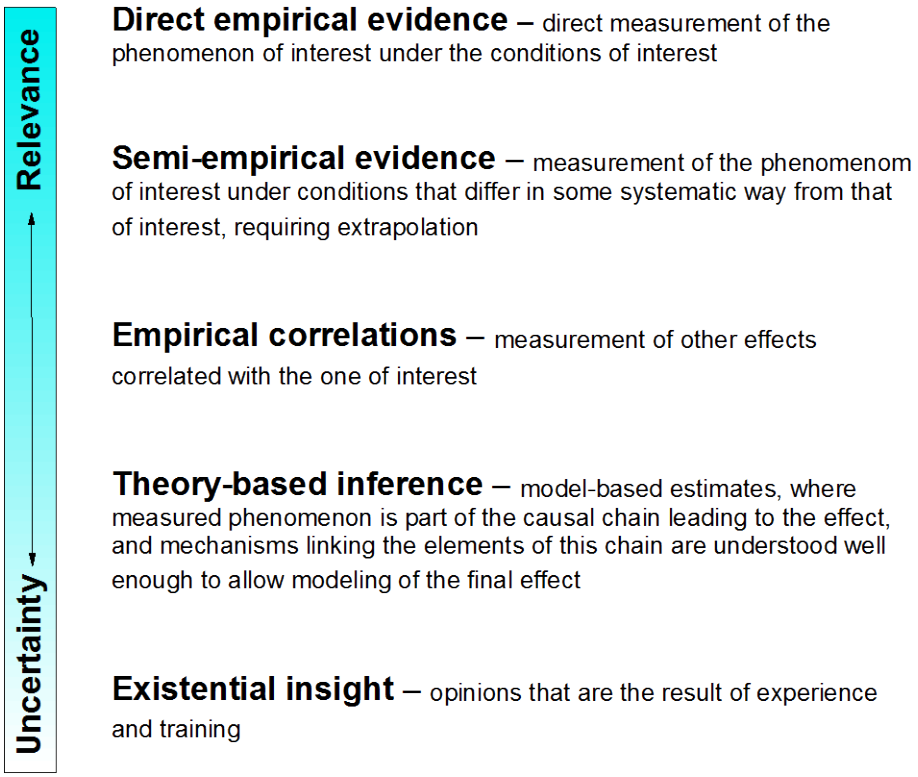


Figure 4-1. Five Types of Evidence

A key consideration for assessing the adequacy of empirical data and the possible utility of EE is to evaluate the representativeness of existing data (e.g., studies are limited to animal data, non-U.S. populations, or unique subpopulations). EPA's *Report of the Workshop on Selecting Input Distributions for Probabilistic Assessments* provides useful methods for assessing and handling data with respect to its representativeness (USEPA, 1999a). The option of working with real data, given their limitations, should be considered as an alternative to EE. One should consider whether there are suitable approaches to adjust data so that they are sufficient to support a specific decision. Furthermore, EE may be a useful approach allowing experts to provide judgments based on adjusting empirical data that are not entirely relevant, such as in determining how data from another population may be used to represent the U.S. population.

4.3.2.2 What is the Quality of the Available Information?

Knowledge of the type and relevance of data is insufficient to evaluate the need for EE. Knowledge of data quality also is needed. One may have a study that represents direct empirical evidence, but the quality of the study may be poor (e.g., poor level of detection). In some cases, EPA has only a very limited database of information on a critical element of an analysis (e.g., very small sample size measuring a poorly characterized population). For example, there may be only five known measurements of a certain emissions rate, and additional measurements may be very costly to obtain. Even assuming the data points are representative, which is not necessarily the case, such a small sample provides very limited information. The average value can be estimated, as can be the degree of uncertainty over that average. Separating variability from uncertainty in this case would present another challenge.

EPA has utilized and documented statistical methods for estimating distributions that describe parameter uncertainty and variability in a variety of analyses (*Options for Development of Parametric Probability Distributions for Exposure Factors*, USEPA, 2000c). Techniques also exist that can provide useful information about uncertainty using even very small data sets (i.e., less than 20 data points) (e.g., Cullen and Frey, 1999). These methods should be explored as options before launching into an EE to solve the problem of data limitation.

4.3.2.3 Are There Critical Data Gaps?

The many uncertainties within an assessment will have varied impacts on the overall assessment or confidence in that assessment. Sensitivity analysis can identify the most critical uncertainties and these should be the focus of uncertainty analysis.

As described in Section 5.2.1, an EE typically will be of most value when it is possible to define a very limited number of critical uncertainties to provide insights about a particular event,

quantity, parameter, or relationship. Because defensible EE efforts can be very resource intensive, it is best to identify those areas that would most benefit from this concentrated analysis. Various techniques are available to identify the relative contribution of various uncertainties and could be used to help identify appropriate targets for an EE (Renn, 1999; Wilson, 1998; Warren-Hicks and Moore, 1998; and Stahl and Cimorelli, 2005).

4.3.2.4 Is the State of Uncertainty Acceptable?

Inferences from data and the quality of those inferences are based on the entire body of evidence. Typically, multiple categories of evidence are available, and it may be difficult to determine how to combine these bodies of evidence. This combination can be accomplished by the judgments of scientific experts who evaluate the categories of evidence, the weights given to those categories, and the quality of evidence within each particular category.

Decision-making under uncertainty is inherent and unavoidable. Overall uncertainty affects the correctness and acceptance of decisions. By using uncertainty analysis, understanding of evidence, its impact on the overall estimate, and confidence in the decisions informed by an estimate can be improved.

In general, the acceptability of uncertainties is judged by decision-makers, scientists, and stakeholders after the data are placed in the context of, or used for, a decision. This is because the extent to which data uncertainty impacts the availability and value of decision options is not known until the context of the decision is clear. Hence, the acceptability of uncertainty is necessarily a contextual decision. Further discussion about the role of uncertainty in a decision context is available elsewhere (Jamieson, 1996; Renn, 1999; Wilson, 1998; Warren-Hicks and Moore, 1998; Harremoes et al., 2001; Stahl and Cimorelli 2005).

4.3.2.5 Is Knowledge So Limited That EE Would Not Be Credible?

In some situations with scant empirical data, expert judgments along with sound theoretical foundations can form the basis for EE. In situations in which knowledge is scant, however, the EE should not be seen as a proxy for creating “data” where none exist. In such circumstances, it may be appropriate to use EE to support scenario analysis for ranking rather than a formal quantification of options. Quantifying options may give the appearance of greater precision than is defensible.

Within complex decisions, debate and opposition often are related to specific concerns of stakeholders and the diverse manner in which they frame issues. Because individuals react to risks on multiple levels, including analytical, emotional, and political, differing approaches can lead to conflict. This is supported by recent brain imaging research that explores how people react to risk and uncertainty. Hsu et al. (2005) have shown that neural responses are different

when reacting to risk (risk with a *known* probability, based on event histories, relative frequency, or accepted theory) vis-à-vis ambiguity (risk with *unknown* probability, or uncertainty about risk levels—meager or conflicting evidence about risk levels or where important information about risk is missing). Attempting to quantify uncertainties may be at odds with how people perceive and react to a situation. In fact, the International Risk Governance Council (2005) recognized the importance of this factor in the development of its integrated framework, in which knowledge is categorized to allow distinction between "simple," "complex," "uncertain," and "ambiguous" risk problems.

This does not imply that EE is inappropriate for these purposes; rather, this means that care must be taken to conduct a decision-making process in which stakeholders will view EE as a credible process. In these circumstances it is worth considering whether the major stakeholders would view EE as credible. For example, such an approach was used successfully to forecast the potential microbial impact of a Mars landing (North et al., 1974). This problem involved an area with no history or expectation of empirical data. In this case, the success of the EE exercise may have been because it was a technical exercise with limited stakeholders. EPA, however, tends to be involved with problems that are complex and involve numerous diverse stakeholders (e.g., regulated community, environmental nongovernmental organizations, community members, and independent scientists). EE can be used to support scenario analysis (such as focusing on mental models), but as it becomes more quantitative, it may become challenging. This challenge will be particularly critical when data are scant, theoretical foundations on which to base judgments are limited, and/or multiple stakeholders are involved in the discourse. In addition, the use of uncertain data tends to be more accepted when stakeholders know how and for what purpose that data will be used. Placing uncertain data within a particular decision context may limit the use of those data and increase stakeholder acceptance for using those data for that limited purpose.

4.3.2.6 Can Additional Data Be Collected To Reduce Uncertainty?

The possibility of seeking additional data should be considered carefully in any decision about using EE. In addition, one should also evaluate the option of working with imperfect data (e.g., not representative) by using suitable approaches to adjust them. As for any analysis, the proposed use of EE data should be clearly defined in the context of the assessment. Another option is using techniques to obtain useful uncertainty assessments for very small data sets (e.g., less than 20 data points). These methods for using imperfect data should be considered before the data limitation is established as the rationale for using EE. One should bear in mind, however, the time and resources needed to augment or improve data, which itself can be a lengthy process. In many circumstances, EE may provide information more promptly.

4.4 WHAT ROLE MAY CONTEXT PLAY FOR AN EE?

EE is suitable for many EPA activities, including identification of research needs, strategies, and priorities; risk assessments for human or ecological health; and cost-benefit analyses to support major regulatory decisions. The context of each potential use may indicate whether it is appropriate to use EE, including the level of scientific consensus, the perspectives of anticipated stakeholders, and the intended use of results.

4.4.1 What is the Degree of Consensus or Debate?

Another consideration about whether to rely on existing data or to conduct an EE is the degree of consensus in the scientific community. One of EE's strengths is that it provides the carefully considered and fully described views of several highly respected experts who are affiliated with diverse institutions and perspectives. Obtaining these cross-institutional viewpoints may be preferable to relying on the views of an in-house expert, judgments from an advisory committee, or otherwise limited data. When evaluating the status of scientific consensus or debate, the following factors may indicate that EE is applicable:

- Conflicting empirical evidence and lack of consensus on selecting analytical options.
- No clear consensus exists, and there is substantial debate among experts.
- The problem concerns an emerging science challenge and/or the scientific controversies include model and/or data selection or use.
- The range of views are not easily articulated or captured by EPA's existing professional judgment processes (e.g., analytical approaches and external peer review).
- Problems are complex and multidisciplinary and hence need methodical deliberation by a group of experts to become tractable.

4.4.2 Will Stakeholders View Expert Elicitation as Credible?

Given the novelty of EE to many potential user communities, stakeholder concerns should be considered early in the process. One potential concern is that subjective assessments may be viewed as unreliable and unscientific. Other stakeholders may have concerns regarding potential bias and manipulation of expert judgments (e.g., American Bar Association, 2003; Natural Resources Defense Council, 2005). Consequently, if EE is to be used in regulatory decisions, transparency is critical for credibility. The presentation of results should take into account what is known about effective risk communication (Bloom et al., 1993; Johnson, 2005; Morgan et al., 2001; Slovic et al., 1979; Slovic, 1986; Thompson and Bloom, 2000; Tufte, 1983; USEPA, 1994, 1998, 1999b, 1999c, 2001a, 2001b). This topic is discussed further in Chapter 6.

As part of this communication strategy, stakeholders may need to be shown that subjective judgment (the linking of facts and judgments) is a component of many environmental analyses, not just EE. Nevertheless, some stakeholders may be unfamiliar with the central role that subjective judgment plays in EE. Therefore, one should consider stakeholder perspectives as part of the overall process of deciding whether and how to conduct an EE. Early interactions and effective communication with stakeholders may help to garner support and satisfy their desire to participate in this process. This may be particularly true for stakeholder groups with limited resources and those that lack familiarity with methods of quantitative analyses.

OMB's *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies* (USOMB, 2002) defined "quality" as encompassing objectivity, utility, and integrity. Objectivity is a measure of whether information is accurate, reliable, and unbiased as well as presented in an accurate, clear, complete, and unbiased manner. Furthermore, the guidelines highlight the need for reproducibility so that "independent analysis of the original or supporting data using identical methods would generate similar analytic results, subject to an acceptable degree of imprecision or error." These OMB guidelines recognize that some assessments are impractical to reproduce (e.g., ones that concern issues of confidentiality) and hence do not require that all results be reproducible. In any case, a carefully conducted and transparent EE can meet these goals for objectivity and reproducibility. To demonstrate the feasibility of reproducibility, some research (Wallsten et al., 1983) showed that expert judgments are stable within a reasonable period.

In general, OMB has been very supportive of EE as a means for EPA to improve uncertainty characterization, as indicated by the following statements and actions:

- "In formal probabilistic assessments, <u>expert solicitation</u> is a useful way to fill key gaps in your ability to assess uncertainty."(*Circular A-4*, USOMB, 2003b)
- OMB played an active role in the nonroad diesel EE pilot to demonstrate how one might comply with *Circular A-4*.

4.4.3 What is the Nature of Review or Dialogue of the Overarching Activity?

The nature of the decision-making process and its context can affect whether EE is appropriate. Many EPA decisions are multifactorial, including technical, social, political, and economic elements that involve multiple stakeholders. EE can be used to address specific scientific or technical issues, but if the decision is inherently political, then EE may not provide helpful insights. When values rather than science are critical, other decision analytic methods may be preferred. Some other methods that facilitate decision-making and promote stakeholder interaction include the Analytical Hierarchy Process, Multicriteria Decision-Making, and Multi-

Attribute Utility Theory. The approaches that are used to support Agency decisions differ in their degree of stakeholder involvement and participation. Although the EE process can be transparent and offers stakeholder review, it generally lacks an opportunity for active stakeholder input. If direct stakeholder interaction is desired, it can be added as a supplemental activity to the basic EE process.[19]

4.4.3.1 Is There a Perception of Imbalance?

The success or acceptance of an EE may depend on the quality of stakeholder communications, transparency, and efforts taken to address potential imbalance. The analyst conducting the EE should seek experts who represent a range of technically legitimate opinions and are not tainted by ethical issues including, but not limited to, conflict of interest and appearance of a loss of impartiality. Then, if the process is transparent, any imbalances and their source should be evident.

4.4.3.2 What is the Role of Peer Review in the Overall Effort?

EPA's *Peer Review Handbook* (3rd Edition) outlines the process for determining when to conduct a peer review and for selecting appropriate mechanisms and procedures to conduct peer reviews (USEPA, 2006a). EPA frequently uses formal peer review for assessments or portions of assessments that support regulatory decisions. According to OMB guidelines, highly influential information should undergo external peer review to ensure its quality.

Peer review of an EE should include subject matter experts and experts in the use of EE. As with any peer review, it may be challenging to obtain peer reviewers if the technical domain is small and the pool of relevant experts is very limited. It is important to note that an expert who participates in an EE becomes part of the analysis and would be ineligible to serve as a peer reviewer for the resulting product. This is because peer reviewers should be independent of the process that developed the product. In some cases, when there are few experts available, it may be a challenge to find a sufficient number of experts for both the EE process and the peer reviewer of the product. See Chapter 5 for a more detailed description of the process and criteria that might be used to select experts for an EE.

[19] Although stakeholders may nominate experts, review the protocol, and review the final EE, they may seek a more participatory process than is available via EE.

4.4.4 How Does EPA Expect To Use Expert Elicitation Results?

EPA faces many decisions with varying levels of quality requirements[20] for which EE is potentially applicable. For example, EE may be relevant for the following decisions: identify research needs, develop research priorities, make regulatory decisions, make major regulatory decisions (greater than $100 million impact)—with increasing importance and therefore increasing requirements for information quality. EEs can be costly and resource intensive undertakings. It generally would be a mistake, however, to reduce cost and time requirements by eliminating EE elements that control for biases and heuristics (see Section 4.5). This will reduce the overall quality and advantages of the EE as compared to other, less costly methods. In the final analysis, whether the diminished quality is critical depends on the planned use of the EE's results.

As with all analytic activities, the use of results should guide the design of the protocol and the purpose for which they were developed. When considering a secondary use of results, one should consider the rigor of the protocol design and whether the necessary elements were elicited. For example, if the EE was developed for internal deliberative purposes (e.g., to identify research needs), depending on design protocol, it may be inappropriate to use the results for a regulatory decision that has higher information quality requirements. If it is expected that the results may be used for other purposes (especially uses with higher quality requirements), this may be considered during protocol design. It is prudent to consider whether demands to use results beyond the intended purpose may exist, the potential impact for any misuse, and whether additional resources are needed to ensure the robustness of results.

4.5 WHAT ARE THE RESOURCE IMPLICATIONS WHEN CONDUCTING AN EXPERT ELICITATION?

EEs as defined by this White Paper are generally ambitious undertakings. As described in more detail in Chapter 5, careful attention should be given to the design and conduct of any EE effort to minimize the impact of heuristics and biases. The cost of conducting an adequate defensible EE includes both EPA resources and time and, typically, a significant level of contractor support. Table 4-3 provides a general outline of the various portions of the EE and considerations regarding time and effort.

Predicting resource needs for EE is not straightforward. The resources for an EE, like those for any complex analysis, depend on the design, scope, and rigor desired. As previously discussed, a well-designed EE controls for heuristics and biases, thereby elevating the quality

[20] OMB Information Quality Guidelines, 2002, states "We recognize that some government information may need to meet higher or more specific information quality standards than those that would apply to other types of government information. The more important the information the higher the quality standards to which it should be held."

and credibility of results. Controlling for such heuristics and biases usually requires extensive planning and protocols. Although numerous methodological adjustments can be implemented that would lower the level of effort and, hence, resource needs, such adjustments can affect the overall quality and/or acceptability of results.

This section briefly describes the resource implications and considerations for conducting an EE (see Chapter 5 for additional discussion).

Table 4-3. Potential Pre-Expert Elicitation Activities, Responsibilities, and Time

Activity	Responsibility	Time
Defining the Problem		
Problem definition/scope	EPA Project Manager Senior Managers	2 months
Structuring/decomposing the problem	EPA Project Manager	2–6 months
Identify expertise needed for EE	EPA Project Manager	2 months
Contracting		
Contract planning, secure funding, contract capacity, expertise in EE, write Statement Of Work, bids, selection	EPA Project Manager and Contracting Staff	3–5 months
Financing contractor	EPA Contracting Officer	1–2 years (entire project)
Selecting Experts		
Development of selection approach and criteria, identification, and recruitment	EPA Project Manager and/or Contractor	1–3 months
Review of nominated experts, selection of experts, evaluation of conflicts of interest	EPA Project Manager with Contractor's Project Officer and/or Contractor	1–2 months

4.5.1 How Long Does an Expert Elicitation Take?

A well-conducted and rigorous EE that adequately controls for biases and heuristics can be a lengthy process and require several years. Table 4-3 provides a general picture of the time needed for the individual steps. If the EE includes more complex analyses, additional resources may be required. The time estimates in this table are part of a continuum of activities, from simple to sophisticated. The level of effort for any given EE is contingent on the type of complexity of the assessment, which is influenced by how the information will be used.

4.5.2 What Skills are Necessary to Conduct an Expert Elicitation?

Conducting an EE requires a breadth of activities and skills, each with its own resource implications. The necessary skills that are summarized here and described in more detail in Chapter 5 can be broken down into two categories: organizational and technical. First, an EE involves an expenditure of EPA resources (work years, contracting funds, etc.). As a result, the EE project manager identified in Table 4-3 is involved in many steps of the project. Secondly, significant technical skills are needed to support an EE effort. In addition to expertise about the subject matter, staff with experience should be included in the EE process itself. To properly capture and document expert judgment, one should have both a thorough understanding of the state-of-science for the particular discipline and expertise in the EE process (e.g., cognitive psychology). Whether these expertise can be met internally or require contractor support may depend on the skills and availability of EPA staff. Another important consideration is that if a contractor establishes, controls, and manages the EE process and its experts, FACA requirements do not apply. Typically, the steps outlined in Chapter 5 require involvement of both EPA staff and a specialized contractor. As described in Chapter 5, a team approach may be most effective to conduct the actual elicitations.

4.5.3 How Much Does an Expert Elicitation Cost?

As is detailed in Table 4-1, a well-conducted rigorous EE may require a large resource commitment. Past EPA experiences indicates that such efforts can range from $200,000 to $2 million, depending on the level of effort and rigor. This is generally consistent with the range of external cost estimates—$100,000 to $1 million—for conducting an EE with individual face-to-face elicitations (Hora and Jenssen, 2002; and Moss and Schneider, 2000, respectively). Adjustments can be made to the process that may provide cost savings. As discussed in Section 5.3.4, alternatives to face-to-face elicitation may reduce costs. Such adjustments, however, will typically lessen the rigor that controls for heuristics and biases, and/or reduce transparency, thereby diminishing the overall quality of the results.

4.6 SUMMARY

As described above, there are many technical, administrative, political, and procedural factors that influence whether to conduct an EE. This section summarizes circumstances in which EE might or might not be appropriate. In most cases, EE is but one of several methods that can be used to characterize or address critical uncertainties or data gaps. For these situations, this chapter discussed how various factors may be evaluated to select a preferred method. In some cases, the decision to conduct an EE is unclear and may be influenced by many factors. For a given project, analysts and decision-makers need to integrate the numerous factors discussed above to facilitate a decision on whether to conduct an EE.

4.6.1 What Conditions Favor Expert Elicitation?

The following conditions tend to favor EE:

- The problem is complex and more technical than political but not so complex that experts would be unable to sufficiently target their expertise in the area of interest.
- Adequate data (of suitable quality and relevance) are unavailable or unobtainable in the decision time framework.
- Reliable evidence or legitimate models are in conflict.
- Appropriate experts are available, and EE can be completed within the decision timeframe.
- Necessary financial resources and skills are sufficient to conduct a robust and defensible EE.

4.6.2 What Conditions Suggest Against Expert Elicitation?

The following conditions tend to suggest against EE:

- The problem is more political than technical
- A large body of empirical data exists with a high degree of consensus.
- The findings of an EE will not be considered legitimate or acceptable by stakeholders.
- The information that the EE could provide is not critical to the assessment or decision.
- The cost of obtaining the EE information is not commensurate with its value in decision-making.
- Available financial resources and/or expertise are insufficient to conduct a robust and defensible EE.
- Other acceptable methods or approaches are available for obtaining the needed information that are less intensive and expensive.

5.0 HOW IS AN EXPERT ELICITATION CONDUCTED?

This chapter summarizes the major steps and important factors to consider when conducting an EE. It also describes good practices for conducting an EE based on a review of the literature and actual experience within EPA and other federal agencies. These practices may apply to EEs conducted by EPA or by outside parties for submission to EPA. In general, these good practices consist of the following elements: (1) clear problem definition, (2) appropriate structuring of the problem, (3) appropriate staffing to conduct EE and select experts, (4) protocol development and training, including the consideration of group processes and methods to combine judgment when appropriate, (5) procedures to check expert judgments for internal consistency (verify expert judgments), (6) clear and transparent documentation, and (7) adequate peer review.

These elements are not intended to constitute a prescriptive "checklist." In practice, the implementation of EEs should be flexible, and a range of approaches may be appropriate. Hence, the protocol design for any particular EE involves considerable professional judgment. As stated by Morgan and Henrion (1990), "the process of expert elicitation must never be approached as a routine procedure amenable to cookbook solutions…. Each elicitation problem should be considered a special case and be dealt with carefully on its own terms." This EE White Paper provides goals and criteria for evaluating the success of an EE, which can be achieved by multiple approaches. As noted in Chapter 3, EE may be used to estimate a quantity or parameter for an uncertain relationship or event or to estimate confidence in a parameter, data set, or model. In the discussion that follows, the phrase "uncertain quantity" is used to represent any of these circumstances.

5.1 WHAT ARE THE STEPS IN AN EXPERT ELICITATION?

There are a number of different approaches or types of EE processes. This White Paper generally focuses on EEs that involve individual interviews of the experts. Some EEs, however, entail group processes (e.g., Delphi, group survey, and nominal group technique). In the sections below, the primary discussion concerns EE processes that focus on individuals. In addition, many, but not all, of the elements discussed below are relevant for both individual and group process EEs. Significant differences in the application of individual or group processes are discussed in Sections 5.2 to 5.4.

Based on decision analysis literature (Morgan and Henrion, 1990; Spetzler and Staehl von Holstein, 1975), Figure 5-1 provides an overview—not a one-size-fits-all approach—of the various steps included in a full EE. This approach is intended to be flexible and allow for

innovations that may suit particular applications. The overall process includes pre-elicitation activities (see Section 5.2), conducting the elicitation (see Section 5.3), and postelicitation activities (see Section 5.4). Some steps should be followed in all EEs, whereas other steps are optional (see Figure 5-1). The value added for these optional steps, which may entail additional time and expense, is discussed in the following sections. Many of these individual steps (e.g., problem definition, selection of experts, and conditioning) are components of other analytic methods in addition to EE. Their application in these other methods may be similar or different to their use in EE. Regardless, EE is unique because of how these steps are joined and used together. Additionally, the strengths and weaknesses of EE identified and discussed in Chapter 3 should be considered fully when designing an EE protocol.

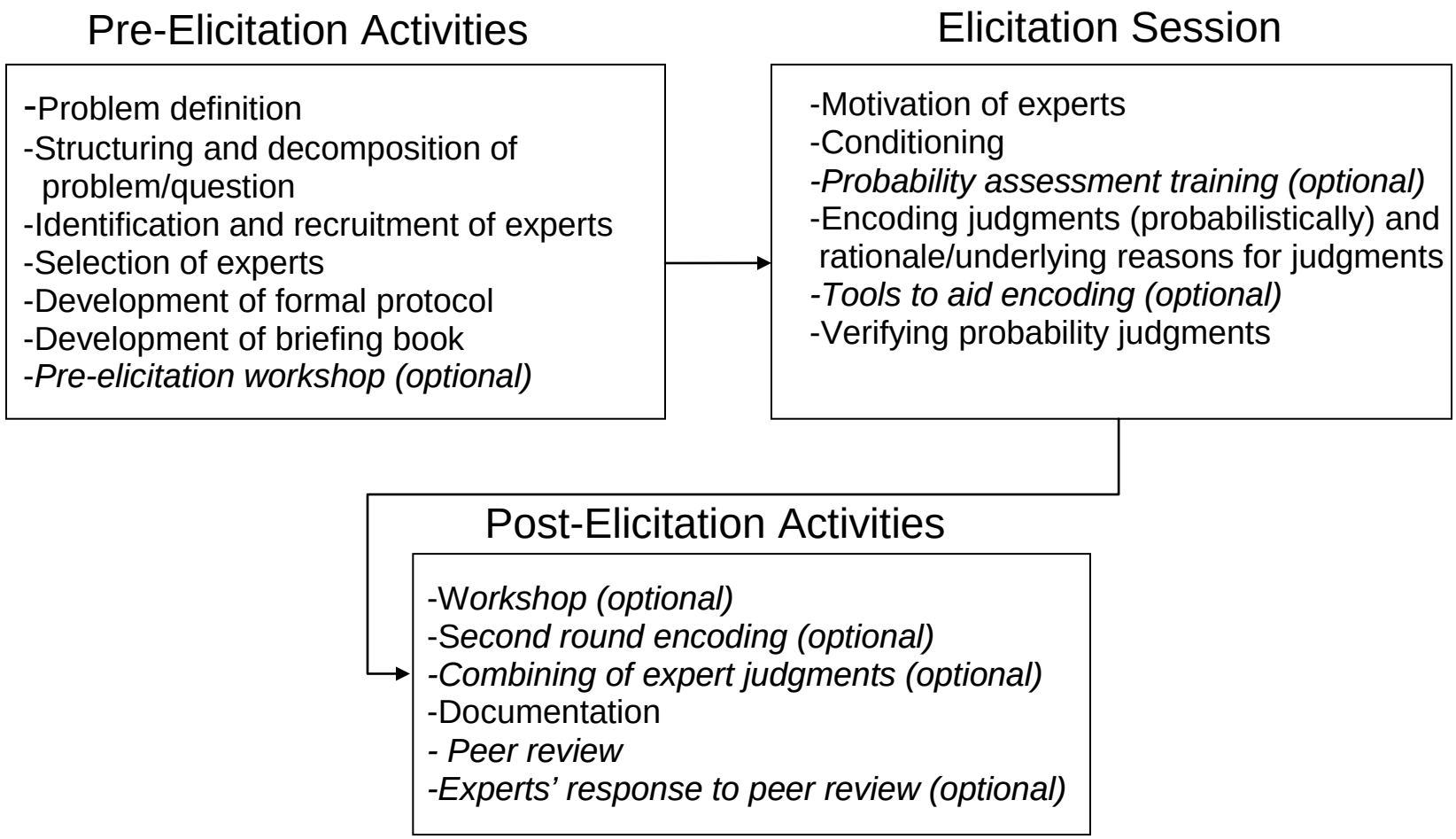


Figure 5-1. Overview of Expert Elicitation Process.

5.2 WHAT ARE THE PRE-ELICITATION ACTIVITIES?

Pre-elicitation activities are shown in Figure 5-1 and discussed in the sections below. Some of these steps can be carried out in parallel; others have prerequisites that require proper sequencing.

5.2.1 What is a Problem Definition?

The initial step of an EE is to craft a problem definition that describes the objectives precisely and explicitly. What is the purpose of the EE, and what are the questions that need to be addressed? Is the purpose of the elicitation to inform a regulatory decision, guide/prioritize research needs, or help characterize uncertainty in a RIA? Laying out the objectives of the elicitation is critical to the design of the EE in guiding the choice of experts, determining how information is presented to them, and determining the form of the judgments that will be required (USNRC, 1996).

Regardless of the purpose of the EE, it is critical to its success that the uncertain quantity of interest is clearly and unambiguously defined. The quantitative question must pass the so-called clarity or clairvoyant test (Morgan and Henrion, 1990). This "test" is fulfilled if a hypothetical clairvoyant could theoretically reveal the value of the uncertain quantity by specifying a single number or distribution without requesting any clarification. Where possible, EE should focus on quantities that are measurable (at least in principle, if not in practice). Although quantities that are "knowable" to a "clairvoyant" are acceptable, the problem must be characterized such that experts can understand how they are to apply their knowledge and experience. This demands that all of the significant assumptions and conditions that could impact the expert's response are well-specified. The focus on "measurable" quantities (at least in principle) also highlights that EE should not focus on societal values and rather be technical in nature.

One also should define the uncertain quantity in such a way that the expert is able to apply his knowledge as directly and fully as possible without necessitating mental gymnastics (Morgan and Henrion, 1990). Because values often are elicited for quantities that are part of a larger system, the EE should be explicit about the values of other quantities that are relevant to the quantity being elicited. This is important for two reasons. First, an expert's judgment about the value of a dependent quantity may depend on whether other input quantities are fixed or are also uncertain. If the input quantities are fixed, the mental challenge for the experts is less; but, if the input quantities are uncertain, the expert must use a second level of judgment to incorporate the effect of this uncertainty on the input value of the elicited quantity. Hence, for

input quantities that are uncertain, it may be useful to guide the expert by considering how the elicited quantity might vary over a mutually agreeable set of discrete values for the input quantities. Second, when multiple quantities are elicited, there may be a complex web of dependencies. For example, multiple elicited values may depend on a third value of some common factor. Also, two elicited values may be mutually dependent. In the complex systems that are often the subject of EEs, these types of relationships are difficult to avoid. Hence, it may be helpful to present their dependency graphically using an influence diagram (see Section 6.5.3.1).

In addition, it is important that the uncertain quantity be specified in such a way that it adequately addresses the policy or analytical question(s) of interest. For example, in the chronic ozone lung injury elicitation, as briefly described in Section 2.5.1.3 (Winkler et al., 1995), the definition of mild lesions in the centriacinar region of the lung were those that could be detected by sophisticated measurement methods such as an electron microscope, whereas moderate lesions were those that could be detected with the naked eye. These definitions were readily understood by the toxicology experts that participated in this elicitation.

5.2.2 How is a Problem Structured and Decomposed?

The problem definition can be designed with either an “aggregated” or “disaggregated” approach. The choice of one of these approaches will be influenced by the type of experts available, their perceived breadth of knowledge, and their ability to use integrative analysis. For the two approaches, questions will be structured and posed differently. For the aggregated approach, the uncertain relationship of interest will be obtained through a single complex question. For example, if the quantity of interest is the probability that exposure to chemical *x* at concentration *y* will lead to a 10 percent increase in mortality, one could ask this question directly to experts. Alternatively, if following a disaggregated approach, there will be a series of simpler, more granular questions to the experts. Multiple types of experts may be used so that each can be asked a specialized question for their expertise. For example, dosimetry experts would be asked the probability that exposure to chemical *x* at concentration *y* will result in a given internal dose. Then, health scientists would be asked to provide the probability that a specified internal dose of chemical *x* will result in a 10 percent increase in mortality. In the first (aggregated) approach the question integrates multiple processes. In the latter (disaggregated) approach, the larger question is broken down to more elemental questions.

The degree of aggregation/disaggegation is important for many reasons. Breaking down a complex model into its component parts and eliciting parameter values may be infeasible because of the complexity. A clairvoyant would know the actual values of all these factors, but

an expert cannot. Furthermore, the degree of dependence among these factors may be difficult to know or comprehend. Using independent marginal distributions (ignoring correlation) for multiple uncertain parameters in a model can produce misleading outputs. Whatever the degree of (dis)aggregation used, one must be clear about which other factors are controlled and their judgments are conditional on these. Disaggregation can also be problematic when experts are unfamiliar with or even reject the specific model being used.

Greater aggregation may be warranted when the relationship between parameter values and outcomes or model outcomes as a reflection of the natural observed variation is measurable. To illustrate, consider the elicitation of an expert's judgment about the maximum hourly ozone concentration in Los Angeles the following summer. Maximum hourly ozone depends on temperature, wind speed and direction, precipitation, motor-vehicle emissions, and other factors. Depending on the purpose of the elicitation, the distribution of some of these may be specified. A clairvoyant would know the actual values of all these factors, but the expert cannot. One could look at historical trends or natural variations, however, and develop some estimate or prediction along with some description of the major assumptions or factors considered in the elicitation.

Using a highly aggregated approach can minimize the potential for "probabilistic inversion." Consider the following example from Jones et al. (2001). The spread of a plume from a point source is often modeled as a power-law function of distance [$\sigma(x) = P \times Q$], where σ is the lateral plume spread, and x is downwind distance from the source. P and Q are parameters whose values depend on atmospheric stability at the time of release. This model is not derived from physical laws but provides a useful description when parameters are estimated using results of tracer experiments. Experts have experience with values of $\sigma(x)$ measured in radioactive tracer experiments and values of lateral spread at multiple distances from the source can be elicited. The problem of "probabilistic inversion" (i.e., identifying probability distributions on P and Q that, when propagated through the model, produce the elicited distributions for lateral spread), however, is difficult; indeed, there may not be any solution or the solution may not be unique (Jones et al., 2001; Cooke and Kraan, 2000). It is unreasonable to expect an expert to be able to perform this probabilistic inversion in the context of an EE, and one might be better served to elicit estimated concentration based on observed variation instead.

In general, analysts try to present the questions at a level of aggregation (or disaggregation) for which the experts are familiar and comfortable. One might expect that decomposing complex processes to obtain uncertain quantities would aid experts in making judgments on more familiar quantities. Morgan and Henrion (1990), however, report that results are mixed as to whether decomposition actually improves outcomes. Thus, the extent to which a

particular elicitation uses a more or less aggregated approach is a matter of professional judgment. Early interaction between the analysts structuring the assessment and substantive experts can help to guide decisions on the extent of aggregation that is most appropriate for a given elicitation. Additional research in this area could help inform and guide analysts on the appropriate level of aggregation for EE projects.

5.2.3 What are the Staffing Needs?

An EE project typically requires three types of staffing. First, it needs *generalists* who are familiar with the overall problem(s) or question(s) of interest and who are responsible for the general management of the EE. The second staffing need is for *analysts* or *facilitators* (often called "*normative experts*") who are proficient in the design and conduct of EEs. These normative experts have training in probability theory, psychology, and decision analysis and are knowledgeable about the cognitive and motivational biases discussed in Chapter 3. They are proficient with methods to minimize these biases in an effort to obtain accurate expert judgments. The third staffing need is for substantive *domain* or *subject matter* experts who are knowledgeable about the uncertain quantity of interest and any relevant theories and processes.

5.2.4 How are Experts Selected?

A critical element of an EE is the selection and invitation of the subject experts who will provide the required probabilistic and other judgments. The process for selecting these experts should ensure that the panel of experts will have all appropriate expertise to address the questions and represent a broad and balanced range of scientific opinions. In practice, EPA may use a contractor to establish, control, and manage the EE process, including the selection of experts. Using a contractor in this way may prevent the group of experts from being subject to FACA.

Previous EE studies (e.g., Hawkins and Graham, 1988) have identified several additional criteria for the expert selection process, including: explicit and reproducible, reasonably cost-effective, and straightforward to execute. Transparency in the selection process is essential for an EE that will be used as part of a government science or regulatory activity.

5.2.4.1 What criteria can be used to select experts?

The selection of experts for an EE should be based on criteria that help to identify and choose experts who span the range of credible views. Because multiple disciplines may bring credible expertise to the EE, the selection criteria should seek to ensure that those disciplines are represented equitably (i.e., technical balance). Keeney and von Winterfeldt (1991) indicate that when an EE is intended to be an input to a highly influential assessment, the substantive experts "should be at the forefront of knowledge in their field, and they should be recognized as leaders

by their peers." As cautioned in Chapter 3 by Shackle (1972b), EE study designers should be mindful that, although it is ideal to obtain experts representing the entire range of opinions in their field, this might be difficult to achieve in practice.

Whether the EE is established, controlled, and managed by EPA staff or contractors, it is necessary to select experts who are free of "conflicts of interest" and "appearance of a lack of impartiality." A "conflict of interest" (18 USC ∫208) is concerned with matters of financial interest and occurs when there is nexus between a person's private interests and public duties. An "appearance of a lack of impartiality" (5 CFR 2635) can be financial of not financial in nature.[21]

The selection criteria for EE experts may also vary depending on the objectives of the assessment. For example, is the goal of the assessment to characterize the range of credible views or to obtain a central tendency estimate?

In general, EE practitioners agree that transparency in the expert selection process helps to assemble a group with an appropriate range of expertise and perspectives. Technical expertise and no conflict of interest or appearance of a lack of impartiality are critical characteristics for selecting experts. Although persons who are familiar with and have a substantial reputation in the field may be called upon to be experts, it is important to maintain balance with new people who bring fresh perspectives.

There is usually a continuum of views on any issue. To the extent practicable, selected experts should have technically legitimate points of view that fall along the continuum. The group of reviewers should be sufficiently broad and diverse to fairly represent the relevant scientific and technical perspectives and fields of knowledge. For some EEs, it may be important that the selection criteria reflect a balance among institutional affiliations.

The USNRC in its Branch Technical Position (NUREG-1563) (USNRC, 1996) has developed guidance for selecting an appropriate set of experts. It states that a panel selected to participate in an EE should include individuals who: "(a) possess the necessary knowledge and expertise; (b) have demonstrated their ability to apply their knowledge and expertise; (c) represent a broad diversity of independent opinion and approaches for addressing the topic; (d) are willing to be identified publicly with their judgments; and (e) are willing to identify, for the record, any potential conflicts of interest" (USNRC, 1996). Much of this guidance resembles that contained in EPA's *Peer Review Handbook* (USEPA, 2006a). It is common practice for an

[21] Additional details on "conflicts of interest" and "appearance of a lack of impartiality" are available in EPA's *Peer Review Handbook* (2006) and its Addendum (2009). Both are available at http://www.epa.gov/peerreview.

EE to report the names and institutional affiliation of the participating experts. It is not always desirable, or necessary, however, to attribute each particular elicited judgment to its respective expert. The issue of anonymity with respect to individual EE judgments is discussed further in Section 5.4.4.

5.2.4.2 What Approaches are Available for Nominating and Selecting Experts?

A number of approaches have been cited in the literature for nominating and selecting the substantive experts. Some approaches use literature counts as a rough measure of expertise. Others use participation on relevant NAS or SAB committees as a proxy for expertise. Another approach is to ask scientists who have published in the area of interest to recommend experts for participation. It may also be helpful to ask professional and academic societies, or other institutions that do not have a financial or other interest in the EE's outcome, to submit nominations. In practice, it is possible to use a combination of these approaches. For example, in the pilot $PM_{2.5}$ mortality EE, literature counts were used to identify which individuals should be asked to nominate experts for potential participation. In the full PM mortality EE, this process was modified further to include nominations from the nonprofit Health Effects Institute. This allowed the pool of experts to include additional expertise in toxicology and human clinical studies that were not represented adequately by the initial approach. Having a carefully designed approach for nominating and selecting experts is advantageous because the entire process is transparent and can be reproduced. This is desirable when there is a need to augment or replace experts in an assessment or to replicate the process for another study.

A currently unresolved issue among EE practitioners is whether the sponsor of the EE (e.g., EPA) should be directly involved in the nomination and selection of experts or should allow a third party to conduct this critical element of the EE process. This challenge is also relevant for many EPA peer reviews and can have implications on whether the group of peer reviewers constitutes a federal advisory committee under FACA. On one hand, the sponsor may wish to exert direct control over the quality and credibility of the process. On the other hand, minimizing involvement by the sponsor has the benefit of greater objectivity. Hawkins and Graham (1988) advocate that the selection process should minimize the level of control of the researcher who is conducting the elicitation.

For highly influential EEs that are likely to attract controversy, it may be worthwhile to adopt a process or take additional steps to help establish that the selection of experts was done carefully to represent the range of credible viewpoints. One possible approach is to use a transparent process with public input similar to the nomination process used to form new SAB panels. Although this approach is more time consuming, the SAB allows outside groups to participate in the expert selection process by soliciting and considering nominations from the

public. Information about the design of peer review panels and selection of participants is presented EPA's *Peer Review Handbook* (USEPA, 2006a).

5.2.4.3 How Many Experts are Necessary?

The number of experts involved in an EE is determined primarily by the complexity of the EE, time and financial constraints, availability of credible experts, and the range of institutional affiliations or perspectives that is sought. There have been only limited efforts to develop mathematical theory for optimizing the number of experts used in studies (Hogarth, 1978; Clemen and Winkler, 1985; Hora, 2004).

It is possible to obtain an estimate of the number of experts that are needed by examining the number of experts that have been used in past EEs. A recent informal survey (Walker, 2004) based on 38 studies found that almost 90 percent of the studies employed 11 or fewer experts. Nearly 60 percent of the studies relied on 6 to 8 experts, and the largest number of experts used in any of these studies was 24. This survey is not intended to be representative of all EE studies but provides some insight. Of the 38 studies in this survey, 27 were from a database provided by Roger Cooke from his work while at the University of Delft, Netherlands (Cooke and Goossens, 2008). The remaining 11 studies were obtained from a literature search. All of the studies elicited probability distributions or confidence intervals to describe uncertainty.

Clemen and Winkler (1985) argue that there can be diminishing marginal returns for including additional experts in an EE assessment. Their observations are based on a number of theoretical examples. The simplest example evaluated the impact of dependence between experts on the equivalent numbers of experts to achieve a particular level of precision in an estimate. Their findings show that when the experts are completely independent, $\rho = 0$, the number of equivalent and actual experts are the same. The practical implication is that, the more different the experts are, the more experts are needed.[22] As dependence among the experts increases, the value of additional experts drops off markedly.

Clemen (1989) describes Hogarth (1978) as using test theory as a basis for discussing the selection of experts. He concluded that between six and 20 different forecasters should be consulted. Furthermore, the more the forecasters differed, the more experts that should be included in the combination. Libby and Blashfield (1978), though, reported that the majority of the improvement in accuracy was achieved with the combination of the first two or three forecasts. Steve Hora has argued often that "three and seldom more than six" experts are

[22] The value of ρ is the correlation and can range from 0 (indicating there is no correlation and, thus in this case, the experts are completely independent) to 1 (indicating that the experts are completely correlated or dependent).

sufficient. Clemen and Winkler (1985) suggest that five experts are usually sufficient to cover most of the expertise and breadth of opinion on a given issue.

It may be necessary to include more experts in the process if an EE seeks to not only characterize the range of judgments but also provide an estimate of the central tendency among the overall scientific community. In addition, it may be necessary to develop additional procedures to address questions about the representativeness of the group of experts. One suggestion has been to precede or follow an EE with a survey of the broader scientific community that is knowledgeable about the issue of concern, combined with appropriate statistical techniques such as factor analysis. This will allow the analyst to compare the judgments from the EE experts with the views of this broader scientific community. To more fully develop and demonstrate this approach requires further research, development, and evaluation.

The requirements of the Paperwork Reduction Act (PRA) are an additional consideration for EEs conducted or sponsored by EPA or another federal government agency. OMB has indicated that EE activities are subject to the requirements of the PRA. The PRA provides that if more than nine persons external to the sponsoring agency participate in a survey, that agency generally must submit an information collection request to OMB under the PRA. This effort can add substantial amounts of time and cost to the completion of the EE. The administrative requirements of the PRA may by themselves be reason enough for EPA to limit the number of experts involved in an EE. Where it appears the PRA may apply to an EE, EPA staff are encouraged to consult with the PRA legal experts in the Office of General Counsel's (OGC) Cross-Cutting Issues Law Office about the PRA to better clarify whether and how its requirements apply to EEs and potentially minimize the impacts of the PRA on the EE process.

5.2.5 What is an EE Protocol?

The development of an EE protocol is one of the most resource intensive steps in the conduct of an EE. This step is particularly demanding when dealing with a complicated issue that is informed by different perspectives and disciplines. An EE protocol serves several purposes and includes the following:

1. Overall issue of interest and any relevant background information;
2. Motivation or purpose for the assessment and, at least in a general sense, the role of the EE in any larger modeling or decision process;
3. The quantitative question of interest and definition of any conditions or assumptions that the experts should keep in mind; and

4. Information about heuristics and biases that are characteristic of EEs (see Section 3.5.5) and guidance on how to minimize these problems.[23]

As an example, the protocol that was developed recently for EPA's pilot $PM_{2.5}$ mortality EE project is available from IEC (IEC, 2004).

5.2.6 Why Should Workshops Be Considered?

Pre-elicitation workshops that bring the project team and experts together are a helpful, though not required, element of many EEs. There are three major reasons why holding one or more pre-elicitation workshops is advisable. First, these pre-elicitation workshops can be used to share information on the technical topics of the EE. This will help to assure that all of the experts are familiar with the relevant literature, different perspectives about the research, and how these views might relate to the EE's questions. This can alleviate some criticism regarding the importance of a common body of knowledge (discussed in Chapter 3) and could ultimately reduce the need for more experts in future EEs on similar topics.

Second, feedback on the draft protocol that is obtained at a pre-elicitation workshop can be used to refine or restructure the EE's question. Finally, the pre-elicitation workshop can be used to introduce the concepts of judgmental probability, heuristics, and biases to the experts. The workshop provides an opportunity to train the substantive experts in the techniques used to elicit probability judgments. During the training, the project team and experts can practice the use of techniques to reduce bias (USNRC, 1996). This can address the findings (as discussed in Chapter 3) that experts who do not understand these techniques tend to give more haphazard responses (Hamm, 1991). Training is further discussed below in Section 5.2.8.

5.2.7 What Is a Briefing Book and What Is Its Role in an Expert Elicitation?

Another important pre-elicitation step is the development of a "briefing book." The briefing book is a binder (and/or contained on electronic media) with journal articles and other technical information relevant to the topic of the EE. To promote balance, this briefing book should include representative papers from all technically legitimate perspectives. According to NUREG-1563, this background material should be selected so that "a full range of views is represented and the necessary data and information are provided in a uniform, balanced, and timely fashion to all subject-matter experts" (USNRC, 1996). The experts should be given the opportunity to add pertinent information to the briefing book, including unpublished data that they are willing to share.

---

[23] Appendix A to this White Paper, "Factors to Consider When Making Probability Judgments," is an example of the type of material that should be included either in the protocol or briefing book.

5.2.8 What Type of Training Should Be Conducted?

Pre-elicitation training can facilitate EEs in several ways. According to USNRC recommendations, training subject matter experts prior to elicitation has the following benefits: "(a) familiarize them with the subject matter (including the necessary background information on why the elicitation is being performed and how the results will be used); (b) familiarize them with the elicitation process; (c) educate them in both uncertainty and probability encoding and the expression of their judgments using subjective probability; (d) provide them practice in formally articulating their judgments as well as explicitly identifying their associated judgments and rationale; and (e) educate them with regard to possible biases that could be present and influence their judgments" (USNRC, 1996). Training helps to ensure that the expert judgments accurately represent the experts' states of knowledge about the problem of interest. In addition, this training provides an opportunity to level the knowledge base among the experts and can help to clarify the problem definition.

5.2.9 What is the Value of Pilot Testing?

After the draft protocol and briefing book are complete, a pilot test can provide valuable feedback on the quality of the protocol and help identify any obstacles. The objective of this step is to improve the clarity of the protocol and determine whether the questions are framed appropriately. Ideally, pilot testing should be conducted with substantive experts who are not among the pool of experts that will participate in the actual EE. Pilot testing can include several experts, but it is essential to pilot test the draft protocol with at least one person.

5.2.10 How Should an EE Be Documented?

It is absolutely critical that all significant aspects of the steps listed above be documented clearly. This documentation is intended to chronicle the process as it occurred and does not seek to encourage any type of consensus. Clear documentation is essential to establishing the credibility of EE results and to assure a transparent process as stipulated in EPA's Information Quality Guidelines.

5.3 WHAT APPROACHES ARE USED TO CONDUCT EXPERT ELICITATIONS?

Three different approaches for conducting EEs have been demonstrated and documented (see pp.141–154 of Morgan and Henrion [1990] for a more detailed description). These approaches for eliciting expert judgment, often referred to as "probability encoding," include: (1) the approach used by Wallsten and Whitfield in EEs carried out for EPA's OAQPS (see Wallsten and Whitfield [1986] for an example); (2) the approach used by Stanford/SRI, pioneered by Howard, North, and Merkhoffer, and described in Spetzler and Staehl von Holstein

(1975); and (3) the approach used by Morgan and his colleagues at Carnegie Mellon University (Morgan et al., 1984). Although these approaches have some case-specific characteristics and other features that differ based on the tools chosen by the analyst, most EE practitioners agree about general principles that constitute good practice for the encoding process.

The encoding process is typically divided into five phases:

- ***Motivating*:** Rapport with the subject is established, and possible motivational biases are explored.
- ***Structuring*:** The structure of the uncertain quantity is defined.
- ***Conditioning*:** The expert is conditioned to think fundamentally about judgments and to avoid cognitive bias.
- ***Encoding*:** This judgment is quantified probabilistically.
- ***Verifying*:** The responses obtained from the encoding session are checked for internal consistency.

The encoding session is conducted in a private setting (e.g., typically the expert's office) so that the subject is comfortable and the discussion can be uninterrupted and candid. As discussed in Section 5.2.5, the EE protocol is used to guide the encoding session so that the topics covered and responses to experts' questions asked are treated consistently among the several experts. Responses and other feedback from the subject matter experts are documented thoroughly with one or more of the following: written notes, transcripts, and audio or video tape.

5.3.1 What are the Staffing Requirements for the Encoding Session?

In general, a minimum of two individuals are required to conduct the encoding session. These usually include at least one subject matter expert and one analyst (see Section 5.2.3 for a description of the roles of these individuals). In addition, the project team requires a generalist who will be responsible for the general management of the EE.

5.3.2 What Methods and Tools Are Available To Aid an Elicitation?

A variety of innovative methods and tools to aid the elicitation of probabilities during the encoding session have been developed and used by Morgan and Henrion (1990), Spetzler and Staehl von Holstein (1975), Wallsten and Whitfield (1986), and Hamm (1991). These methods include: (1) fixed values, (2) fixed probabilities, (3) probability wheel, and (4) specialized software or models for incorporating and/or displaying judgments for feedback to experts (Morgan and Henrion, 1990; Hamm, 1991).

- ***Fixed value methods*:** In this method, the probability that the quantity of interest lies within a specified range of values is assessed. Generally, this involves dividing up the range of the variable into equal intervals.
- ***Fixed probability methods*:** In this method, the values of the quantity that bound specified fractiles or confidence intervals are assessed. Typically, the fractiles that are assessed include the median (0.5), quartiles (0.25, 0.75), octiles (0.125, 0.875) and extremes such as (0.1, 0.99).
- ***Specialized software or models*: S**oftware (e.g., Analytica®) have been developed to facilitate the elicitation of probability judgments. These software tools use graphics to illustrate the implications expert judgments. In addition, software (e.g., Excalibur) have been developed to address the calibration or performance assessment of experts.
- ***Probability wheel and other physical aids:*** The probability wheel (Figure 5-2) is among the most popular physical aids used for EEs. This colored wheel (e.g., orange) supports the visualization of probabilities by using an adjustable pie shaped wedge (e.g., blue) and a spinner with a pointer. To use the probability wheel, the expert varies the pie shaped wedge, changing the blue/orange proportions of the wheel until the probability that the spinner will end up on blue equals the probability of occurrence for the event of interest. The back of the probability wheel shows the fractional proportions of the two colors. The analyst uses these numbers for the expert's response. See Morgan and Henrion (1990, pp. 126–127).

Figure 5-2. Probability Wheel

5.3.3 What Group Processes Can Be Used for Expert Elicitation?

Many methods exist for assembling groups and obtaining judgments. These methods vary in how they encourage interaction, allow iteration, and seek consensus. In addition, the application of each method depends on the number of group members, the time available, the financial resources, and the context of the participation. Obtaining collective judgments using any of these methods may trigger FACA requirements. The Delphi method and Q-Methodology are described herein. Detailed discussions of many other methods are available elsewhere.[24]

5.3.3.1 Delphi

The Delphi method is a common technique for assembling a group, in a manner similar to focus groups, to obtain their judgments. One distinguishing feature of the Delphi method is that its members generally do not meet as a group. In a study by the Delphi method, participants are selected because they have expertise about a specific domain of interest. In most cases, correspondence is remote (i.e., mail, e-mail, fax, or telephone); however, face-to-face interviews also can be used. The initial interview session with each participant is conducted by a facilitator. The facilitator serves as a clearinghouse of the panelists' responses. For subsequent iterations of the interview session, each participant sees and reacts to the anonymous views expressed by the other participants. Through a series of iterations of the interview, the panelists share and generate new ideas with the objective that consensus will emerge. Through this process, the Delphi method generates ideas and facilitates consensus among participants despite the fact that they are not in direct contact with each other.

The advantages of the Delphi method are that it encourages the sharing of ideas and promotes consensus with a large number of stakeholders who may be geographically distant from each other while minimizing bias that results from circumstantial coercion (e.g., rank or seniority) or group think. It is a transparent and democratic technique that may be appropriate for highly technical issues in which the goal is to obtain a consensus judgment. Its shortcomings are that: (1) it may be resource intensive (time and money); (2) it can require large amounts of data to be assessed and distributed; (3) by emphasizing consensus, its final judgment may not characterize the full range of uncertainty; and (4) if participants are too rigid or their commitment wanes, it may be impossible to obtain consensus judgments.

---

[24]For example, see SMARTe's (Sustainable Management Approaches and Revitalization -electronic) Tool for Public Participation (Go to: http://www.smarte.org → select: "Tools" → select: "Public Participation"), or: http://www.smarte.org/smarte/tools/PublicParticipation/index.xml?mode=ui&topic=publicinvolvementaction

5.3.3.2 Q-Methodology

The Q-Methodology is a social science technique that was invented in 1935 by William Stephenson, a British physicist-psychologist (Stephenson 1935a; 1935b; 1953). It provides a quantitative basis for determining the subjective framing of an issue and identifies the statements that are most important to each discourse. This method provides a quantitative analysis of subjectivity (i.e., "subjective structures, attitudes, and perspectives from the standpoint of the person or persons being observed") (Brown, 1980, 1996; McKeown and Thomas, 1988).

In Q-Methodology, participants map their individual subjective preferences by rank-ordering a set of statements (typically on a Likert-like scale with endpoints usually representing "most agree" to "most disagree" and zero indicating neutrality). This is accomplished by sorting statements that are printed on cards, using a traditional survey instrument and numerical scale, or, more recently, with an Internet software application. The sorting often is performed according to a predetermined "quasi" normal distribution (i.e., the number of allowable responses for each value in the scale is predetermined with the greatest number of responses in the middle of the scale and fewer responses at either end). The collection of sortings from the participants form a kind of cognitive map, or mental model, of subjective preferences about the particular issue.

Following the sorting phase, participants generally are asked to reflect on their responses. Participants' individual preferences then are correlated against each other and factor analyzed. Factor rotation commonly is conducted either judgmentally, based on theoretical considerations, or by using varimax rotation. The factor outputs, as indicated by individual loadings, represent the degree to which the study participants are similar and dissimilar in their responses. Factor scores represent the degree to which each statement characterizes the factor and can be used to construct a narrative that describes (subjectively) the specific discourse associated with the factor.

5.3.4 How Can Other Media Be Used To Elicit Judgments From Remote Locations?

Though face-to-face interviews often are the preferred method for EEs, constraints on time and money may necessitate conducting the interviews via another medium. Whether questionnaires (by mail or e-mail) (Arnell et al, 2005), telephone, video conference, or some combination (Stiber et al., 2004; Morgan, 2005) are used, the essential elements are consistency and reproducibility. To the degree possible, each expert should be presented with identical information in a standardized order. Answers to questions from the experts should be uniform and ideally should be prepared in advance for anticipated queries. Questions from experts to the interviewers (e.g., requesting clarifications) and responses should be documented and shared

with all experts to ensure each has equal access to and understanding of the information. It is important that the elicitation process produces a common experience for the experts so that their responses reflect a shared understanding of the questions.

Although face-to-face interviews are the standard and preferred approach, there are important advantages to other media. In addition to being less expensive and permitting greater flexibility with schedules, eliciting judgments remotely can engender greater consistency. Because there is less body language (e.g., hand movements and facial expressions), the emphasis is on the content of the interview, and this can be more easily standardized for all of the interviews. In addition, if experts are responding to written surveys, they may feel less pressured, and as a result, may be more thoughtful in their responses. Also, if follow-up questions become necessary, they can be handled via the same format as the original questions (e.g., telephone or written).

5.4 WHAT POSTELICITATION ACTIVITIES SHOULD BE PERFORMED?

5.4.1 How and When Should Final Judgments Be Verified?

As soon as practical after the elicitation, the subject matter experts should be provided with their elicitation results (USNRC, 1996) as well as any clarifications requested by and provided to experts during the elicitations. Then, the analysts can query the experts to ensure that the experts' responses have been represented accurately and even-handedly. It is the job of the project team to determine whether any revision or clarification of the experts' judgments and rationale is needed. Any revisions should be obtained in a manner consistent with the original elicitation and documented carefully. Finally, the experts can confirm concurrence with their final judgments and associated qualitative discussions of rationale.

5.4.2 What Methods Should Be Used To Ensure the Quality of Judgments?

As has been demonstrated by Cooke (1991), it is possible to evaluate empirically the quality of alternative methods for combining distributions when the values of the quantities that are elicited become known. For example, Clemen (2008) and Cooke and Goossens (2008) compared the performance of alternative methods for combining experts' distributions for seed variables. Similarly, one could evaluate the quality of alternative combinations of expert judgments in cases in which the values of the target quantities become known (Hawkins and Evans, 1989; Walker et al., 2003). This practical evaluation can be used to select a method for combining distributions that optimizes the use of quality judgments while rejecting (or reducing the weight of) judgments that were determined empirically to be of lower quality.

5.4.3 When are Postelicitation Workshops and/or Follow-Up Encoding Appropriate?

Conducting a postelicitation workshop is not required for every EE; however, the EE project team should consider and weigh its potential value added against the additional cost (i.e., resources and time). A postelicitation workshop provides an opportunity for all of the subject matter experts to view the judgments of their peers. Typically, the probabilistic judgments and the reasoning behind those judgments are shared. At a workshop, the experts and the project team can probe reasons for differences in judgments. The exchange of views may unearth new insights (i.e., new data, theory, or perspectives) that could influence experts to modify their judgments. This reconsideration and modification of judgments is consistent with the goal of obtaining the most accurate representation of the experts' beliefs based on their understanding of the state of information. If resources or timing preclude an in-person postelicitation workshop, an alternative is to meet via audio or video conference.

A postelicitation workshop in which the experts can reflect on and change their views has two potential additional benefits. First, the experts can change and refine their responses so that the elicitation results more accurately represent their judgments. Second, when movement toward consensus is possible, it becomes more likely that uncertainty can be reduced by more closely representing the collective judgment of the relevant experts for that discipline.

5.4.4 Should Individual Judgments Be Combined, and If So, How?

Many decision-makers want a single unambiguous result, not a complex spectrum of findings. When an EE uses multiple experts, as is typical, the result is many independent sets of judgments, each representing the beliefs of a single expert. These sets of judgments are the "experimental" results of the EE exercise. These data, however, are very different from traditional experimental results. With traditional scientific experiments, if the process that produced the results, including measurement errors and uncertainties, is known, the results may be combined into a single aggregate finding. Handling the results of an EE is more complex. Although each expert was prepared similarly for the elicitation (pre-elicitation workshop), was presented with the same data, and was elicited by essentially the same process (elicitation protocol), the experts differ in their training, experiences, and the manner of considering the relevant information to produce beliefs. Consequently, one should account for both theoretical and practical constraints when considering whether and how to combine multiple expert beliefs. In many cases, combining expert judgments may not be theoretically defensible or practical.

Despite cautions, the practical nature of decision-making may motivate analysts to produce a single aggregate result. This section examines when it is appropriate to aggregate the judgments of multiple experts and describes how this can be done while preserving the richness

and individual nature of the original EE data. The decision to combine expert judgments and the selection of a method for doing so must consider the attributes of the particular elicitation and how the findings will be used. The application of a method to combine the judgments of multiple experts can be project-specific. This section focuses on the mechanics of expert aggregation. See Sections 3.4.3 and 3.4.4 for a more detailed discussion of the theoretical advantages and disadvantages, as well as FACA considerations, of combining expert judgments.

Whether individual judgments should be combined and how it should be done may depend on the objective of the particular study. If the EE was conducted to estimate project costs or to assess the expected value of an action, it might be rational to combine the individual expert's outcomes using a method that is mathematically optimal; however, if the EE's outcomes will be used to inform research investments or to assess health benefits for a proposed regulatory action, it may be better to present the individual results without combining them.

Whether or not judgments are aggregated, there is a question about the desirability of identifying each individual's judgments by name or whether it is sufficient to list the experts who participated and identify the individual judgments via an anonymous lettering system. This issue was raised in the USNRC's Branch Technical Position guidance (USNRC, 1996) and is still debated in the decision analysis community. Cooke (1991) takes the perspective that EE should be consistent with scientific principles and has argued that the goal of accountability requires each judgment to be explicitly associated with a named expert. Others consider EE to be a trans-scientific exercise. From this perspective, preserving anonymity for the specific judgments made by an expert best serves the overarching goal of obtaining the best possible representation of expert judgment. Given the current norms within the scientific community, experts may be unwilling to participate and share their judgments honestly if they fear a need to defend any judgments that divert from the mainstream or conflict with positions taken by their institutions.

5.4.4.1 Why are Judgments Different?

When considering the possible combination of expert judgments, the first step is to ask the question: "Why are judgments different?" Unlike the conventional scientific method in which a technique for combining results can (and should) be selected before data are collected, with an EE it is necessary to see the results before determining if aggregation is appropriate.

Understanding the source of differences between experts can lead to insights, consensus, and/or revision of the elicitation protocol (Morgan and Henrion, 1990). Indeed, this understanding about the source of the differences can be more valuable than any aggregate finding. Furthermore, for many situations, variability among experts is not a problem, but the objective of the elicitation. Many scientific questions are unanswerable but have a spectrum of

legitimate approaches providing different answers. EE often is undertaken to obtain a feel for the range of potential answers. Hence, diversity of judgments may be a good thing, and it would be inappropriate to replace this richness outcome with a crude average.

The experts' judgments may be different for a number of reasons, including: unused information, differences in weighting the importance of information, misunderstandings about the question, different paradigms among experts, and motivational biases. In some cases, it may be possible to re-elicit a mistaken expert (e.g., who did not understand the question or who was unaware of some important information) to rectify the irregularity. If the elicitation process is uniform, there may be less variance in responses.

In other cases, the differences in response may result from different paradigms by which the experts view the world and the data. This often is true when the experts come from different disciplinary backgrounds. Experts tend to trust data obtained through methods with which they have direct experience. For example, when one is trying to estimate the relationship between exposure to a substance and increased morbidity or mortality, epidemiologists may tend to find epidemiological data compelling while being more suspect of toxicological studies on animals. Toxicologists may have the opposite preference. In this situation, the variability among the findings represents a spectrum of beliefs and weights that experts from different fields place on the various types of evidence. In such cases, reconciling the differences may be imprudent.

Although one of the goals in the selection of experts is to obtain an impartial panel of experts, there still may be cases of motivational bias. Identifying such experts and determining how to use their beliefs is best handled on a case-specific basis.

Evaluating the source of differences among experts is intended to produce insights that may obviate any needs to aggregate. These insights may lead to improvements in the elicitation protocol, understanding about varying disciplinary perspectives, or ideas for future research that can (ideally) reduce the interexpert differences. In any case, the knowledge gained from insights may be more valuable than the benefits of a single aggregate finding.

5.4.4.2 Should Judgments Be Aggregated?

The next step is to determine whether judgments should be aggregated, or, in less normative terms, if it is appropriate to aggregate. In many situations, part of the answer to this question depends on the relative value of the uncertainty of each individual's judgments with respect to the difference between the individual judgments. Aggregation may be appropriate if the interindividual variability is less than each individual's uncertainty (see Figure 5-3). In this case, knowledge sharing may result in convergence because aggregation may average out the "noise" in the characterization of the different judgments.

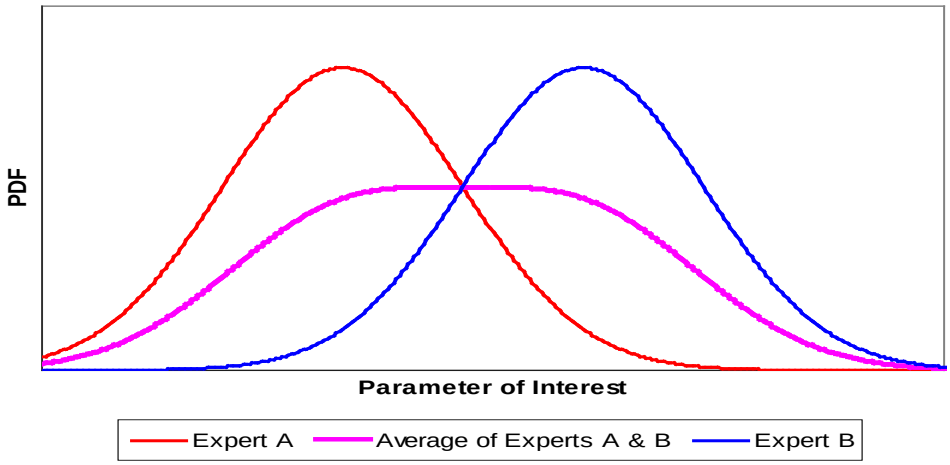


Figure 5-3. Experts with Similar Paradigm but Different Central Tendency

It may be inappropriate to aggregate when the inverse is true; the interindividual variability is greater than each individual's uncertainty (see Figure 5-4). If it appears that the experts' judgments are based on fundamentally different paradigms, aggregation may create an average set of judgments that lacks any phenomenological meaning. The dissonance between the experts itself may be the insight that the EE provides. The existence of heterogeneity is itself important and should encourage alternative modes of policy analysis (Keith, 1996). If the decision-maker expects to receive a single value outcome, and it does not appear appropriate based on the considerations of individual uncertainty and interindividual variability, the analyst should explain these limitations to the decision-maker. If it is possible to combine the results to provide a meaningful characterization, these judgments may be aggregated. The analyst, however, should exercise caution and try to share sensitivity analyses with the decision-maker. In some cases, the analyst may want to consult with the experts to better understand the sources for differences of opinion.

In any situation when multiple judgments are combined, sensitivity analyses could be used to examine the effects of each expert on the aggregate outcome. Moreover, the presentation of results should include the individual responses along with the combined response. In general, the decision to aggregate is case-specific and "depends on the individual circumstances and what is meant to be accomplished" (Krayer von Krauss et al., 2004).

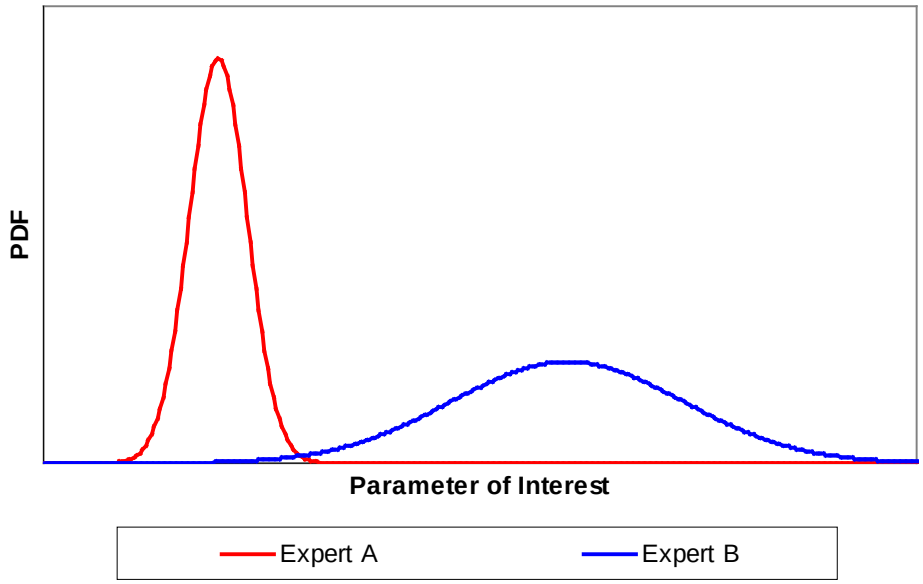


Figure 5-4. Experts with Different Paradigms

5.4.4.3 How Can Beliefs Be Aggregated?

Once the decision has been made to aggregate expert judgments, several methods are available. There are mathematical methods of combination, including simple averages, weighted averages, the Classical Model, and Bayesian approaches (Clemen and Winkler, 1999; Ayyub, 2000, 2001). Simple averages are the easiest to implement and often have good and equitable performance. Weighted averages can vary in sophistication and scheme for developing weighting factors. In the Classical Model *calibration* and *information* are used for a statistical test for a performance-based method for combination (Bedford and Cooke, 2001; Clemen, 2008). Bayesian approaches are discussed in numerous references (Genest and Zidek, 1986; Jouini and Clemen, 1996; Stiber et al., 2004).

Behavioral methods offer a means by which group interaction can produce consensus or at least generate a better understanding of differences between experts and their sources. Included in these behavioral methods are the Delphi method (Dalkey, 1969; Linstone & Turoff, 1975; Parenté and Anderson-Parenté, 1987) and the Nominal Group Technique (Delbecq et al., 1975; Kaplan, 1990). Although a group can lead to a consensus among experts, this is not always possible or desirable. Interexpert variability is a source of richness in EE studies and often leads to insights. The analyst should not seek to remove these differences without first understanding them.

Another important aspect to consider is: At what level of analysis should the combination occur? Should the judgments themselves be combined prior to any calculations or model simulations, or should each expert's judgments be used to construct a model and then combine the outputs of each expert's model? In many cases, it would seem that the most logical and phenomenologically rational approach is to combine the individual judgments. These quantities represent the direct beliefs of the experts and were the quantities sought through the EE process. If the experts were aware of and understood the model that was used to run their judgments, however, their responses may have incorporated knowledge and expectations about the model. In such a situation, it may be appropriate to combine the model outputs rather than the individual experts' responses.

5.5 WHEN AND WHAT TYPE OF PEER REVIEW IS NEEDED FOR REVIEW OF EXPERT ELICITATION?

The purpose of conducting a peer review of an EE project is to evaluate whether the elicitation and other process elements were conducted in a professional and objective manner and whether the protocol was adequate and was followed. The judgments provided by the experts are not subject to evaluation by peer review. The mechanism for peer review should be selected in consultation with EPA's *Peer Review Handbook* (2006a) and in consideration of the intended use of the EE results. In some circumstances, it may also be appropriate to conduct a peer review to provide advice on how the results of an EE should be considered relative to other analyses or scientific assessments in a given regulatory context.

5.6. SUMMARY

This chapter discussed important factors to consider when conducting an "acceptable" EE. It presented "good" practices for conducting an EE, whether it is conducted (or sponsored) by EPA or by outside parties for submission to EPA. The discussion of "good" or "acceptable" practice was based on a review of the literature and actual experience within EPA and other federal agencies. In general, the degree to which practices are "good" or "acceptable" depends substantively on the following: (1) clear problem definition, (2) appropriate structuring of the problem, (3) appropriate staffing to conduct EE and selection of experts, (4) protocol development and training, including the consideration of group processes and methods to combine judgment, if appropriate, (5) procedures to verify expert judgments, (6) clear and transparent documentation, and (7) appropriate peer review for the situation. Although this White Paper presents what the EE Task Force believes can constitute "good" practice for EPA EEs, the Task Force recognizes that a range of approaches currently are used among EE practitioners. Hence, the *design of a particular EE* involves considerable professional judgment.

6.0 HOW SHOULD RESULTS BE PRESENTED AND USED?

6.1 DOES THE PRESENTATION OF RESULTS MATTER?

The presentation of results not only provides the findings of an EE, it also serves as a window to the methods, assumptions, and context of the EE itself. For many stakeholders, viewing the results will be their first impression of the EE. Hence, it is important that results are presented thoughtfully and with attention to the particular needs of the intended audience. At EPA, it is expected that EE results will be used to support regulatory decision-making. In this context, the decision-maker will be one of the primary recipients of EE findings. The presentation of results will inform decision-makers about EE findings and help them to consider the EE results along with many other sources of information. To this end, results should be presented in ways that will enable their understanding and promote their appropriate use. In addition, regulatory decisions employ a transparent process; therefore, EE results also should be presented in a manner that will enhance their understanding by members of the public.

6.2 WHAT IS THE STAKEHOLDER AND PARTNER COMMUNICATION PROCESS?

Communicating about an EE is a two-way process involving the transfer of information between parties with different professional training, data needs, motivations, and paradigms for interpreting results. Hence, EE analysts should consider how to present results for each intended audience. Table 6-1 summarizes potential stakeholders for EPA EEs. Presentation of EE results should consider the stakeholder perspectives, knowledge of the EE subject matter, and the context in which the EE is being developed and used.

Table 6-1. List of Stakeholders and Partners for EPA Expert Elicitations

EPA risk managers/assessors
Members of the public
State and local environmental and health agencies
Federal agencies (e.g., Health and Human Services, U.S. Geological Survey, Department of the Interior)
The Office of Management and Budget
Tribal governments
Regulated community
Scientific community
Managers of federal facilities (e.g., department of Defense, Department of Energy)
(After: USEPA, 2001a)

Communication of technical information to the public involves the establishment of "trust and credibility" between the Agency and the Stakeholders. Approaches to establishing trust and credibility include continued community involvement (USEPA, 1994) and providing information at the appropriate level to aide an understanding of results (USEPA, 2002). A wide range of literature is available regarding stakeholder perceptions of risk (Slovic et al., 1979) and sponsors of an EE need to consider these perceptions in developing communication strategies for EE results. For example, it may be helpful to discuss the EE results within the context of other studies or regulation decisions.

EPA has developed several guidance documents on public involvement and communication that may be considered when developing EE communication strategies (USEPA, 1995, 1997, 2000a) along with the large research literature (Covello, 1987; Deisler, 1988; Fischhoff, 1994, 1995, 1998; Hora, 1992; Ibrekk and Morgan 1987; Johnson and Slovic 1995; Kaplan 1992; Morgan et al., 1992; Ohanian et al., 1997; Thompson and Bloom, 2000). Depending on the complexity of information presented in the EE, it may be necessary to test and review communication materials to improve the clarity and effectiveness of presentation.

6.3 HOW CAN COMMUNICATIONS BE STAKEHOLDER-SPECIFIC?

Each stakeholder has different needs that should influence the content and form of communications products. The types of communication and information products that would be appropriate for the major types of stakeholders are listed below.

Risk managers and state/federal officials may have technical and/or policy training but probably have a limited understanding of EE. Useful products for this audience include executive summaries, bulleted slides, and briefing packages with a short description of the EE process. One may consider placing the results of the EE in the broader context of the decision when developing these communication materials. This may include an overview of the reasons and importance of the decision, previous decisions, and the positions of major stakeholders. The presentation also should include an appropriate discussion of the uncertainties and their potential impact on the decision (Thompson and Bloom, 2000).

Researchers with technical domain knowledge and/or expertise in EE will be the most literate about EEs. This audience typically has a more in-depth scientific knowledge of the EE process and related issues. Useful products include technical reports, peer-reviewed scientific papers, and presentations at professional meetings.

Community members generally have a limited knowledge of EE and may need more background and discussion of the EE process. Documents may include fact sheets, press releases, and presentations that summarize the key issues and conclusions of the EE in a context

and language(s) that is appropriate for the community. Synopsis and simplification do not mean simplistic products. Consideration of the user's knowledge base is important.

6.4 WHAT IS IN A TECHNICAL SUPPORT DOCUMENT?

The Technical Support Document (TSD) provides the basis for development of all information products. The TSD contains all relevant information for the EE, including background, methods, data, analysis, and conclusions. Appendices to the TSD may include the EE protocol, prepared questions and answers for the interviews, list of experts with their affiliation, and other information documenting the process.

The following sections provide a checklist of suggested topics for the TSD of an EE. It covers the introductory information that should be included in documentation (Section 6.4.1), the technical details of the EE (Section 6.4.2), and finally, examples of means to summarize the results of the EE (Section 6.5). This template for information that should be covered does not preclude other relevant requirements for data quality or peer review. Most of the items in this checklist are self-explanatory and do not require a detailed description. Descriptions are only included when they may be helpful.

6.4.1 What is in the Introduction of a Technical Support Document?

The Introduction of a TSD should include:

- The data or information gap(s) that this EE addresses. The quantities that are the subject of the elicitation may be poorly defined or variously conceptualized (Hora, 2004) and should be clarified. A well-crafted objective statement should specify the problem that the EE addresses, define the meaning of elicited values, and identify the intended audiences of the findings.
- A brief summary of what EE is and what it is not (especially compared to "expert judgment" and "peer review").
- The rationale for using EE in the context of the larger policy or research question. Describe the policy/research question to which these elicitation results will be applied.
- How the methods used in this instance compare to "gold standard" methods.
- What the results of EE mean.
- What are the limitations/cautions of the elicitation?

When the documentation uses words to describe uncertainty, the audience may have varying understandings of what is meant. Hence, the authors need to be sensitive that even

quantitatively presented “probabilities” or “likelihoods” often are misunderstood and can create confusion or misunderstanding for the decision-maker about what presented results may mean. Research by cognitive scientists indicates that a presenter must take care in the introduction of a presentation to define what is meant by “probability” or “likelihood” (Anderson, 1998; Morgan and Henrion, 1990). The TSD should describe what is meant in the EE study by terms like “probability” or “likelihood,” and these meanings should correspond to those held by the experts in the elicitation.

Anderson (1998) provides a useful summary of cognitive research that demonstrates different people may interpret the word “probability” in very different ways. She abbreviates the concepts as in Table 6-2 and asserts that people use different solution algorithms or heuristics to take meaning from a provided probability, depending on which definition of “probability” they are using. Likewise, she stresses that it is important for an audience to know how probability was understood by the experts.

The research on how this affects elicited information implies that the results of the EE must be presented with attention to the format of results and to encourage the audience to use the results correctly. As discussed below, Anderson recommends formatting and presentation of information in light of these challenges.

Table 6-2. Classification of Subjective Concepts Associated with Probability

For most ecologists and statisticians, the word "probability" seems to have a clear meaning. Cognitive scientists, however, recognize that subjective meanings vary depending on context. Teigen (1994) classified several ideas associated with probability and uncertainty. Each of the subjective concepts implies its own calculus of "probability," and each seems to be processed by a different cognitive mechanism. It is important for Bayesian analysts to realize which idea they are activating when they refer to "probability" in a paper or ask an expert for a probability estimate.

Concept	Definition of "Probability"
Chance	The frequency of a particular outcome among all outcomes of a truly random process.
Tendency	The tendency of a particular outcome to occur or how "close" it is to occurring.
Knowledge	It is allocated among the set of known hypotheses.
Confidence	The degree of belief in a particular hypothesis.
Control	The degree of control over particular outcomes.
Plausibility	The believability, quantity, and quality of detail in a narrative or model.

Adapted by Anderson (1998) from Teigen (1994)

6.4.2 What Technical Details of the Expert Elicitation Methods are in the TSD?

The previous section covered what might be included in the introduction for an EE's TSD or other EE communication products. This section addresses what technical details of the EE methods should be included in the body of the document. The documentation should cover:

- The process (protocol) used for the EE and reasons for selection of key elements of the process.
- What criteria were used in selecting the experts (both criteria for individuals such as type of expertise and overall group criteria such as representing the range of credible viewpoints or range of disciplines). It also should identify who selected the experts.
- How well the set of experts selected meets the criteria set forth for the elicitation:
  - Identification of the list of experts, with affiliation and discipline/field, who were selected and agreed to participate and who, if any, did not.
  - Any potential conflicts of interest concerns (or appearances of conflict) and how any were addressed.
- Clear characterization of what experts were asked:

  - Which uncertainties, parameters, relationships, and so forth that the experts addressed.
  - The information elicited from the experts.
  - What data on which the experts may have based their judgments, as well as identified key data gaps. Presenting these alongside the EE results, however, might misleadingly imply an "apples to apples" comparison. The EE may address a broader question or may introduce complexities that cannot be analyzed with available data; or, if the EE includes a wider range of sources of uncertainty, one would expect the uncertainty bounds to be wider.
  - The degree to which the elicited results conform to axioms of probability theory and to the available empirical data.
  - Where biases may have been introduced and, if possible, insights into the likely direction of any biases (individually and overall).
  - How well the extreme values are likely to be represented (especially if there are potential catastrophic or nonlinear effects).
  - Possible correlations with non-elicited components of the overall analysis or policy question.
  - Text or graphics (e.g., influence diagrams or frequency-solution diagrams) that describe the mental models of the experts.
- Presentation of results.
- Findings of uncertainty and sensitivity analysis, including sensitivity of results to different methods of aggregating expert judgments from the elicitation.
- Insights/explanations for differences in judgments among experts:
  - Degree of consensus or disagreement.
  - Degree to which views changed from initial judgments to final judgments—how much exchange and clarification of definitions and issues helped to resolve differences.
  - Principle technical reasons for difference in views, especially for outlying judgments. These may reflect different conceptual models, functional relationships, or beliefs about the appropriateness and/or weight of evidence or parameters. This qualitative explanation is an important complement to the quantitative presentation of results (Morgan, 2005).
- Whether this is the first EE on this parameter or whether there is a history or evolution of judgments that would be helpful to understand.
- Remaining uncertainties and weaknesses—possible future strategies (e.g. data development) to reduce important uncertainties or to eliminate possible biases.

- Summarize any peer review comments and what was done (and not done) to address them, including preliminary peer review conducted on methods.

Table 6-3 summarizes the technical details that should be included in the TSD.

Table 6-3. Summary of Technical Details of Expert Elicitation Methods

EE Process	Key Elements	Additional Data
Process	Description of EE process Description of reasons for elicitation and elements	Appendix to Technical Report
Expert Selection	Expertise requirements Range of affiliations and disciplines/fields Comparison of experts and how they met criteria	Appendix with criteria and basis List of Experts in Appendix Criteria for determining potential conflicts of interest
EE Questions	Charge Questions summarized Definitions of uncertainties, parameters, relationships, etc. experts addressed Definition of information elicited from experts Definition of data gaps Justification for aggregating data	Appendix with detailed questions and supporting information
EE Results	Raw data tables and post-processed results that apply elicited values to calculate relevant quantities.	Appendix with all elicited probabilities and qualitative responses.
EE Analysis	Comparison of how elicited results conform to axioms of probability theory and to empirical data Biases introduced? Direction of Biases Extreme value presentation Correlations with non-elicited components of overall analysis or policy questions Mental models	Appendix with detailed analysis of calculations, graphics, etc. Calculations and detailed analyses in Appendix Appendix with influence diagrams and other figures
EE Conclusions	Insights/explanations for differences in judgments among experts Degree of consensus and disagreement Analysis of changes in views from initial judgments to final judgments (how exchange and clarification of definitions and issues resolved differences). Technical reasons for differences in views and outlying judgments Results in context – history or evolution that would be helpful to understand	Appendix may include dissenting opinions if appropriate
Uncertainties and Weaknesses	Future strategies (e.g., develop data to reduce uncertainties and eliminate possible biases).	

6.5 WHAT ARE EXAMPLES OF EFFECTIVE EXPERT ELICITATION COMMUNICATIONS?

Many alternatives are available for conveying results to the users of EE findings. The following section provides examples of how results of EEs may be displayed qualitatively and quantitatively in text, figures, and tables. These examples are intended to demonstrate effective communication, but the suitability of any presentation depends on the particular results, context, and audience. These different presentations of results require a variety of software and range of expertise to create and interpret. When considering different displays, the analysts should consider the technical level of the audience and the aspect of the results to be highlighted. This section also identifies areas in which research suggests particularly effective means of communicating results.

6.5.1 How Can Probabilistic Descriptions Be Used?

If qualitative terms (e.g., “likely” and “probably”) are used, they should be associated with their quantitative meaning in the EE. Wallsten et al. (1986 and 2001), and other researchers have demonstrated that the quantitative probability associated with a term of likelihood varies substantially from person to person. To overcome that interindividual variability, some researchers have proposed systematizing the uses of specific terms (Moss and Schneider, 2000; Karelitz et al., 2002). For example, the terminology system of the IPCC (2005) is shown in Tables 6-4 and 6-5. Table 6-4 shows quantitatively calibrated levels of confidence. These can be used to characterize uncertainty that is based on expert judgment as to the correctness of a model, an analysis, or a statement.

Table 6-4. Quantitatively Calibrated Levels of Confidence

Terminology	Degree of Confidence in Being Correct
Very High confidence	At least 9 out of 10 chance of being correct
High confidence	About 8 out of 10 chance
Medium confidence	About 5 out of 10 chance
Low confidence	About 2 out of 10 chance
Very Low confidence	Less than 1 out of 10 chance

Table 6-5 shows a likelihood scale. This refers to a probabilistic assessment of some well defined outcome having occurred or occurring in the future—fuzzy boundaries.

Table 6-5. Likelihood Scale

Terminology	Probability of Occurrence/Outcome
Virtually certain	> 99%
Very likely	> 90%
Likely	> 66%
About as likely as not	33–66%
Unlikely	< 33%
Very unlikely	< 10%
Exceptionally unlikely	< 1%

6.5.2 How Can Text Summaries Be Used?

To improve the audience's understanding of results, Anderson et al. (1998) recommend that results should be presented:

- As a frequency (as 43 out of 10,000, rather than 0.0043);
- Within a well-defined "reference class," such as the general population to which the frequency might apply;[25] or
- By keeping constant the denominator of the frequency statement (i.e., the size of the population) constant across comparisons (such as 43 out of 10,000 and 2082 out of 10,000, rather than 43 out of 10,000 and 2 out of 10).

Anderson and other researchers have concerned themselves with the ways in which humans receive and process information, specifically with respect to uncertainties, and conclude that humans have difficulty in interpreting probabilities expressed as decimals between 0.0 and 1.0. They note that frequencies are taught early in elementary school mathematics, being part of set theory, classification, and counting, whereas probabilities generally are not taught until advanced math in high schools or universities. Consequently, the heuristics needed for an audience to interpret results correctly from an EE are more easily available to most people when presented as frequencies.

From a survey of summaries of EEs, it appears that few researchers provide simple, quantitative summaries of results. Instead, the summaries rely on graphical or tabular presentations. Policy-makers and others may have difficulty reading and understanding these technical presentations. Three examples of effective textual summaries of EE results are provided in the following box:

[25] Modified slightly from Anderson's example: "If there were 100 similar populations of Spectacled Eiders nesting in eastern arctic Russia, how many would you expect to exhibit a rate of population increase of less than -0.05?"

IEI (2004): *"…the experts exhibited considerable variation in both the median values they reported and in the spread of uncertainty about the median. In response to the question concerning the effects of changes in long-term exposures to PM2.5, the median value ranged from values at or near zero to a 0.7 percent increase in annual non-accidental mortality per 1 μg/m3 increase in annual average PM2.5 concentration. The variation in the experts' responses regarding the effects of long-term exposures largely reflects differences in their views about the degree of uncertainty inherent in key epidemiological results from long-term cohort studies, the likelihood of a causal relationship, and the shape of the concentration-response (C-R) function."*

Morgan and Keith (1995): *"Of the 13 responses received, 4 of the means lie in the interval -1 to ≤0 and 9 lie in the interval -2 to ≤ - 1."*

Titus and Narayanan (1995): *"Global warming is most likely to raise sea level 15 cm by the year 2050 and 34 cm by the year 2100. There is also a 10 percent chance that climate change will contribute 30 cm by 2050 and 65 cm by 2100."*

6.5.3 How Can Figures Be Used Effectively?

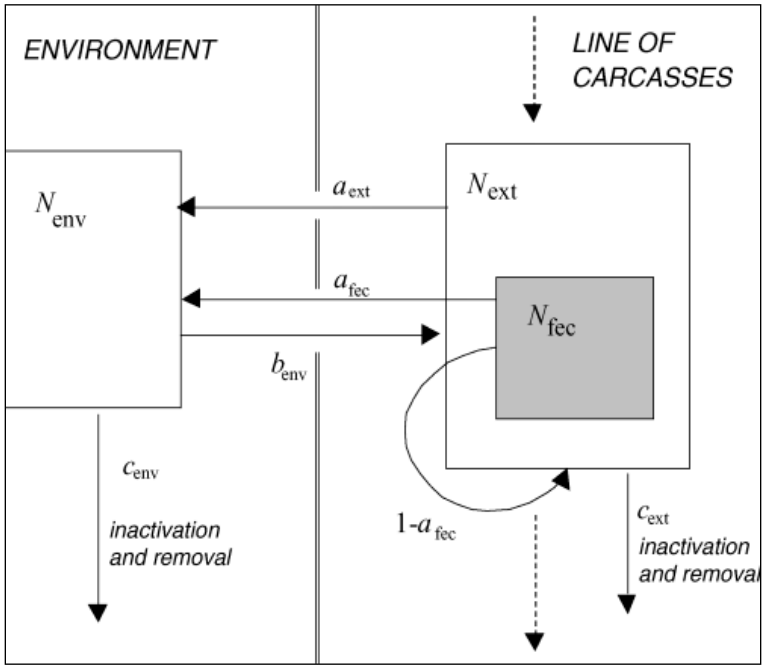
EE results should be presented in a way that promotes a clear and unambiguous understanding of the elicitation questions. Hora and Jensen (2002) note that the attribute, parameter, or relationship that is the subject of an elicitation may itself generate debate among experts. "In a sense, this is logical: if there were nothing unclear about a quantity, it would probably not be selected for elicitation. The mere fact that it was chosen in the first place implies that it is critical in some sense, and perhaps the difficulties extend to its definition." Because it is difficult to present uncertain quantities with text alone, diagrams and figures can lead to more effective communication.

6.5.3.1 Influence Diagram

"Influence diagrams" are frequently used to illustrate the question, the important parameters, and relationships that are understood to compose the elicited question. Because many EE contexts include a complex web of dependency among parameters (see Section 5.2.1), providing an influence diagram for the experts can prepare them to make judgments by placing various results in proper context.

In Figure 6-1, Nauta et al. (2005) provides a useful influence diagram that captures the uncertain model parameters to be elicited within the graphic presentation. This type of presentation improves the clarity of the model and parameters by helping to facilitate both

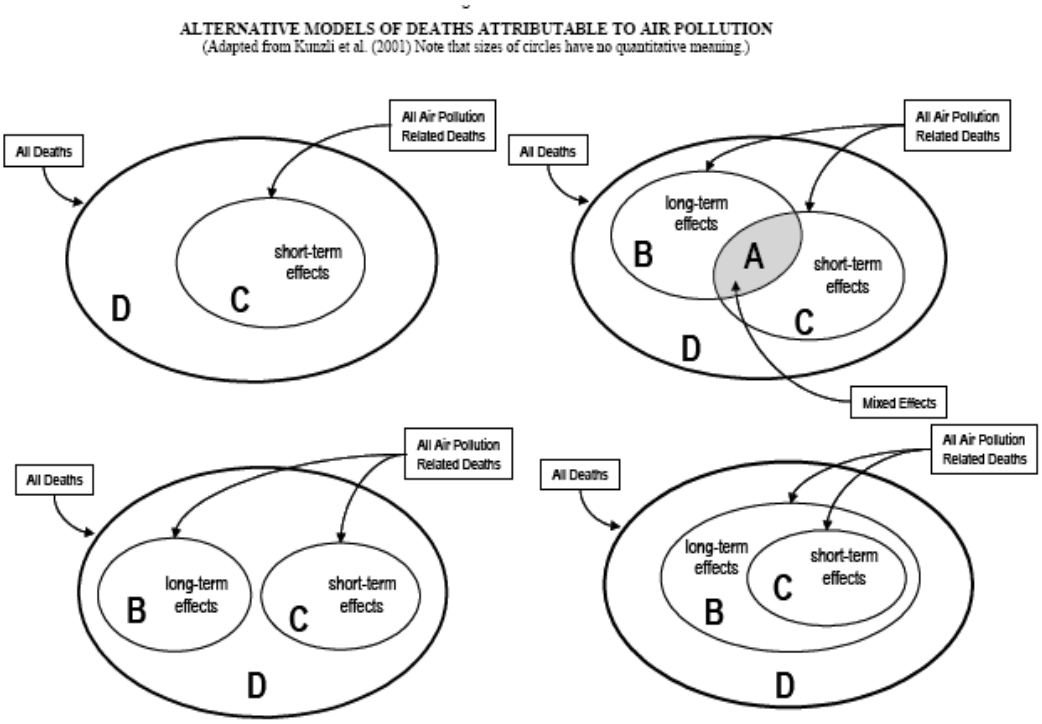
common conceptual models (or identification of differences) and the meaning of the parameters elicited.



Source : Nauta et al. (2005)

Figure 6-1. Example Influence Diagram

In Figure 6-2, IEC (2004) presents a set of influence diagrams that were adapted from Kunzli et al. (2001). The original model, in the upper right of Figure 6-2, was adapted to illustrate variant conceptual frameworks (mental models) for describing the relationship among different causes of mortality.

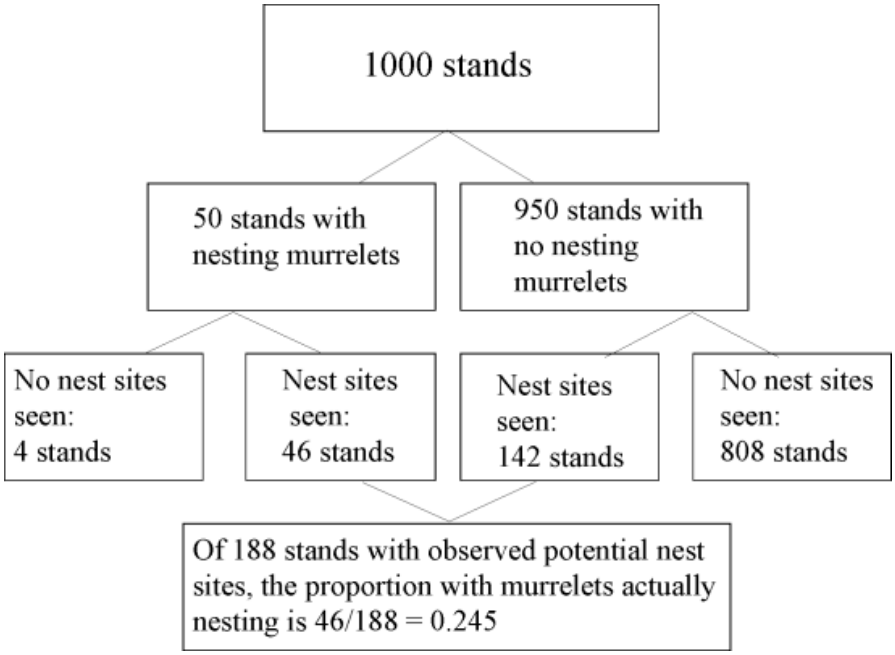


Source: Adapted from Kunzli et al. (2001) by IEC (2004)

Figure 6-2. Example of Alternative Mental Models Held of Different Experts

6.5.3.2 Frequency-Solution Diagram

An example of a frequency-solution diagram, including hypothetical statements of frequencies and the reference population, is provided in Figure 6-3 (adapted from Anderson, 1998). Anderson (1998) found that "presentation of the data in frequency format seems to encourage mental imagery and facilitate estimation of the correct answer."

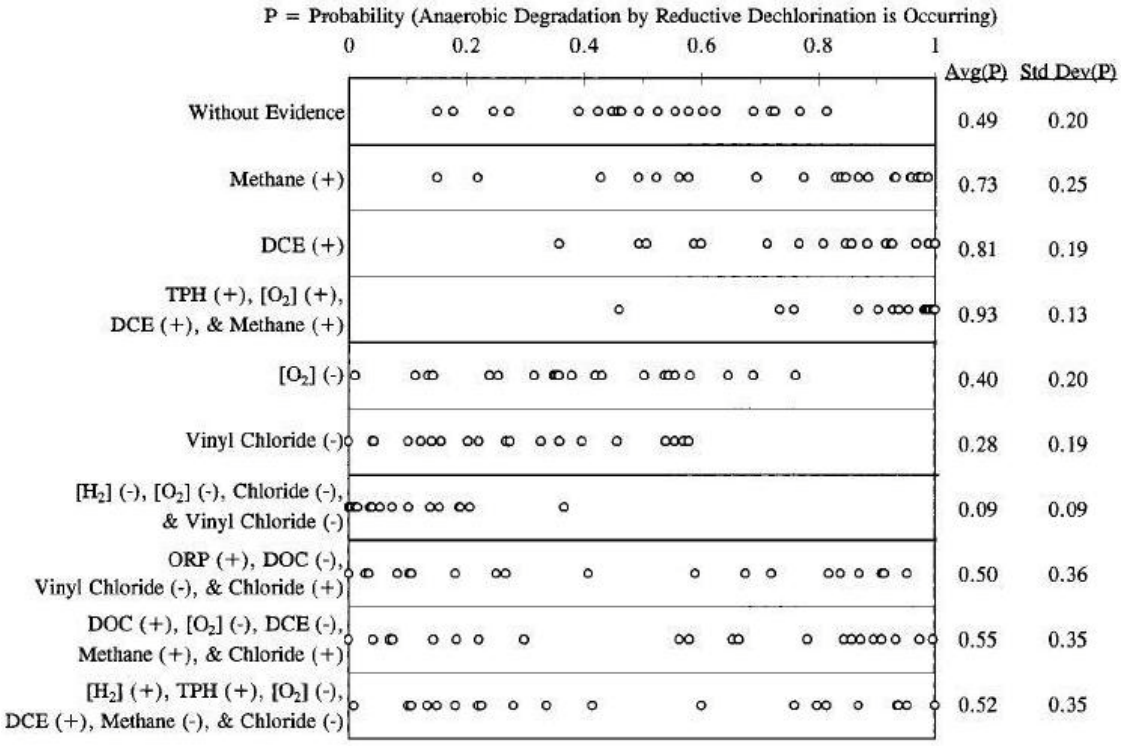


Source: Anderson (1998)

Figure 6-3. Example of a Frequency-Solution Diagram

6.5.3.3 Results of Multiple Models Based on Experts' Probabilities

The probabilities obtained from experts may be used as the quantitative parameters of models. These multiple models (one for each expert) can be used to provide results under different scenarios of evidence. Figure 6-4 shows an example from Stiber et al. (1999) in which the outputs of 22 models built from the probabilities obtained from 22 experts are compared for different scenarios (or cases of evidence).



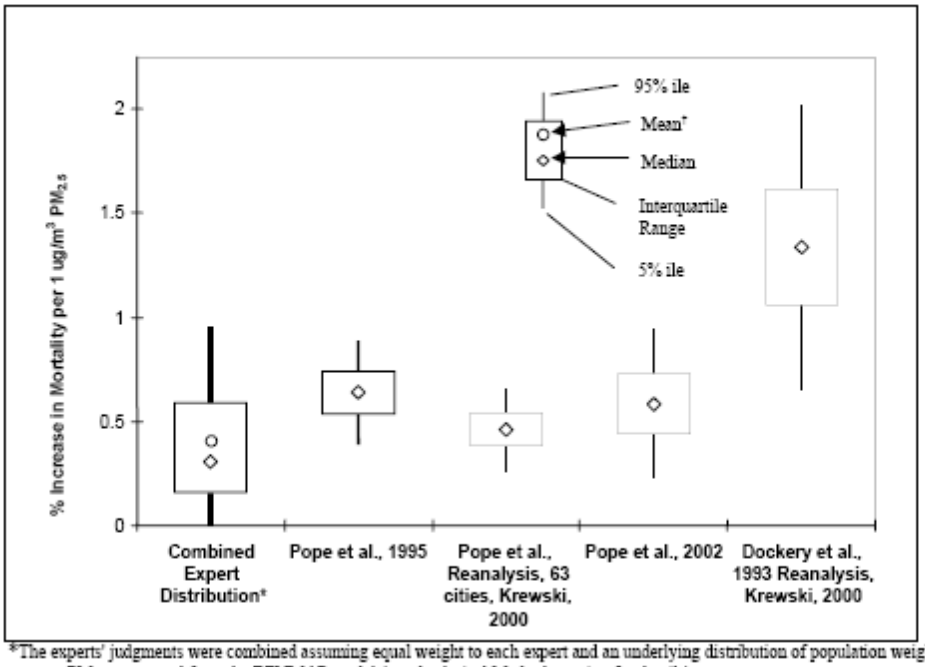
Source: Stiber et al. (1999)

Figure 6-4. Distribution of Expert Models' Predictions for Different Cases of Evidence

6.5.3.4 Box and Whisker Graphs

Sometimes experts are asked to specify a probability density function or a cumulative probability function. Below are several examples by which such results may be presented. The "box and whisker" diagrams are perhaps the most intuitively understood of the formats. Figure 6-5 is a complex but effective box-and-whisker graphic showing multiple responses from the multiple experts and providing combined expert distribution.

Comparison of Combined Expert Judgment Distribution to Results from Selected Studies: Percent Increase in Annual Non-Accidental Mortality Associated with a 1 μg/m³ Increase in Annual Average $PM_{2.5}$



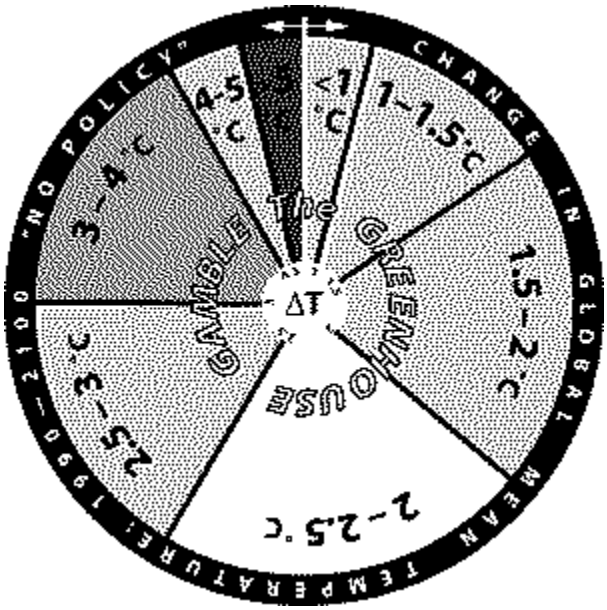
*The experts' judgments were combined assuming equal weight to each expert and an underlying distribution of population weighted annual average $PM_{2.5}$ generated from the BENMAP model (see Analytical Methods section for details).
†No mean values were elicited from experts. All means were estimated using Monte Carlo sampling methods.

Source: IEC (2004)

Figure 6-5. Example Presentation of a Simple Box and Whisker Diagram Comparing Expert Elicitation Results with Other Studies

6.5.3.5 Roulette or Probability Wheel

Because it is based on a familiar gambling paradigm, the roulette or probability wheel is a useful and intuitive display (Figure 6-6). People intuitively understand that the likelihood of any particular outcome is proportional to its area on probability wheel. This display provides an effective presentation of a future uncertain event because one of these states of the world will occur, but one does not know which.



Source: http://web.mit.edu/globalchange/www/wheel.degC.html

Figure 6-6. The Roulette or Probability Wheel to Express EE Results

6.5.3.6 Cumulative Density Functions (CDFs) and Probability Density Functions (PDFs) or Both

In the Thompson and Bloom study (2000) of EPA decision-makers, the focus group liked "the format of showing risk as a distribution, although several members indicated a preference for seeing a cumulative distribution function instead of, or in addition to, a probability density function. They expressed some confusion about the level of aggregation of the distribution (i.e. whether it was representing variability in the distribution of all the maximum individual risks for the source category, or uncertainty for the maximum individual risks for one source). Most said that they would need more information about what the distribution represents and the underlying assumptions."

The results of Pb-induced health effect from the elicitations of multiple experts can be easily compared with CDFs[26] as in Figure 6-7.

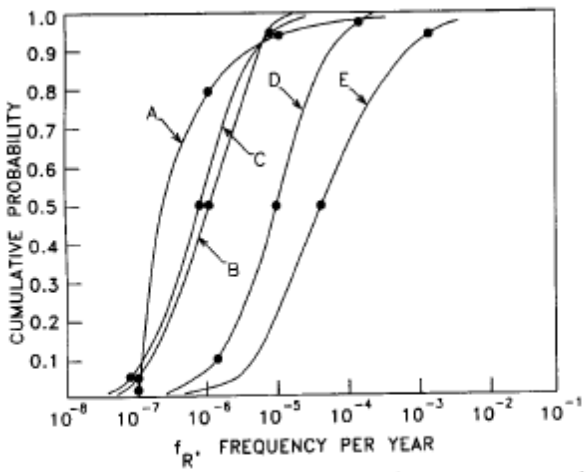


Fig. 2. Cumulative probability distributions for five experts over annual RCS rupture frequency.

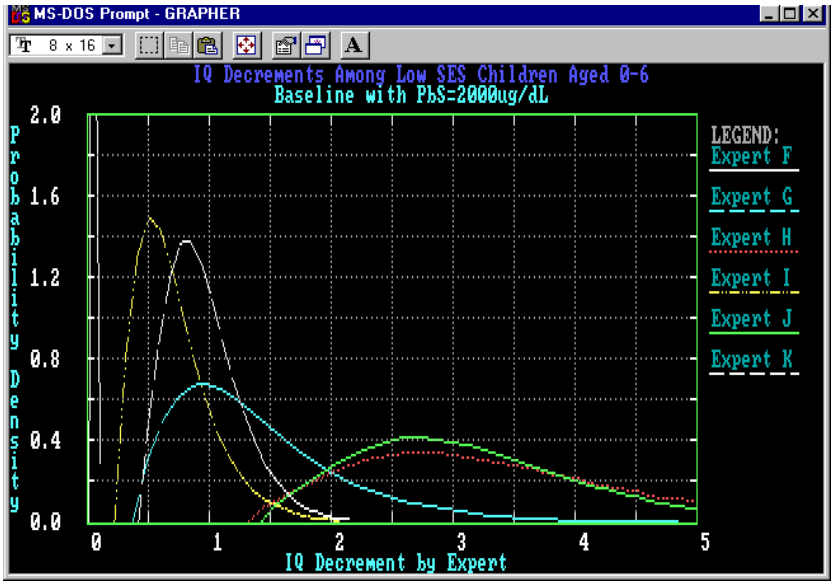
Source: Keeney and von Winterfelt (1991)

Figure 6-7. Example Presentation of CDFs of Multiple Experts in an Elicitation[27]

For the same study as in the figure just above, the Argonne National Laboratory ran the elicited probability distributions through their model and presented the alternative results (in this case, estimated IQ decrement) that would result from application of each expert's estimates (Figure 6-8).

[26] The CDF expresses the probability from 0 to 1 that the random variable (in Figure 6.7: "Frequency per Year") is less than or equal to the value on the x-axis (abscissa).

[27] The CDFs in this figure show rupture frequencies for the reactor coolant system of a nuclear reactor. Nuclear safety has been a key topic for expert elicitation specifically and decision analysis in general because nuclear accidents may be low-frequency and high-consequence events.



Source: Whitfield and Wallsten (1989)

Figure 6-8. Sample Presentation of Results of Each Expert from an Elicitation[28]

[28] The PDF expresses the probability that the random variable (in Figure 6.8: "IQ Decrement") is equal to the value on the x-axis (abscissa).

6.5.4 How Can Tables Be Used?

Some elicitations, or parts of an elicitation, may not seek quantitative data, and even when quantitative results are elicited, a table may be used to summarize the differences in paradigms among experts, helping to explain differences among elicited values. Table 6-6 provides a concise but excellent nonquantitative summary of expert responses.

Table 6-6. Sample Compilation of Qualitative Expert Responses

A	**Theoretical Construct for Long-and Short-term Exposure Effects on Premature Mortality**
	See discussion of F1 and F2 in text

B1.	**What do you believe to be the major causes of death associated with long-term exposure to $PM_{2.5}$?** ***(In order of importance)***
A	• Cardiovascular disease • lung cancer • Respiratory disease
B	• Cardiovascular disease • Respiratory disease • Not cancer – does not believe PM is likely to be a significant contributor to cancer risk
C	• Cardio-respiratory diseases probably constitute the bulk of the effects of PM but because cardiac deaths represent a very substantial portion of all deaths in the U.S. "But then our air pollution related effects are a very small part of that total." • "I think our data is highly uncertain with regard to the issue of lung cancer associated with contemporary levels of airborne particulate material and I think even more uncertain with regard to other cancers." • He thinks that PM exposure does not create a unique disease related to PM exposure. Instead, it "adds to the wear and tear of life."

D	• Broad Category of Effects: Cardio-respiratory deaths
	  • Heart Disease (CHD)
	  • COPD (particles likely contribute, but are not a major contributor)
	  • stroke, possibly
	• Cancer

E	• cardiovascular deaths • respiratory deaths (COPD, pneumonia, flu, infectious disease) • lung cancer

Source: USEPA (2004b, technical appendix)

Alternatively, a complex but clear presentation of quantitative EE results in tabular format is provided in Table 6-7.

Table 6-7. Example Table for Responses of Multiple Experts

Summary of expected extinctions relative to the number of extinctions experienced during historical times, estimated by the 11 forest ecosystem experts. The notations 'w/o CC' and 'w/CC' refer to extinctions without and with a 2 × [CO_2] climate change

Expert	Condition	North America above 45° N			Tropical forests between 20° N and 20° S		
		Mammals	Birds	Vascular plants	Mammals	Birds	Vascular plants
1	w/o CC	2	1.3	4	3.3	8	2
	w/ CC	5	1.4	8	6	10	4
2	w/o CC	0.5	3.3	1.1	2	10	10
	W/ CC	0.5 + ε	3.3 + ε	3.9	2 + ε	12	12
3	w/o CC	1	1	0.2	10	2	3
	w/ CC	2	1–2.5	0.2	10	2–30	3–20
				Note 1	Note 2	Note 2	Note 2
4	w/o CC	1	1.1	1.1	2	1.2	1.2
	w/ CC	1.3	1.2	1.2	2	1.2	1.3
5	w/o CC	10	3.3	1.05	250	10	2
	w/ CC	10	5	1.05	250	12	2
6	w/o CC	dk	dk	1	dk	dk	2.5
	w/ CC	dk	dk	1	dk	dk	6
7	w/o CC	2	6.7	3	10	>10	dk
	w/ CC	2 + ε	10	4.5	10 + ε	>10 + ε	1.5–2(na)
8	w/o CC	1–1.5	2.5	2.5–3	4	4	6
	w/ CC	1–1.5	4.5	2.5–3	4 + ε	4 + ε	6 + ε
9	w/o CC	1	1	'	10	10	10
	w/ CC	1	1.2	1.3	11	10 + ε	11
10	w/o CC	~0	dk	0.5	1	dk	1
	w/ CC	ε to 10	dk	2	1 + ε	dk	2
		Note 3		Note 4			
11	w/o CC	~0	0.33	~0	5	2	~2
	w/ CC	0.5	0.33 + ε	2	5 + ε	2 + ε	~2 + ε

dk = don't know.
Note 1: Some North American vascular plants may go extinct in the wild, but people will continue to propagate them.
Note 2: If there is a climate state change that results in the loss of the tropical forests tropical species loss could be much greater, but the climate scenario considered is too modest to produce this change.
Note 3: Expert considered all of North America including mountains of the SW. Without human intervention mammalian extinctions could be 10 times historical level. Human interventions could largely eliminate this.
Note 4: There would be seed bank efforts.

Source: Morgan et al. (2001)

As will be discussed in Section 6.4, the technical documentation and other presentations also should present important sensitivity analyses. One presentation that is particularly useful is an analysis of the overall importance of the uncertainty in the elicited results to the research question or decision at hand. Given the controversy potentially surrounding aggregation of experts' beliefs, another very useful presentation is the sensitivity analysis to different methods of combining expert judgments or of using them individually. The results of sensitivity analyses can be summarized in tables and in graphics and are often a critical component of high-level and public presentations in addition to technical documentation. The decision-maker will want to understand the implications—or the lack of influence—of such methodological choices on the support for decision choices. Table 6-8 illustrates such a sensitivity analysis.

Table 6-8. Example Table Presenting a Sensitivity Analysis for Combining Experts

Table D-5

Sensitivity Analysis of Combined Results for Effects of Long-term $PM_{2.5}$ Exposure to Individual Expert Results

Percentiles	Combined Results All Experts[a]	Percent Change in Combined Results				
		Minus A	Minus B	Minus C	Minus D	Minus E
95th	0.94	-2%	-0.4%	15%	-4%	-20%
75th	0.59	-5%	13%	18%	-0.5%	-26%
50th	0.30	-19%	21%	21%	-3%	-35%
25th	0.15	-22%	21%	21%	5%	-41%
5th	0.00	0%	0%	0%	0%	0%
Mean (Estimated)[b]	0.40	-8%	10%	20%	-3%	-28%

a. The combined values are averages across experts at each percentile. The method gives equal weight to each expert's distribution. Combination method uses population-weighted distribution of annual mean $PM_{2.5}$ concentrations in U.S. (from BENMAP model).

b. The mean is estimated by combining the distributions using Monte Carlo simulation (see text for full discussion of methodology).

Source: USEPA (2004b)

6.6 HOW CAN EXPERT ELICITATIONS BE TRANSPARENT, DEFENSIBLE, AND REPRODUCIBLE?

EPA's *Information Quality Guidelines* (USEPA, 2002) specify that all information disseminated by the Agency meet a high standard of quality. This rigorous attention to quality is particularly relevant for EEs that were conducted as part of a regulatory process. When an EE is a component in a regulatory analysis, it may receive significant public attention.

In accordance with EPA's *Peer Review Handbook* (2006a), influential scientific and technical work products used in decision-making will be peer reviewed. The mechanism of

review for a work product depends on its significance, the decision-making timeframe, level of public interest, and other factors. Regardless of the peer review mechanism selected, it is important that the reviewers (whether they are other EPA employees, independent external experts, or members of the public) are able to follow and understand the process of an EE.

The methods selected for analyzing EE data are of interest to peer reviewers. Given that many methods are available and the choice of a particular method could influence the outcome, peer reviewers are certain to examine this process. A method selected for arbitrary, subjective reasons is sure to attract criticism.

6.7 SHOULD EXPERT JUDGMENTS BE AGGREGATED FOR POLICY DECISIONS?

In Section 5.4, the appropriateness and conduct of multi-expert aggregation was discussed within the context of the EE process. As was noted, not all EEs are amenable to meaningful aggregation (Morgan and Henrion, 1990). Even when a particular project and its results are well suited for aggregation, the analysts should preserve and present the richness of each individual expert's beliefs while showing the aggregated results. This section provides additional thoughts on issues concerning the aggregation of multiple experts.

Given potential impact that aggregation can have on the interpretation of expert results, the use of aggregation should be decided on a case-by-case basis. Such decisions also may be addressed by a peer review, as well as the experts themselves. To support a peer review, any decision to combine experts should be well documented to explain why and how the experts were aggregated. This documentation should include a well-developed rationale for the decision, methods selected, a discussion of the influence of aggregation on findings, and an analysis of the sensitivity of the results to aggregation by different methods (or not aggregating). Meta-analytic techniques can be used to estimate relative importance of differences among expert views or of combining elicited judgments.

6.8 HOW CAN EXPERT ELICITATION RESULTS AND OTHER PROBABILITY DISTRIBUTIONS BE INTEGRATED?

Risk assessments and policy analyses frequently require the synthesis of multiple types of disparate data. Because these data, including EE results, are derived from multiple disciplines and sources, it is important to integrate the data with caution. After integrating the elicited results with other parameters in a larger analysis, the analyst should critically evaluate the outcome to consider whether it is physically, biologically, and logically plausible.

Bayesian updating is a useful method for integrating the results of an EE with a prior distribution (Stiber et al., 1999). In simple terms, Bayesian updating may be described as having

one set of data or one distribution—or "prior"—then updating that prior as new data become available. The state of knowledge before the EE may be used as a prior. The findings of the EE then can provide an update on the prior. As better observations become available, they should be used.

6.9 HOW CAN AN EXPERT ELICITATION BE EVALUATED *POST HOC*?

The importance of presenting sensitivity and uncertainty analyses to provide insight into the strengths and weaknesses of the EE already has been discussed. Additional evaluations of the EE serve both future research design and the influence of the EE for various decision options.

A *post hoc* analyses should consider choice of model, distributions, bounding of the parameters, method of combination (if any), and the parameters themselves. Whether a "back of the envelope" analysis or a more formal approach is used will depend on the importance of the findings. Do the results of the EE sufficiently answer the question that was at issue? Did the EE reduce or resolve controversy and assist in determining what should (or should not) be done? If not, has the EE revealed how a different analysis might resolve the controversy, or has it exposed the need for different or new data or models? Was the original question well-posed, or should the question be restated or better defined?

6.10 SUMMARY

The protocol is developed, the experts are selected, and the probabilities are elicited, but the findings of an EE only are beneficial after they are presented. Because the presentation of results is what most readers and decision-makers will see and read, it is used to judge the findings, form opinions, and make decisions. Although a TSD contains all relevant details, pertinent findings must be abstracted for effective presentation to different users. The manner of presentation is critical because users have various backgrounds, preferences, and paradigms for using data. Hence, the presentation of results should ideally be part of a communication strategy that focuses on users and their needs but also does not distort the findings, their strengths, or limitations of the EE. This chapter provided some examples for communicating EE results via probabilistic descriptions, text, figures, and tables. These examples were effective in their contexts and similar presentation can be considered by other practitioners. In addition, this chapter discussed issues concerning how to make defensible decisions about the aggregating expert judgments, combining EE results with other data, and providing peer review of findings.

7.0 FINDINGS AND RECOMMENDATIONS

The purpose of this Task Force was to initiate a dialogue within the Agency about the choice, conduct (including selection of experts), and use of EE and to facilitate future development and appropriate use of EE methods. The Task Force facilitated a series of discussions to explore the potential utility of using EE and to evaluate and address issues that may arise from using this approach. Based on those discussions, the Task Force has developed a set of findings and recommendations concerning: (1) when it is appropriate to use EE (i.e., what constitutes "good practice" in deciding whether to conduct an EE), (2) how to plan and conduct such assessments, and (3) how to present and use the results of such assessments. The Task Force has also identified various recommended steps to facilitate future development and application of these methods within EPA. Sections 7.l and 7.2 summarize the Task Force's findings and recommendations, respectively.

7.1 FINDINGS

The findings of the Task Force are as follows:

7.1.1 What is Expert Elicitation?

- For the purposes of this White Paper, the Task Force has developed an operational definition of EE as the formal, systematic process of obtaining and quantifying expert judgment on the probabilities of events, relationships, or parameters. This definition also applies to expert judgment as identified in existing or proposed guidelines, including OMB's *Circular A-4* and EPA's revised *Cancer Risk Assessment Guidelines*.
- EE is recognized as a powerful and legitimate tool. It can enable quantitative estimation of uncertain values and can provide uncertainty distributions when data are unavailable or inadequate. In addition, EE may be valuable for questions that are not necessarily quantitative, such as model conceptualization or selection, and the design of observational systems.
- EE is one type of expert judgment activity. In general, expert judgment is an inherent and unavoidable part of many EPA assessments and decisions. Expert judgment is employed in many stages of EPA analyses (e.g., problem formulation, model selection, study selection, estimation of input values) and for the interpretation and communication of results. In addition, expert judgment is a component in the external peer review of EPA assessments.
- EE concerns questions of scientific information rather than societal values or preferences. In the broader set of tools for expert judgment, there are methods for capturing and incorporating values and preferences. In addition, the elicitation of preferences or economic

valuation (e.g., willingness to pay for avoided risks) are related topics but were not the focus of the Task Force and are not included in this White Paper's operational definition of EE.

- The results of an EE provide a characterization of the current state of knowledge for some question of interest. This is useful when traditional data are unavailable or inadequate. Because an EE does not include measurements, observations, or experiments of the physical environment, however, it does not create new empirical data. Rather, it provides subjective estimates from experts that characterize the state of knowledge about some uncertain quantity, event, or relationship. Hence, the decision to conduct an EE should consider the relative cost and/or feasibility of performing research and collecting additional data.

7.1.2 What is the Role of Expert Elicitation at EPA?

- Past experience with EE at EPA (e.g., in OAQPS since the late 1970s) indicates that it can provide useful, credible results. NAS has highlighted these past efforts as exemplary and recommended that EPA continue in the direction established by these precedents.
- The use of EE is appropriate for some, but not all, situations. Factors favoring the use of EE include: inadequate information to inform a decision, lack of scientific consensus, and the need to characterize uncertainty. See Section 4.6 for a summary of these factors. Typically, an EE requires a significant investment of resources and time to provide credible results.
- EE can work well when a scientific problem has a body of knowledge but lacks a consensus interpretation, such as for an emerging scientific challenge or one that depends on uncertain future events. In such cases, expert beliefs about the value and meaning of data can provide valuable insights. When a problem has abundant relevant empirical data and relative consensus exists in the scientific community, however, there is probably little need to conduct an EE. At the other end of the spectrum, if data are inadequate for the experts to develop judgments, an EE may not be worthwhile.
- Given that EPA uses other more familiar approaches to characterize uncertainty, the application and acceptance of EE at EPA will likely grow with experience. If early EE efforts are well-designed and implemented, this will promote the credibility and endorsement of EE within the Agency and by external stakeholders.
- The nature of the regulatory process (i.e., legal, political, financial, technical, and procedural considerations) will influence whether and how to conduct an EE and how to communicate and use results. Within the regulatory process, EPA can use EE to encourage transparency, credibility, objectivity (unbiased and balanced), rigor (control of heuristics and biases), and relevance to the problem of concern.

7.1.3 What Factors Are Considered in the Design and Conduct of an Expert Elicitation?

- EE design and interpreting results generally are case-specific and context-specific. Hence, the conduct of an EE does not lend itself to a rigid "cookbook" approach, and it is important that practitioners exercise innovation when implementing EE. Nevertheless, following a number of fundamental steps can help to promote a credible and defensible EE (Chapter 5).
- An EE includes distinct roles for the members of the project team (generalist, analyst, and subject matter expert) and the experts whose judgments are the subject of the EE.

7.2 RECOMMENDATIONS

The recommendations of the Task Force are as follows:

7.2.1 What Challenges Are Well-Suited for an Expert Elicitation?

- EE is well-suited for challenges with complex technical problems, unobtainable data, conflicting conceptual models, available experts, and sufficient financial resources.
- There are two primary motivations to consider the use of EE. The first is to characterize uncertainty where existing data are inadequate and additional studies are infeasible. The second is to fill data gaps where conventional data are unobtainable in the timeframe needed for a decision.
- EE results can provide a proxy for traditional data, but it is not equivalent to valid empirical data. EE should not be used as a substitute for conventional research when empirical data can be obtained within the available time and resources..
- Before deciding to conduct an EE, managers and staff should engage in discussions about:
    - The goals of the EE and the basis for the selection of this approach;
    - The anticipated output from the EE and how it may be used in the overall decision;
    - EEs that have been used for similar types of decisions;
    - The outcome and presentation of the EE on completion; and
    - The time and cost of conducting a defensible EE.
- Investigators considering the use of EE should consider alternative methodologies to characterize uncertainty or fill data gaps (Chapter 4).

7.2.2 How Should an Expert Elicitation Be Designed and Conducted?

- The design and conduct of EE should address the following issues:
    - Standards of quality for EEs that are a function of their intended use (e.g., to inform research needs, to inform regulatory decisions) and a minimum set of best practices.
    - How to interpret the quality of the results and the EE process.

  - How to review or interpret efforts in the context of their use (i.e., how does acceptability depend on context?).
  - The role of stakeholders early in the EE planning process to provide input on relevant questions or issues.
  - Appropriateness of secondary application of EE results (i.e., the use of results beyond the purpose intended when the study was designed).
  - Under what circumstances and how should experts' judgments be combined.
  - Comparison of quality/usefulness of various types of research findings: empirical data, external expert recommendations, and EE result.
  - Whether the judgments of individual experts should be weighted differentially to produce an aggregate judgment. If so, what criteria measures are most equitable?
  - How results should be used and communicated to decision-makers.
- Until appropriate EE resources are developed, those considering and/or conducting an EE within EPA should carefully consider the issues, examples, concerns, and references presented in this White Paper.
- EEs should focus on those aspects of uncertainty that cannot be adequately described by empirical data. The EE protocol should avoid overlap between what the experts are asked and what the data adequately describe.
- For questions that require characterization of a quantity that encompasses several aspects of uncertainty, it may be appropriate to disaggregate the problem and ask the experts to assess each aspect separately.
- The long-term success of EE may depend heavily on whether its early applications are considered credible and helpful by decision-makers and stakeholders. Hence, a broader application and acceptability of EEs at EPA can be facilitated by well-designed and implemented studies.

Therefore, the Task Force recommends that:

  - Early efforts by EPA program or regional offices with little EE experience should include collaboration with knowledgeable staff within the Agency (e.g., members of this Task Force) and/or external EE specialists. These efforts should cogitate the approaches and considerations outlined in Chapter 5 for design and conduct of an EE.
  - Given that the success of the EPA/OAQPS 1980s efforts (cited by the 2003 NAS panel as exemplary efforts) benefited significantly from early collaborations with SAB and external EE specialists, similar collaborations may be highly desirable in early efforts by other offices. This collaborative approach can help to ensure the quality, credibility, and relevance of these efforts.

  - If feasible, EPA EE experts should develop training materials to teach EE basics. Those involved in the design, conduct, or use of EEs should draw on these materials to promote familiarity with EEs and to obtain the advice of those with greater experience.
- To facilitate learning among EPA staff and offices about EEs, EPA should consider providing EE resources, such as:
  - Examples of well-conducted EEs, including protocols, criteria used for selecting experts, peer reviews, and so forth.
  - Documentation for these EEs should include discussion of advantages and limitations and lessons learned.
  - Internal tracking of ongoing and planned EE efforts.
  - Establishment of an EE Community of Practice, where staff can learn informally from each other on how to improve their work and identify opportunities for collaboration.
- Additional research and discussion is recommended to better determine the appropriate level of disaggregation for EE questions.

7.2.3 How Should Experts Be Selected?

- For highly influential and potentially controversial EEs, additional steps are recommended to establish that experts were selected without bias and include the range of scientific perspectives. This is of special importance if an EE project may aggregate expert judgments.
- Whether the EE is established, controlled, and managed by EPA staff or contractors, it is necessary to select experts who are free of "conflicts of interest" and "appearance of a lack of impartiality."
- The involvement of the EE's sponsor (e.g., EPA) in the process of nominating and selecting experts should be decided on a case-by-case basis. On one hand, sponsor may wish to exert direct control over the quality and credibility of the process. On the other hand, if the sponsor does not participate, the perception of objectivity may increase. As a default practice, an EE project team should give this issue careful consideration.
- To comply with PRA, EPA may need to submit an information collection request (ICR) to OMB if more than nine experts will participate in an EE and if the questions are "identical" as defined by PRA. To avoid triggering PRA, it may be expedient to use a maximum of nine experts or work with the OGC PRA attorney to ensure that the questions do not trigger the ICR requirement. When making this decision, the importance of the EE, the availability of experts, and the range of perspectives that are sought should be taken into account.

7.2.4 How Should Expert Elicitation Results Be Presented and Used?

- Experts who participate in an EE should be identified by name and institutional affiliation, but to promote candid responses, their actual judgments may be kept anonymous. A record of all judgments, however, should be maintained and provided for any required auditing or if needed for peer review.
- EPA should link and/or integrate its EE efforts with its ongoing efforts to promote the use of probabilistic risk assessment. Lessons about communicating to decision-makers and other stakeholders can be derived from common efforts that are related to probabilistic analysis.

7.2.5 What is the Role of Peer Review and Peer Input in Expert Elicitation Projects?

- Peer review of any EE exercise should focus on the EE process, including how the experts were selected, what information they were provided, how the EE was conducted (including controlling for heuristics and biases), and how the results were analyzed and are presented. The peer review should include subject matter experts and EE specialists. The purpose of peer reviewing an EE is to review the process, not to second-guess the expert judgments.
- Depending on the purpose of the EE, a peer review of the expert selection process, the EE methods, and the results of any pilots may be appropriate prior to conducing the actual EE. Peer input about the EE protocol (i.e., prior to conducting the elicitations) may be very useful. Receiving this *ex ante* consultation can improve the quality of the EE and maximize resource efficiency.

7.2.6 What Outreach, Research, and Future Steps Are Recommended?

- This White Paper may serve as a guide for discussions on how EE might enhance Agency decision-making, as appropriate. For example, EPA could identify cross-cutting scientific issues where an improved characterization of uncertainty could impact multiple Agency assessments (e.g., a parameter or relationship that affects numerous exposure analyses) and which are good candidates for EE.
- EPA should consider supporting research on EE methods development and evaluation that are related its environmental and regulatory mission (e.g., the appropriate use and limitations of probabilistic and nonprobabilistic EE methodologies).
- EPA should continue to support efforts to co-sponsor or participate in workshops, colloquia, and professional society meetings to promote dialogue, encourage innovation, and improve the quality and appropriate use of EE assessments.

REFERENCES

American Bar Association, 2003. Comment Letter from William Funk, Chair-Elect of ABA Section of Administrative Law and Regulatory Practice to Loraine Hunt, Office of Information and Regulatory Affairs, OMB. April 24, 2003.

Amaral, D.A.L., 1983. *Estimating Uncertainty in Policy Analysis: Health Effects from Inhaled Sulfur Dioxides*. Doctoral Dissertation, Carnegie Mellon University.

Anderson, J. L., 1998. Embracing Uncertainty: The Interface of Bayesian Statistics and Cognitive Psychology. *Conservation Ecology* online: 1(2). (http://www.ecologyandsociety.org/vol2/iss1/art2/#AdaptingBayesianAnalysisToTheHumanMind:GuidelinesFromCognitiveScience).

Arnell, N.W., E. L. Tompkins, and W.N. Adger, 2005. Eliciting Information from Experts on the Likelihood of Rapid Climate Change. *Risk Analysis* 25(6):1419–1431.

Ayyub, B.M., 2000. *Methods for Expert-Opinion Elicitation of Probabilities and Consequences for Corps Facilities*. IWR Report 00-R-10. Prepared for U.S. Army Corps of Engineers Institute for Water Resources, Alexandria, VA. (http://www.iwr.usace.army.mil/docs/iwrreports/00-R-101.pdf).

Ayuub, B.M., 2001. *A Practical Guide on Conducting Expert-Opinion Elicitation of Probabilities and Consequences for Corps Facilities*. IWR Report 01-R-01. Prepared for U.S. Army Corps of Engineers Institute for Water Resources. Alexandria, VA. (http://www.iwr.usace.army.mil/docs/iwrreports/01-R-01.pdf).

Bedford, T. and R. M. Cooke, 2001. *Probabilistic Risk Analysis: Foundations and Methods*.

Bloom, D.L. et. al., 1993. *Communicating Risk to Senior EPA Policy Makers: A. Focus Group Study*. U.S. EPA Office of Air Quality Planning and Standards, Research Triangle Park, NC.

Brown, S. R., 1980. *Political Subjectivity: Application of Q Methodology in Political Science*. Yale University Press, New Haven.

Brown, S. R., 1996. Q Methodology and Qualitative Research. *Qualitative Health Research* 6(4): 561–567.

Bruine de Bruin, W., B. Fischhoff, S. G. Millstein, and B. L. Halperm-Felsher, 2000. Verbal and Numerical Expressions of Probability: "It's a Fifty-Fifty Chance." *Organizational Behavior and Human Decision Processes* 81:115–131.

Bruine de Bruin, W., P. S. Fischbeck, N.A. Stiber, and B. Fischhoff, 2002. What Number is "Fifty-Fifty"?: Redistributing Excessive 50% Responses in Elicited Probabilities. *Risk Analysis* 22:713–723.

Chechile, R. A., 1991. Probability, Utility, and Decision Trees in Environmental Analysis. In: *Environmental Decision Making: A Multidisciplinary Perspective*, R. A. Chechile and S.Carlisle, eds. Van Nostrand Reinhold, New York, pp. 64–91.

Clemen, R., 1996. *Making Hard Decisions*. Duxbury Press, Belmont, CA

Clemen, R. and R. L. Winkler, 1999. Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis* 19 2):187–203.

Clemen, R.T., 1989. Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting* 5:559–583.

Clemen, R.T., 2008. A Comment on Cooke's Classical Method. *Reliability Engineering and System Safety* 93(5):760–765.

Clemen, R.T and R.L. Winkler, 1985. Limits for the Precision and Value of Information From Dependent Sources. *Operations Research* 33(2):427–442

Cooke, R.M., 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science.* Oxford University Press, New York.

Cooke, R. M. and L.H. J. Goossens, 2000. Procedures Guide for Structured Expert Judgment in Accident Consequence Modelling. *Radiation Protection Dosimetry* 90(3):303–309.

Cooke, R.M. and B. Kraan, 2000. Uncertainty in Computational Models for Hazardous Materials— A Case Study. *Journal of Hazardous Materials* 71:253–568.

Cooke, R.M. and L.H. J. Goossens, 2008. TU Delft Expert Judgment Database. *Reliability Engineering and System Safety* 93(5): 657–674.

Covello, V.T., 1987. Decision Analysis and Risk Management Decision Making: Issues and Methods. *Risk Analysis* 7(2):131–139.

Crawford-Brown, D., 2001. Scientific Models of Human Health Risk Analysis in Legal and Policy Decisions. *Law and Contemporary Problems* 64(4):63–81.

Cullen, A. C. and H. C. Frey, 1999. *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing With Variability and Uncertainty in Models and Inputs*. Plenum, New York.

Dalkey, N. C., 1969. *The Delphi Method: An Experimental Study of Group Opinion*. Rand Corp, Santa Monica, CA.

Delbecq, A. L., A. H. Van de Ven, and D. H. Gustafson, 1975. *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*. Scott Foresman and Company, Glenview, IL.

Deisler, P.F., 1988. The Risk Management-Risk Assessment Interface. *Environmental Science and Technology* 22:15

Dutch National Institute for Public Health and the Environment (RIVM), 2003. *Uncertainty Analysis for NO_x Emissions from Dutch Passenger Cars in 1998: Applying a Structured Expert Elicitation and Distinguishing Different Types of Uncertainty*. RIVM Report 550002004/2003.

Ehrlinger, J., T. Gilovich, and L. Ross, 2005. Peering Into the Bias Blind Spot: People's Assessments of Bias in Themselves and Others. *Personality and Social Psychology Bulletin* 31(5):1–13.

Ehrmann, J. R. and B. L. Stinson, 1999. Joint Fact-Finding and the Use of Technical Experts. In: *The Consensus Building Handbook: A Comprehensive Guide to Reaching Agreement*, L. E. Susskind, S.McKearnan and J.Thomas-Larmer, eds. Sage Publications, Thousand Oaks, CA, pp. 375–399.

European Commission, 2000. *Procedures Guide for Structured Expert Judgment*. EUR 18820EN. Nuclear Science and Technology, Directorate-General for Research.

Evans J.S., G.M. Gray, R.L. Sielken Jr, A.E. Smith, C. Valdez-Flores, and J.D. Graham, 1994. Use Of Probabilistic Expert Judgment in Uncertainty Analysis of Carcinogenic Potency. *Regulatory Toxicology and Pharmacology* 2:15–36.

Feagans, T. and W.T. Biller, 1981. Risk Assessment: Describing the Protection Provided by Ambient Air Quality Standards. *The Environmental Professional* 3(3/4):235–247.

Finkel, A., 1990. *Confronting Uncertainty in Risk Management: A Guide for Decision Makers*. Center for Risk Management, Resources for the Future, Washington, DC.

Fischhoff, B. 1994. What Forecasts (Seem to) Mean. *International Journal of Forecasting* 10:387–403.

Fischhoff, B., 1995. Risk Perception and Communication Unplugged: Twenty Years of Process. *Risk Analysis* 15(2):137–145.

Fischhoff, B., 1998. Communicate Unto others. *Journal of Reliability Engineering and System Safety* 59:63–72.

Fischhoff, B., 2003. Judgment and Decision Making. In: *The Psychology of Human Thought*, R. J. Sternberg and E.E. Smith, eds. Cambridge University Press, Cambridge, UK, pp.153–187.

Garthwaite, P. H., J. B. Kadane, and A. O'Hagan, 2005. Statistical Methods for Eliciting Probability Distributions. *Journal of the American Statistical Association* 100:680–701.

Genest, C. and J.V. Zidek, 1986. Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science* 36:114–148.

Gokhale, A.A., 2001. Environmental Initiative Prioritization With a Delphi Approach: A Case Study. *Environmental Management* 28:187–193.

Hamm, R.M., 1991. *Modeling Expert Forecasting Knowledge for Incorporation Into Expert Systems*. Institute of Cognitive Science Tech Report 91-12. Institute of Cognitive Science, University of Colorado, Boulder, CO.

Hammond, K. R., R. M. Hamm, J. Grassia, and T. Pearson, 1987. Direct Comparison of the Efficacy of Intuitive and Analytical Cognition in Expert Judgment. *IEEE Transactions on Systems, Man and Cybernetics* SMC-17:753–770.

Harremoes, P., D. Gee, M. MacGarvin, A. Stirling, J.Keys, B. Wynne, and S. G. Vaz, 2001. Twelve late lessons. In: *Late Lessons from Early Warning: The Precautionary Principle 1896-2000*, P. Harremoes, D.Gee, M.MacGarvin, A.Stirling, B.Wynne and S.G.Vaz, eds. European Environment Agency, Copenhagen, Denmark, pp. 168-194.

Hattis, D. and D. Burmaster, 1994. Assessment of Variability and Uncertainty Distributions for Practical Risk Analyses. *Risk Analysis* 14(5):713–730.

Hawkins N.C. and J.D. Graham, 1988. Expert Scientific Judgment and Cancer Risk Assessment: A Pilot Study of Pharmacokinetic Data. *Risk Analysis* 8:615–25.

Hawkins, N. C. and J. S. Evans, 1989. Subjective Estimation of Toluene Exposures: A Calibration Study of Industrial Hygienists. *Applied Industrial Hygiene* 4:61–68.

Hoffmann, S., P. Fischbeck, A. Krupnick, and M. McWilliams, 2007. Using Expert Elicitation to Link Foodborne Illness in the U.S. to Food. *Journal of Food Protection* 70(5):1220-1229.

Hogarth, R.M., 1978. A Note on Aggregating Opinions. *Organizational Behavior and Human Performance* 21:40–46.

Hora, S.C., 1992. Acquisition of Expert Judgment: Examples form Risk Assessment. *Journal of Energy Engineering* 118(2):136–148.

Hora, S. and M. Jensen, 2002 Expert *Judgment Elicitation*. SSI report ISSN 0282-4434. Department of Waste Management and Environmental Protection.

Hora, S. C., 2004. Probability Judgments for Continuous Quantities: Linear Combinations and Calibration. *Management Science* 50:597–604.

Hsu, M., M. Bhatt, R. Adolphs, D. Tranel, and C. Camerer, 2005. Neural Systems Responding to Degrees of Uncertainty in Human Decision Making. *Science* 310:1680–1683.

Ibrekk, H. and M.G. Morgan, 1987. Graphical Communication of Uncertain Quantities to Non-Technical People. *Risk Analysis* 7:519–529.

Industrial Economics, Inc., 2004. *An Expert Judgment Assessment of the Concentration Response Relationship Between $PM_{2.5}$ Exposure and Mortality*. Prepared for U.S. EPA Office of Air Quality Planning and Standards. (http://www.epa.gov/ttn/ecas/regdata/Benefits/pmexpert.pdf).

International Risk Governance Council, 2005. *Risk Governance Deficits: An Analysis and Illustration of the Most Common Deficits in Risk Governance*.

Intergovernmental Panel on Climate Change (IPCC), 2001. Quantifying Uncertainties in Practice (Chapter 6) *In: IPPC Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories*. GPGAUM-Corr.2001.01. (http://www.ipcc-nggip.iges.or.jp/public/gp/english/6_Uncertainty.pdf).

IPCC, 2005. Guidance Notes for Lead Authors of the IPCC Fourth Assessment Report on Addressing Uncertainties. July 2005. (http://www.ipcc.ch/pdf/supporting-material/uncertainty-guidance-note.pdf).

Jamieson, D., 1996. Scientific Uncertainty and the Political Process. *The Annals of the American Academy of Political and Social Science* 545:35–43.

Jelovsek, F. R., D. R. Mattison, and J. F. Young, 1990. Eliciting Principles of Hazard Identification From Experts. *Teratology* 42:521–533.

Johnson B.B., 2005. Testing and Explaining a Model of Cognitive Processing of Risk Information. *Risk Analysis* 25:631.

Johnson, B.B. and P. Slovic, 1995. Presenting Uncertainty in Health Risk Assessment: Initial Studies of its Effects on Risk Perception and Trust. *Risk Analysis* 15:485–494.

Jones J.A., J. Ehrhardt, L.H.J Goossens, J. Brown, R.M. Cooke, F. Fischer, I. Hasemann, B.C.P. Kraan, A. Khursheed, and A. Phipps, 2001. *Probabilistic Accident Consequence Uncertainty Assessment Using COSYMA: Uncertainty from the Dose Module*. EUR-18825.

Jouini, M.N. and R.T. Clemen, 1996. Copula Models for Aggregating Expert Opinion. *Operations Research* 44(3):444–457.

Kahneman, D., P. Slovic, and A. Tversky, eds., 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York.

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

Karelitz, T. M., M. K. Dhami, D. V. Budescu, and T. S. Wallsten, 2002. Toward a Universal Translator of Verbal Probabilities. In: *Proceedings of the 15th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. AAAI Press, pp. 298–502.

Keeney, R. L. and D. von Winterfeldt, 1991. Eliciting Probabilities from Experts in Complex Technical Problems. *IEEE Transactions on Engineering Management* 38(3):191–201.

Keith, D. W., 1996. When is it Appropriate To Combine Expert Judgments? *Climactic Change* 33:139–143.

Krayer von Krauss, M.P., E.A. Cashman, and M.J. Small, 2004. Elicitation of Expert Judgments of Uncertainty in the Risk Assessment of Herbicide-Tolerant Oilseed Crops. *Risk Analysis* 24(6):1515–1527.

Kunzli N., S. Medina, R. Kaiser, P. Quenel, F. Horak, M. Studnicka, 2001. Assessment of Deaths Attributable to Air Pollution: Should We Use Risk Estimates Based on Time Series or on Cohort Studies? *American Journal of Epidemiology* 153:1050–5.

Laden, F., J. Schwartz, F.E. Speizer, and D.W. Dockery, 2006. Reduction in Fine Particulate Air Pollution and Mortality. *American Journal of Respiratory and Critical Care Medicine* 173:667–672.

Libby, R. and R.K. Blashfield, 1978. Performance of a Composite as a Function of the Number of Judges. *Organizational Behavior and Human Performance* 21:121–129.

Lichtenstein, S., P. Slovic, B. Fischhoff, M. Layman, and B. Combs, 1978. Judged Frequency of Lethal Events. *Journal of Experimental Psychology: Human Learning and Memory* 4:551–578.

Linstone, H. A. and M. Turoff, 1975. *The Delphi Method: Techniques and Applications*. Addison-Wesley, Reading, MA. (http://www.is.njit.edu/pubs/delphibook/)

Martin, S.A., T.S. Wallsten, and N.D.Beaulieu, 1995. Assessing the Risk of Microbial Pathogens: Application of a Judgment-Encoding Methodology. *Journal of Food Protection* 58 (3):289–295.

Mazur, A., 1973. Disputes Between Experts. *Minerva* 11:243–262.

McKeown, B. and D. Thomas, 1988. *Q Methodology*. Sage Publications, Newbury Park, CA.

Merkhofer, M. W. and R. L. Keeney, 1987. A Multiattribute Utility Analysis of Alternative Sites for the Disposal of Nuclear Waste. *Risk Analysis* 7:173–194.

Meyer, M. A. and J. M. Booker, eds. 2001. *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Society for Industrial and Applied Mathematics, American Statistical Association, Philadelphia, PA.

Morgan, M.G., 1998. Uncertainty in Risk Assessment. *Human and Ecological Risk Assessment* 4(1):25–39.

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

Morgan M.G. and D.W. Keith, 1995. Subjective Judgments by Climate Experts. *Environmental Science and Technology* 29(10):468A–476A.

Morgan, M.G., S.C. Morris, M. Henrion, D.A.L. Amaral, and W.R. Rish, 1984. Technical Uncertainty in Quantitative Policy Analysis: A Sulfur Air Pollution Example. *Risk Analysis* 4:201–216.

Morgan K.M., D.L. DeKay, P.S. Fischbeck, M.G. Morgan, B. Fischhof, and H.K Florig, 2001. A Deliberative Method for Ranking Risks (II): Evaluation of Validity and Agreement Among Risk Managers. *Risk Analysis* 21:923–37.

Morgan, K., 2005. Development of a Preliminary Framework for Informing the Risk Analysis and Risk Management of Nanoparticles. *Risk Analysis* 25(6):1621–1635.

Moss, R. and S.H. Schneider, 2000. Uncertainties in the IPCC TAR: Recommendations to Lead Authors for More Consistent Assessment Reporting. In: *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report of the IPCC*, Pachauri, R., Taniguchi, R., and Tanaka, K. eds. World Meteorological Organisation, Geneva, Switzerland, pp. 33–51.

National Academies of Science (NAS), 1983. *Risk Assessment in the Federal Government: Managing the Process*. National Research Council. National Academies Press, Washington, DC.

NAS, 1984. *Pesticides in the Diets of Infants and Children*. National Research Council. National Academies Press, Washington, DC.

NAS, 1994. *Science and Judgment in Risk Assessment*. National Research Council. National Academies Press, Washington, DC.

NAS, 1996. *Understanding Risk: Informing Decisions in a Democratic Society*. National Research Council. Academies Press, Washington, DC.

NAS. 1999. *Upgrading the Space Shuttle*. National Research Council. National Academies Press, Washington, DC.

NAS, 2002. *Estimating the Public Health Benefits of Proposed Air Pollution Regulations.* National Research Council. National Academies Press, Washington, DC.

Natural Resources Defense Council, 2005. White House Weakens EPA Cancer Safeguards to Protect Chemical Industry Instead of Children. Press Release: March 29, 2005.

Nauta, M., I. van der Fels-Klerx, and A. Havelaar, 2005. A Poultry-Processing Model for Quantitative Microbiological Risk Assessment. *Risk Analysis* 25(1):85–98.

North, W., B.R. Judd, and J.P. Pezier, 1974. New Methodology for Assessing the Probability of Contaminating Mars. *Life Sciences and Space Research* 13:103–109.

O'Hagan, A., 1998. Eliciting Expert Beliefs in Substantial Practical Applications. *The Statistician* 47:21–35.

O'Hagan, A., 2005. Research in Elicitation. In: *Bayesian Statistics and Its Applications*. Research Report No. 557/05. Department of Probability and Statistics, University of Sheffield., Sheffield, UK.

O'Hagan, A., C. Buck, A. Daneshkhah, J.R. Eiser, P.H. Garthwaite, D.J. Jenkinson, J.E. Oakley, and T. Rakow, 2006. *Uncertain Judgments: Eliciting Experts' Probabilities*. John Wiley & Sons Ltd., Chichester, UK.

Ohanian. E.V., J.A. Moore, J.R. Fowle, et al., 1997. Risk Characterization: A Bridge to Informed Decision Making. Workshop Overview. *Fundamental and Applied Toxicology* 39:81–88.

Parenté, F.J. and J.K. Anderson-Parenté, 1987. Delphi Inquiry Systems. In: *Judgmental Forecasting* ,G. Wright and P. Ayton, eds. Wiley, Chichester, UK, pp.129–156.

Pope, C.A.,III, R.T. Burnett, M.J. Thun, E.E. Calle, D. Krewski, K. Ito, and G.D. Thurston, 2002. Lung Cancer Cardiopulmonary Mortality and Long-term Exposure to Fine Particulate Air Pollution. *Journal of the American Medical Association* 287(9): 1132–1141.

Renn, O., 1999. A Model for an Analytic-Deliberative Process in Risk Management. *Environmental Science and Technology* 33:3049–3055.

Renn, O., 2001. The Need For Integration: Risk Policies Require the Input From Experts, Stakeholders and the Public At Large. *Reliability Engineering and System Safety* 72**:**131–135.

Richmond, H.M., 1991. *Overview of Decision Analytic Approach to Noncancer Health Risk Assessment*. Paper No. 91-173.1 presented at Annual Meeting of the Air and Waste Management Association, Vancouver, BC.

Rosenbaum, A.S., R.L. Winkler, T.S. Wallsten, R.G. Whitfield, and H.M. Richmond, 1995. An Assessment of the Risk of Chronic Lung Injury Attributable to Long-Term Ozone Exposure. *Operations Research* 43(1):19–28.

Saaty, T. L., 1990. *Multicriteria Decision Making: The Analytic Hierarchy Process*. RWS Publications, Pittsburgh, PA.

Shackle, G. L. S., 1972a. Economic Theory and the Formal Imagination. In: *Epistemics and Economics: A Critique of Economic Doctrines*. Cambridge University Press, Cambridge, UK, pp.3-24.

Shackle, G. L. S., 1972b. Languages for Expectation. In: *Epistemics and Economics: A Critique of Economic Doctrines*. Cambridge University Press, Cambridge, UK. pp.364-408.

Shrader-Frechette, K. S., 1991. *Risk and Rationality: Philosophical Foundations for Populist Reforms*. University of California Press, Berkeley, CA.

Slovic, P., 1986. Informing and Educating the Public About Risk. *Risk Analysis* 6(4):403–415.

Slovic, P., B. Fischhoff, and S. Lichtenstein, 1979. Rating the Risks. *Environment* 21(4):14–20 and 36–39. (Reprinted in P. Slovic (ed.), 2000. *The Perception of Risk*. Earthscan, London, UK).

Slovic, P., B. Fischhoff, and S. Lichtenstein, 1988. Response Mode, Framing, and Information-Processing Effects in Risk Assessment. In: *Decision Making: Descriptive, Normative, and Prescriptive Interactions*, D. E. Bell, H.Raiffa and A.Tversky, eds. Cambridge University Press, Cambridge, UK. pp.152-166.

Southern Research Institute (SRI), 1978. *The Use of Judgmental Probability in Decision Making*. SRI Project 6780. Prepared for the U.S. Environmental Protection Agency's Office of Air Quality, Planning and Standards, Research Triangle Park, NC.

Spetzler, C.S. and C.A.S. Staehl von Holstein, 1975. Probability Encoding in Decision Analysis. *Management Science* 22:3.

Stahl, C. H. and A. J. Cimorelli, 2005. How Much Uncertainty is Too Much and How Do We Know? A Case Example of the Assessment of Ozone Monitoring Network Options. *Risk Analysis* 25:1109–1120.

Stephenson, W., 1935a. Correlating Persons Instead of Tests. *Character and Personality* 4:17–24.

Stephenson, W., 1935b. Technique in Factor Analysis. *Nature* 136(3434):297.

Stephenson, W., 1953. *The Study of Behavior: Q-Technique and its Methodology*. University of Chicago Press, Chicago.

Stiber, N.A., M. Pantazidou, and M.J. Small, 1999. Expert System Methodology for Evaluating Reductive Dechlorination at TCE Sites. *Environmental Science & Technology* 33:3012–3020.

Stiber, N.A., M.J. Small, and M. Pantazidou, 2004. Site-Specific Updating and Aggregation of Bayesian Belief Network Models for Multiple Experts. *Risk Analysis* 24(6):1529–1538.

Teigen, K.H., 1994. Variants of Subjective Probabilities: Concepts, Norms, And Biases. In: *Subjective Probability*, G. Wright and P. Ayton, eds. Wiley, New York, pp. 211–238.

Thompson, K.M. and D.L. Bloom, 2000. Communication of Risk Assessment Information to Risk Managers. *Journal of Risk Research* 3(4):333–352.

Titus, J.G., and V. Narayanan, 1996. The Risk of Sea Level Rise: A Delphic Monte Carlo Analysis in Which Twenty Researchers Specify Subjective Probability Distributions for Model Coefficients Within Their Areas of Experience. *Climatic Change* 33:151–212.

Tufte, E.R., 1983. *The Visual Display of Quantitative Information*. Graphics Press, Chesire, CT.

U.S. Department of Transportation/Federal Railroad Administration, 2003. Chapter 2: Approach to Estimation of Human Reliability in Train Control System Studies. In: *Human Reliability Analysis in Support of Risk Assessment for Positive Train Control*. DOT/FRA/ORD-03/15. Office of Research and Development, Washington, DC.

U.S. Environmental Protection Agency (USEPA), 1984. Risk Assessment and Management: Framework for Decision Making, EPA 600/9-85-002, Washington, DC.

USEPA, 1986a. *Guidelines for Carcinogen Risk Assessment*. EPA/600/8-87/045. Risk Assessment Forum, Washington, DC.

USEPA, 1986b. *Guidelines for the Health Risk Assessment of Chemical Mixtures*. EPA/600/8-87/045. Risk Assessment Forum, Washington DC.

USEPA, 1986c. *Guidelines for Mutagenicity Risk Assessment*. EPA/630/R-98/003. Risk Assessment Forum, Washington DC.

USEPA, 1986d. Guidelines for Estimating Exposure. *Federal Register* 51:34042-34054. Risk Assessment Forum, Washington DC.

USEPA, 1986e. Guidelines for the Health Assessment of Suspect Developmental Toxicants. *Federal Register* 51:34028–34040. Risk Assessment Forum, Washington DC.

USEPA, 1991. Risk Assessment Guidance for Superfund: Volume I – Human Health Evaluation Manual, Supplement to Part A: Community Involvement in Superfund Risk Assessments.

USEPA, 1992. *Guidance on Risk Characterization for Risk Managers and Risk Assessors*. Signed by Deputy Administrator F. Henry Habicht II. February 26, 1992.

USEPA, 1994. *Seven Cardinal Rules of Risk Communication*. EPA-OPA-87-020. Office of Policy Analysis, Washington DC.

USEPA, 1995a. *Guidance for Risk Characterization*. Signed by Administrator Carol M. Browner. March 21, 1995.

USEPA, 1995b. The Probability of Sea Level Rise, Office of Policy, Planning and Evaluation. EPA 230-R-95-008. Washington, DC.

USEPA, 1997. *Guiding Principles for Monte Carlo Analysis*. EPA/630/R-97/001. Risk Assessment Forum, Office of Research and Development, Washington, DC. (http://www.epa.gov/ncea/pdfs/montcarl.pdf).

USEPA, 1998. *Superfund Community Involvement Handbook and Toolkit*. EPA 540-R-98-007. Office of Emergency and Remedial Response, Washington DC.

USEPA, 1999a. *Report of the Workshop on Selecting Input Distributions for Probabilistic Assessments*. EPA/630/R-98/004. Risk Assessment Forum, Office of Research and Development, Washington, DC.

USEPA, 1999b. *Risk Assessment Guidance for Superfund: Volume I—Human Health Evaluation Manual, Supplement to Part A: Community Involvement in Superfund Risk Assessments*. EPA-540-R-98-042. Washington DC. (http://www.epa.gov/oswer/riskassessment/ragsa/pdf/rags-vol1-pta_complete.pdf).

USEPA, 1999c. *Superfund Risk Assessment and How You Can Help: An Overview Videotape*. September 1999 (English version) and August 2000 (Spanish version). English Version: EPA-540-V-99-003, OSWER Directive No. 93285.7.29B. Spanish Version (northern Mexican); EPA-540-V000-001, OSWER Directive No. 9285.7-40. Available Through NSCEP: (800) 490-9198 or (613) 489-8190.

USEPA, 2000a. *Risk Characterization Handbook*. EPA 100-B-00-002. Science Policy Council, Office of Science Policy, Office of Research and Development, Washington, DC. (http://www.epa.gov/osa/spc/pdfs/rchandbk.pdf).

USEPA, 2000b. *Guidelines for Preparing Economic Analyses*. EPA 240-R-00-003. Office of the Administrator, Washington DC. (http://yosemite.epa.gov/ee/epa/eed.nsf/webpages/Guidelines.html).

USEPA, 2000c. *Options for Development of Parametric Probability Distributions for Exposure Factors*. EPA/600/R-00/058. National Center for Environmental Assessment, Office of Research and Development, Washington DC. (http://www.epa.gov/NCEA/pdfs/paramprob4ef/chap1.pdf).

USEPA, 2001a. *Risk Assessment Guidance for Superfund: Volume III—Part A: Process for Conducting Probabilistic Risk Assessment*. EPA 540-R-02-002, OSWER 9285.7-45. Office of Solid Waste and Emergency Response, Washington, DC. (http://www.epa.gov/oswer/riskassessment/rags3adt/index.htm).

USEPA, 2001b. *Early and Meaningful Community Involvement*. Directive No. 90230.0-99Office of Solid Waste and Emergency Response, Washington DC. October 12, 2001. (http://www.epa.gov/superfund/policy/pdfs/early.pdf).

USEPA, 2002. *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Environmental Protection Agency*. (http://www.epa.gov/oei/qualityguidelines).

USEPA, 2004a. *An Examination of Risk Assessment Principles and Practices*. Staff Paper. EPA/100/B-04/001. Office of the Science Advisor, Washington DC. (http://www.epa.gov/osainter/pdfs/ratf-final.pdf).

USEPA, 2004b. *Final Regulatory Analysis: Control of Emissions from Nonroad Diesel Engines*. EPA 420-R-04-007. Office of Transportation and Air Quality, Washington, DC. (http://www.epa.gov/nonroad-diesel/2004fr/420r04007a.pdf).

USEPA, 2005a. *Guidelines for Carcinogen Risk Assessment*. EPA/630/P-03-001B. Risk Assessment Forum, Office of Research and Development, Washington, DC. (http://www.epa.gov/osa/mmoaframework/pdfs/CANCER-GUIDELINES-FINAL-3-25-05%5B1%5D.pdf).

USEPA, 2005b. *Regulatory Impact Analysis for the Final Clean Air Interstate Rule*. EPA-452/R-05-002. Office of Air and Radiation, Washington, DC.

USEPA, 2006a. *Peer Review Handbook*, 3rd Edition. Washington DC. (http://www.epa.gov/peerreview).

USEPA, 2006b. *Regulatory Impact Analysis of the Final PM National Ambient Air Quality Standards*. Washington, DC. (http://www.epa.gov/ttnecas1/ria.html).

USEPA, 2007. Regulatory Impact Analysis of the Proposed Revisions to the National Ambient Air Quality Standards (NAAQS) for Ground-Level Ozone. EPA-452/R-07-008. Washington, DC. (http://www.epa.gov/ttn/ecas/regdata/RIAs/452R07008_all.pdf).

USEPA, 2009. *Addendum to the Peer Review Handbook, 3rd edition: Appearance of a Lack of Impartiality in External Peer Reviews*. Washington DC. (http://www.epa.gov/peerreview).

U.S. Nuclear Regulatory Commission (USNRC), 1996. *Branch Technical Position on the Use of Expert Elicitation in the High-Level Radioactive Waste Program*. NUREG-1563. Division of Waste Management, Office of Nuclear Material Safety and Standards, Washington, DC.

U.S. Office of Management and Budget (USOMB), 2002. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies. *Federal Register*, 67(36):8452–8460, February 22, 2002 (http://www.whitehouse.gov/OMB/fedreg/reproducible2.pdf).

USOMB, 2003a. Draft 2003 Report to Congress on the Costs and Benefits of Federal Regulations. *Federal Register* 68:5492–5527.

USOMB, 2003b. *OMB Circular A-4, Regulatory Analysis, To the Heads of Executive Agencies and Establishments*. Office of Information and Regulatory Affairs, Washington, DC. September 17, 2003. (http://www.whitehouse.gov/omb/circulars_a004_a-4).

USOMB, 2006. *Proposed Risk Assessment Bulletin, Office of Management and Budget, Office of Information and Regulatory Affairs*. January 9, 2006. (http://www.whitehouse.gov/omb/inforeg/proposed_risk_assessment_bulletin_010906.pdf).

van der Fels-Klerx, H.J. Ine. L.H.J. Goossens, H.W. Saatkamp and S.H.S. Horst, 2002. Elicitation of Quantitative Data From a Heterogeneous Expert Panel: Formal Process and Application in Animal Health. *Risk Analysis* 22:67–81.

Walker K.D., J.S. Evans, and D. MacIntosh, 2001. Use of Expert Judgment in Exposure Assessment: Part I. Characterization of Personal Exposure to Benzene. *Journal of Exposure Analysis and Environmental Epidemiology* 11:308–322.

Walker K.D., P. Catalano, J.K. Hammitt, and J.S. Evans, 2003. Use of Expert Judgment In Exposure Assessment: Part 2. Calibration of Expert Judgments About Personal Exposures to Benzene. *Journal of Exposure Analysis and Environmental Epidemiology* 13:1–16.

Walker, K.D., 2004. Memo: Appropriate Number of Experts for the PM EJ Project. Memo to Jim Neumann, Henry Roman, and Tyra Gettleman, IEC, November 11.

Wallsten, T.S., 1986. *Meanings of Nonnumerical Probability Phrases*. Final Report for the Period 16 August 1983 Through 15 August 1986. L.L. Thurstone Psychometric Laboratory Research Memorandum No. 67, Chapel Hill, NC.

Wallsten, T.S., B.H. Forsyth, and D.V. Budescu, 1983. Stability and Coherence of Health Experts' Upper and Lower Subjective Probabilities About Dose-Response Functions. *Organizational Behavior and Human Performance* 31:277–302.

Wallsten, T.S. and D.V. Budescu, 1983. Encoding Subjective Probabilities: A Psychological and Psychometric Review. *Management Science* 29(2):151–173.

Wallsten, T.S. and R.G. Whitfield, 1986. *Assessing the Risks to Young Children of Three Effects Associated With Elevated Blood Levels, ANL/AA-32*. Argonne National Laboratory, Argonne, IL.

Wallsten, T.S., D.V. Budescu, I. Erev, and A. Diedrich, 1997. Evaluating and Combining Subjective Probability Estimates. *Journal of Behavioral Decision Modeling* 10:243–268.

Warren-Hicks, W.J. and D. R. J. Moore, eds., 1998. *Uncertainty Analysis in Ecological Risk Assessment*. SETAC Press, Pensacola, FL.

Whitfield, R.G. and T.S. Wallsten, 1989. A Risk Assessment for Selected Lead-Induced Health Effects: An Example of a General Methodology. *Risk Analysis* 9(2):197–208.

Whitfield, R.G., T.S. Wallsten, R.L. Winkler, H.M. Richmond, S.R. Hayes, and A.S. Rosenbaum, 1991. *Assessing the Risk of Chronic Lung Injury Attributable to Ozone Exposure*. Report No. ANL/EAIS-2. Argonne National Laboratory, Argonne, IL. July.

Whitfield, R.G., H.M. Richmond, S.R. Hayes, A.S. Rosenbaum, T.S. Wallsten, and R.L. Winkler, 1994. Health Risk Assessment of Ozone. In: *Tropospheric Ozone*, David J. McKee, ed. CRC Press, Boca Raton, FL.

Wilson, J.D., 1998. *Default and Inference Options: Use in Recurrent and Ordinary Risk Decisions*. Discussion Paper 98-17, February. Resources for the Future, Washington DC.

Winkler, R., T.S. Wallsten, R.G. Whitfield, H.M. Richmond and A. Rosenbaum, 1995. An Assessment of the Risk of Chronic Lung Injury Attributable to Long-Term Ozone Exposure. *Operations Research* 43(1):19–28.

Supplemental References Regarding Risk Communication and Public Perception

Ariely, D. 2008. *Predictably Irrational: The Hidden Forces that Shape our Decisions*. Harper Collins Publishers, New York.

Ariely, D., W.-T. Au, R.H. Bender, D.U. Budescu, C.B. Dietz, H. Gu, T.S. Wallsten, and G. Zauberman, 2000. The Effects of Averaging Probability Estimates Between and Within Judgments. *Journal of Experimental Psychology: Applied* 6:130–147.

Batz, M. B., M. P. Doyle, J. G. Morris, J. Painter, R. Singh, R. V. Tauxe, M. R. Taylor, M. A. Danilo, and L. F. Wong for the Food Attribution Working Group,, 2005. Attributing Illness to Food. *Emerging Infectious Diseases* 11(7):993–999.

Berger, J. O. and D. A. Berry, 1988. Statistical Analysis and the Illusion of Objectivity. *American Scientist* 76:159–165.

Bruine de Bruin, W., B. Fischhoff, L. Brilliant, and D. Caruso, 2006. Expert Judgments of Pandemic Influenza Risks. *Global Public Health* 1:178–193.

Bruine de Bruin, W., A. M. Parker, and B. Fischhoff, 2007. Individual Differences in Adult Decision-Making Competence. *Journal of Personality and Social Psychology* 92:938–956.

Fischhoff, B. and J. S. Downs, 1998. Communicating Foodborne Disease Risk. *Emerging Infectious Diseases* 3(4):489–495.

Gilovich, T., D. Griffin, and D. Kahneman, eds., 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge University Press, Cambridge, UK.

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

Hoffmann, S., P. Fischbeck, A. Krupnick, and M. McWilliams, 2007. Elicitation from Large, Heterogeneous Expert Panels: Using Multiple Uncertainty Measures to Characterize Information Quality for Decision Analysis. *Decision Analysis* 4(2): 91-109.

Hoffmann, S., P. Fischbeck, A. Krupnick, and M. McWilliams, 2008. Informing Risk-Mitigation Priorities Using Uncertainty Measures Derived from Heterogeneous Expert Panels: A Demonstration Using Foodborne Pathogens. *Reliability Engineering and System Safety* 93(5): 687-698.

Kahneman, D. and A. Tversky, eds., 2000. *Choices, Values, and Frames*. Cambridge University Press, Cambridge, UK.

Morgan, M.G., A. Bostum, L. Lave, and C. J. Atman, 1992. Communicating Risk to the Public. *Environmental Science and Technology* 26(11):2048–2056.

Morgan, M.G., B. Fischhoff, A. Bostrom, and C.J. Atman, 2002. *Risk Communication*: *A Mental Models Approach*. Cambridge University Press, Cambridge, UK.

Morgan, M. G., H. Dowlatabadi, M. Henrion, D. Keith, R. Lempert, S. McBride, M. Small, and T. Wilbanks, eds., 2009. *Best Practice Approaches for Characterizing, Communicating, and Incorporating Scientific Uncertainty in Decisionmaking*. CCSP Synthesis and Assessment Product 5.2. National Oceanic and Atmospheric Administration, Washington, DC.

Schwartz, N. 1996. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Erlbaum Press, Hillsdale, NJ.

Tversky, A. and D. Kahneman, 1974. Judgments Under Uncertainty: Heuristics and Biases. *Science* 185:1124–1131.

USEPA, 2002. *Risk Communication in Action: Environmental Case Studies*. Washington, DC. (http://www.epa.gov/ordntrnt/ORD/NRMRL/pubs/625r02011/625r02011.htm).

Winkler, R.-L. and R.T. Clemen, 2004. Multiple Experts vs. Multiple Methods: Combining Correlation Assessments. *Decision Analysis* 1(3):167–176.

Woloshin, S. and L. M. Schwartz, 2002. Press Releases: Translating Research Into News. *Journal of the American Medical Association* 287:2856–2858.

APPENDIX A: FACTORS TO CONSIDER WHEN MAKING PROBABILITY JUDGMENTS

Introduction

Uncertainty often is associated with conclusions that we draw from research and more generally in our everyday thinking. When the data desired to support a particular decision do not yet exist, are sparse, of poor quality, or of questionable relevance to the problem at hand, subjective judgment comes into play. Formal elicitation of subjective judgments, often conducted with experts in the particular field, attempts to integrate what is known with what is not known about a particular quantity into a comprehensive, probabilistic characterization of uncertainty.

Many sources may contribute to uncertainty about any given issue, and it generally is difficult for most people to consider and integrate them all. When an expert is asked to make probability judgments on socially important matters, it is particularly important that he or she consider the relevant evidence in a systematic and effective manner and provide judgments that represent his or her opinions well.

Several academic traditions—decision analysis, human factor cognitive sciences, experimental psychology, and expert systems analysis—have sought to understand how to elicit probabilistic judgments from both lay people and experts in a reliable way. Researchers have amassed a considerable amount of data concerning the way people form and express probabilistic judgments. The evidence suggests that most people use heuristics (i.e., simplifying rules) and demonstrate certain cognitive biases (i.e., systematic distortions of thought) when considering large amounts of complex information,. These heuristics and biases can lead to systematic biases in the judgments and errors of over- and under-confidence. In particular, many studies indicate that both experts and lay people tend to be overconfident. In probabilistic assessments of uncertainty, overconfidence manifests itself as placing higher probabilities on being correct (or narrower confidence intervals around a prediction) than measures of performance ultimately warrants. Such errors in judgments may have important implications for decisions that depend on them.

The purpose of this paper was to illuminate these heuristics and biases. We first reviewed the most widespread heuristics and biases and then offered some suggestions to help mitigate their effects.

Heuristics and Biases Involved in Expert Judgment

Sequential Consideration of Information

Generally, the order in which evidence is considered influences the final judgment, although logically that should not be the case. Of necessity, pieces of information are considered one by one in a sequential fashion. Those considered first and last tend to dominate judgment, however. In part, initial information has undue influence because it provides the framework that subsequent information is then tailored to fit. For example, people usually search for evidence to confirm their initial hypotheses; they rarely look for evidence that weighs against them. The latter evidence has an undue effect simply because it is fresher in memory.

Related to these sequential effects is the phenomenon of "anchoring and adjustment." Based on early partial information, individuals typically form an initial probability estimate, the "anchor," regarding the event in question. They then make adjustments to this judgment as they consider subsequent information. Such adjustments tend to be too small. In other words, too little weight is attached to information considered subsequent to the formation of the initial judgment.

Effects of Memory on Judgment

It is difficult for most people to conceptualize and make judgments about large, abstract universes or populations. A natural tendency is to recall specific members and then to consider them as representative of the population as a whole. Specific instances that are not necessarily representative, however, often are recalled precisely because they stand out in some way, such as being familiar, unusual, especially concrete, or of personal significance. Unfortunately, the specific characteristics of these singular examples are then attributed, often incorrectly, to all the members of the population of interest. Moreover, these memory effects are often combined with the sequential phenomena discussed earlier. For example, in considering evidence regarding the relationship between changes in ambient fine particulate matter and premature mortality, one might naturally think first of a study one has recently read or one that was unusual and therefore stands out. The tendency might then be to treat the recalled studies as typical of the population of relevant research and ignore important differences among studies. Subsequent attempts to recall information could result in thinking primarily of evidence consistent with the initial items considered.

Estimating Reliability of Information

People tend to overestimate the reliability of information, ignoring factors such as sampling error and imprecision of measurement. Rather, they summarize evidence in terms of simple and definite conclusions, which causes them to be overconfident in their judgments. This

tendency is stronger when one has a considerable amount of intellectual and/or personal involvement in a particular field. In such cases, information often is interpreted in a way that is consistent with one's beliefs and expectations, results are overgeneralized, and contradictory evidence is ignored or undervalued.

Relation between Event Importance and Probability

Sometimes the importance of events, or their possible costs or benefits, influences judgments about the uncertainty of the events when, rationally, importance should not affect probability. In other words, one's attitudes towards risk tend to affect one's ability to make accurate probability judgments. For example, many physicians tend to overestimate the probability of very severe diseases because they feel it is important to detect and treat them. Similarly, many smokers underestimate the probability of adverse consequences of smoking because they feel that the odds do not apply to themselves personally.

Assessment of Probabilities

Another limitation is related to one's ability to discriminate between levels of uncertainty and to use the appropriate criteria of discrimination for different ranges of probability. One result is that people tend to assess both extreme and midrange probabilities in the same fashion, usually doing a poor job in the extremes. It is important to realize that the closer to the extremes (either 0 or 1) that one is assessing probabilities, the greater the impact of small changes. It helps here to think in terms of odds as well as probabilities. Thus, for example changing a probability by 0.009 from 0.510 to 0.501 leaves the odds almost unchanged, but the same change from 0.999 to 0.990 changes the odds by a factor of about 10 from 999:1 to 99:1.

Recommendations

Although extensive and careful training would be necessary to eliminate all of the problems mentioned above, some relatively simple suggestions can help minimize them. Most important is to be aware of natural cognitive biases and to try consciously to avoid them.

To avoid sequential effects, keep in mind that the order in which you think of information should not influence your final judgment. It may be helpful to actually note on paper the important facts you are considering and then to reconsider them in two or more sequences, checking the consistency of your judgments. Try to keep an open mind until you have examined all of the evidence, and do not let the early information you consider sway you more than is appropriate.

To avoid adverse memory effects, define various classes of information that you deem relevant, and then search your memory for examples of each. Do not restrict your thinking only

to items that stand out for specific reasons. Make a special attempt to consider conflicting evidence and to think of data that may be inconsistent with a particular theory. Also, be careful to concentrate on the give probability judgment, and do not let your own values (how you would make the decision yourself) affect those judgments.

To accurately estimate the reliability of information, pay attention to such matters as sample size and the power of the statistical tests. Keep in mind that data are probabilistic in nature, subject to elements of random error, imprecise measurements, and subjective evaluation and interpretation. In addition, the farther you must extrapolate or generalize from a particular study to a situation of interest, the less reliable the conclusion may be and the less certainty should be attributed. Rely more heavily on information that you consider more reliable, but do not treat it as absolute truth.

Keep in mind that the importance of an event or an outcome should not influence its judged probability. It is rational to let the costliness or severity of outcome influence the point at which action is taken with respect to it but not the judgment that is made about the outcome's likelihood. Finally, in making probability judgments, think primarily in terms of the measure (probability or odds) with which you feel more comfortable, but sometimes translate to the alternative scale or even to measures of other events (e.g., the probability of the event not happening). When estimating very small or very large likelihoods, it is usually best to think in terms of odds, which are unbounded, instead of probabilities, which are bounded. For example, one can more easily conceptualize odds of 1:199 than a probability of 0.005.

APPENDIX B: GLOSSARY

Bayesian Analysis: Statistical analysis that describes the probability of an event as the degree of belief or confidence that a person has, given some state of knowledge that the event will occur. Bayesian Monte Carlo combines a prior probability distribution and a likelihood function to yield a posterior distribution. Also called subjective view of probability, in contrast to the frequentist view of probability.

Expert Elicitation: A systematic process of formalizing and quantifying, in terms of probabilities, experts' judgments about uncertain quantities, events, or relationships.

Expert Judgment: An inferential opinion of a specialist or group of specialists within an area of their expertise. Expert judgment (alternatively referred to as professional judgment) may be based on an assessment of data, assumptions, criteria, models, and parameters in response to questions posed in the relevant area of expertise.

Frequentist: A term referring to classical statistics in which the probability of an event occurring is defined as the frequency of occurrence measured in an observed series of repeated trials.

Likelihood Function: A term from Bayesian statistics referring to a probability distribution that expresses the probability of observing new information given that a particular belief is true.

Stochastic Process: A process involving random variables and characterized by variability in space or time.

Uncertainty: Lack of knowledge about specific variables, parameters, models, or other factors. Examples include limited data regarding concentrations of environmental contaminants. Some forms of uncertainty may be reduced through further study.

Variability: True heterogeneity or diversity in characteristics among members of a population or one individual over time.