CNWRA 94-019

BACKGROUND REPORT ON THE USE AND ELICITATION OF EXPERT JUDGMENT

Prepared for

Nuclear Regulatory Commission
Contract NRC-02-93-005

Prepared by

Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas

September 1994

CNWRA 94-019

BACKGROUND REPORT ON THE USE AND ELICITATION OF EXPERT JUDGMENT

Prepared for

Nuclear Regulatory Commission
Contract NRC-02-93-005

Prepared by

A.R. DeWispelare
L.T. Herren
E.J. Bonano
R.T. Clemen

Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas

September 1994

ABSTRACT

The Nuclear Regulatory Commission (NRC) is developing methods to determine compliance with its regulation for the disposal of high-level nuclear waste (HLW) (10 CFR Part 60) (Nuclear Regulatory Commission, 1991) by the U.S. Department of Energy (DOE). Performance of HLW geologic repositories has to be assessed for a regulatory period of 10,000 years. Because of this extremely long period, a combination of experimental methods, studies of natural analogs, and mathematical models will be used in technical assessments and in performance assessments (PAs). Mathematical models are expected to be the primary tools for estimating the long-term future performance of the repository. Identification of external conditions to which the repository will be subjected during its regulatory life is an essential requirement for applying the mathematical models. These external conditions pertain to evolving tectonism, volcanism, seismicity, climate, and others that may affect repository performance.

One compelling fact dominates the HLW program's technical analyses and assessments in the earth and atmospheric sciences: the time and space scales are so large, the interrelationships so complex, and the future so uncertain that obtaining the data to compute reliable estimates is not always feasible.

Certain data gathered for this project will require interpretation and supplementation before it can be incorporated into these mathematical models. Expert judgment elicitation is a potential source of this data interpretation and supplementation. The NRC is currently evaluating the applicability of expert judgment elicitation to the HLW repository program licensing process. Specifically, the NRC is examining the proper mechanics and procedures for conducting formal expert elicitation, as well as examining conditions which warrant the use of expert judgment in the HLW program. Of primary concern to the NRC is gaining experience in the expert elicitation process to aid in the review of DOE use of expert judgment and possibly contributing to the development of guidance on expert elicitation.

Experts will be involved in the design and implementation of activities to understand present site conditions and to predict the behavior of the disposal site. Expert judgment may be used in: (i) setting priorities for data collection, (ii) designing site data collection activities, (iii) determining the level of resources for reduction of uncertainties, (iv) quantifying the uncertainty in numerical values for key parameters, (v) developing scenarios and assigning corresponding probabilities of occurrence, and (vi) formulating approaches for validating conceptual and mathematical models, as well as verifying computer codes.

This report presents a background for the use and elicitation of expert judgment in the HLW repository program. It begins with a discussion of expert judgment and its role in technical endeavors. The discussion then delineates situations that indicate the use of expert judgment. This report also discusses possible methods of obtaining expert judgments. While both informal and formal expert elicitation will be addressed, the emphasis will be on formal expert elicitation.

A section of the report is dedicated to quantitative probabilistic expert judgment because of its extensive historical use and expected continued use in the HLW program. The report presents criteria to assess the quality of a formal expert elicitation procedure and issues affecting the use of the results of a formal expert elicitation. A historical summary of the use of expert judgment in nuclear programs is presented, including coverage of probabilistic risk assessment, as well as potential future uses of expert judgment from both a DOE and NRC perspective. The report concludes with comments on the use of the results of a formal elicitation in support of the HLW program.

CONTENTS

CONTENTS (Cont'd)

CONTENTS (Cont'd)

CONTENTS (Cont'd)

FIGURES

ACRONYMS AND ABBREVIATIONS

BIOMOVS	Biosphere model validation
CDF	Cumulative Distribution Function
CDM	Compliance Determination Method
CDS	Compliance Determination Strategy
CNWRA	Center for Nuclear Waste Regulatory Analyses
DOE	Department of Energy
EB	External Bayesianity
EPA	Environmental Protection Agency
EPCs	Future events, phenomena, and conditions
EPRI	Electric Power Research Institute
FAA	Federal Aviation Administration
FEPs	Factors, events, and processes
HLW	High-Level Nuclear Waste
HYDROCOIN	Hydrologic code intercomparison
IAEA	International Atomic Energy Agency
INTRACOIN	Transport code intercomparison
INTRAVAL	International transport validation
LA	License Application
LARP	License Application Review Plan
LHS	Latin Hypercube Sampling
LLW	Low-Level Nuclear Waste
NEA	Nuclear Energy Agency
NRC	Nuclear Regulatory Commission
OECD	Organization for Economic Cooperation and Development
OSHA	Occupational Safety and Health Administration
PA	Performance assessment
PDF	Probability Density Function
PRA	Probabilistic risk assessment
PSAG	Probabilistic Safety Assessment Group
QA	Quality Assurance
R&D	Research & Development
SKB	Swedish Nuclear Waste Management Company
SKI	Swedish Nuclear Inspectorate
SNL	Sandia National Laboratories
TMI	Three Mile Island
UK	United Kingdom
UKDoE	United Kingdom Department of Environment
WIPP	Waste Isolation Pilot Plant
YM	Yucca Mountain

ACKNOWLEDGMENTS

This report was prepared to document work performed by the Center for Nuclear Waste Regulatory Analyses (CNWRA) for the Nuclear Regulatory Commission (NRC) under Contract NRC-02-93-005. The activities reported here were performed on behalf of the NRC Office of Nuclear Material Safety and Safeguards (NMSS), Division of Waste Management (DWM). The report is an independent product of the CNWRA and does not necessarily reflect the views or regulatory position of the NRC.

The authors wish to thank P. LaPlante and S. Spector for their contributions to this report. The authors also would like to thank M. Federline, N. Eisenberg, J. Buckley, J. Park, R. Winkler, and B. Sagar for their reviews of this document.

1 WHAT IS EXPERT JUDGMENT

1.1 BACKGROUND

The Nuclear Regulatory Commission (NRC) is developing methods to determine compliance with its regulation for the disposal of high-level nuclear waste (HLW) (10 CFR Part 60) (Nuclear Regulatory Commission, 1991) by the U.S. Department of Energy (DOE). Performance of HLW geologic repositories has to be assessed for a regulatory period of at least 10,000 years as previously stipulated in standards developed by the U.S. Environmental Protection Agency (EPA). Because of this extremely long period, a combination of experimental methods, studies of natural analogs, and mathematical models will be used to support technical analyses and performance assessments (PA) used to forecast the long-term behavior of the disposal system. Certain data gathered for this project will require interpretation and supplementation before it can be incorporated into these mathematical models. Expert judgment can have many applications in the HLW repository program, including data interpretation and supplementation. The NRC is currently evaluating the applicability of expert judgment to the HLW repository program licensing process. A fundamental concern to the NRC is gaining an understanding of when the use of expert judgment is warranted and what process of eliciting expert judgments is acceptable. Resolution of these concerns will contribute to the development of guidance in areas related to the use of expert judgment and its elicitation, and aid in the review of the DOE use of expert judgment.

Expert elicitation is a process to obtain data (e.g., judgments, opinions, and information) from subject matter experts related to a specific question, issue, or problem. This type of data, together with data collected by othe methods, contributes to the resolution of the question, issue, or problem of interest. Peer review, a part of expert judgment elicitation, evaluates the resolution of the problem. Expert elicitation can take a variety of forms and can be obtained using different levels of rigor and formalism. Informal expert elicitation is very common in the scientific inquiry process. Formal elicitation is often used when there is a desire to provide traceability as to, for example, the rationale and assumptions underlying an opinion. Formal expert elicitation is warranted when the question, issue, or problem of interest is complex, its solution can have a significant impact, there is a need for thorough scrutiny, and the issue or problem being addressed is potentially controversial. Increasing the level of formality in the expert elicitation is accompanied by a commensurate increase in the resources needed to implement the process. Regardless of the level of formality used in the expert elicitation process, it leverages the existing expert knowledge and cognition in a manner that captures the current state of knowledge about a given issue or problem (i.e., provides a statement of what is known and what is not known about a given question, issue, or problem).

1.2 PURPOSE AND SCOPE OF THE REPORT

This report discusses the background for the use and elicitation of expert judgments in the HLW repository program. It begins with a discussion of expert judgment and its role in technical and programmatic endeavors. The discussion then delineates some situations that indicate the use of expert judgment. This report also covers some possible methods of obtaining expert judgments. While both informal and formal expert elicitation will be addressed, the emphasis will be on formal expert elicitation.

A section of the report is dedicated to quantitative probabilistic expert judgment because of its extensive historical use and expected continued use in the HLW program. The report also presents criteria to assess the quality of a formal expert elicitation procedure and issues affecting the use of the results of

a formal expert elicitation. A historical summary of the use of expert judgment in nuclear programs is presented, including coverage of probabilistic risk assessment as well as potential future uses of expert judgment from both a DOE and an NRC perspective. The report concludes with comments on the use of the results of a formal elicitation in support of the HLW program.

The technical activities associated with the HLW repository program (i.e., site characterization, repository design, waste package design, repository PA, etc.) can be expected to follow normal and accepted scientific and engineering practice. As with all complex projects, time and resource constraints will modify the preferred technical approaches because of the absence of critical knowledge and data. And, like other technical programs, when other feasible options don't exist, the use of experts is expected to fill these information gaps.

Because of limitations in the state of science in many of the fields involved in modeling and long-term performance prediction associated with an HLW repository, the use of expert judgment elicitation in license support activities by the parties involved is expected. The issue of granting a license to construct, operate, and decommission and close an HLW repository has extreme public relevance, and the data and analyses contributing to the decision needs to be open, traceable, and unambiguous. The decision making process that will result in granting or denying a license involves the resolution of both scientific and policy issues. The scientific issues, such as the future climate, seismicity, and volcanism of any proposed repository site, are extraordinarily complex and significant to outcomes, and data supporting these predictions are relatively scarce and sources are sometimes controversial. Similarly, policy issues—such as the need for and consideration of human intrusion, the nature and implementability of the attendant regulatory requirements, etc.—present challenges, the resolution of which is neither straightforward nor uncontroversial. In addition, programmatic decisions associated with the HLW repository will require inputs from various sources including expert judgment. Finally, the documentation required in support of possible litigation is extensive and can only be adequately supplied by a formal process that emphasizes completeness in documentation. These circumstances suggest the use of formal expert elicitation to obtain the relevant judgments. Therefore, this report presents the process of eliciting expert judgments, research supporting the process, and issues in the use of expert judgments in areas associated with the HLW repository program.

1.3 DEFINITIONS

Like most areas of research, the literature on expert elicitation has seen a proliferation of terms to refer to methods of obtaining expert judgment. This section describes the terms that will be used in this report and the relationships between them. These definitions are a synthesis of the current state of thought on expert judgment and elicitation.

(i) **Expert Judgment.** The term expert judgment refers to the data or information that is produced through communication with an expert. Expert judgments can be evaluations of theories, models, or experiments or recommendations for further research. Expert judgments may also be numerical data that can be used in PA models or can be analyzed and interpreted. Expert judgments can be either qualitative or quantitative. Expert judgments can also be judgments about uncertain quantities or judgments about value preferences. An expert providing a probability distribution or a point estimate for a parameter are examples of judgments about uncertain parameters. The weather forecaster's prediction of a 20 percent chance of rain tomorrow is an example of such a

judgment. Value judgments can never be proven true or false. An individual's judgment that an automobile's features do not justify its high price, or a community collectively deciding that promotion of the arts is worth issuing bonds to pay for it are examples of value judgments.

(ii) **Expert Elicitation.** Expert elicitation is the process by which expert judgments are obtained. The degree of formality of the expert elicitation process is an attribute of the method and not of the expert judgment that is obtained. Expert elicitation often involves multiple participants who serve different roles in the process. The experts from whom the judgments are elicited are recognized as such in the field of interest. Normative experts design the elicitation process and have a background in decision theory, probability theory, and psychology. Generalists are typically project staff that help determine the types of judgments that need to be obtained in the elicitation based on their knowledge of the overall goals of the project and of the issues to be addressed by the experts. The more formal the elicitation, the more likely that each of these roles will be filled in the elicitation process.

(iii) **Formal Expert Elicitation.** Formal expert elicitation is a method that is differentiated from an informal process by various factors: (i) selection and decomposition of the problem, (ii) selection of experts, (iii) training and debiasing, (iv) conduct of elicitation sessions, (v) the degree of documentation provided, and (vi) the importance and expected use of the expert judgment produced by the process. In a formal process, the experts participate in a training and debiasing session prior to the elicitation of their judgments. The training and debiasing session prepares the experts for the form in which they will be asked to produce their judgments, clarifies all assumptions that will be shared by the experts, and identifies the biases that might affect their judgments during the elicitation. The experts often engage in exercises to increase their awareness of the biases and how they affect judgment. The training and debiasing session, in principle, increases the experts' ability to not only estimate values but to express their uncertainty in their estimates.

Formal expert elicitation emphasizes complete and comprehensive documentation of the basis for the experts' judgments as well as for every aspect of the elicitation process. The experts provide extensive rationale for their judgments, which is meticulously recorded and included with the analysis of the elicitation. This record allows a comprehensive evaluation of the data and information used by the expert in the judgment process.

Formal expert elicitation also differs from informal elicitation in the situations for which it is used. While both formal and informal processes can yield expert information, opinions, and evaluations, formal expert elicitation is used to collect expert judgments in those situation where the judgments are very important, such as for complex issues or problems that are controversial and for which little or no data exists. The formal process imparts greater confidence in the experts' ability to translate their knowledge and analysis of other sources of data into a representative judgment.

(iv) **Informal Expert Elicitation.** Informal expert elicitation refers to methods of obtaining expert judgments or expert information that do not involve such things as training and debiasing sessions with the experts or a prescribed review of the relevant literature by the

experts prior to the elicitation. In addition, informal expert elicitation can be characterized along a continuum by describing the level of documentation of the basis for the experts' judgments and the degree to which the process involved face-to-face, structured meetings of experts.

(v) **Peer Review.** Peer review consists of a technical review of specific scientific methods and results by qualified peers. It varies considerably in the degree to which it involves extensive documentation of the experts' judgments and whether or not it involves interaction among the participants. A peer review can range from an informal process with minimum documentation to a very formal process where consistency and thoroughness of the evaluation are extensively documented. However, peer review, in its most common form (i.e., pre-publication review of scientific articles), is often a fairly informal process.

1.4 EXPERT JUDGMENT

Expert judgments are types of data provided by an individual in response to given questions, issues, or problems. The questions, issues, or problems can be either technical or nontechnical. The data can be either qualitative, such as a list of possible scenarios, or quantitative, such as possible values of a parameter. Expert judgment is subjective, based on the knowledge, experience, and analysis of the expert. An expert is a peer-recognized authority who has an extensive background in the subject area.

The type of judgments required is driven by the problem being addressed. For example, if the judgments will be the basis for the development of a conceptual model, chances are that non-numerical judgments will suffice. A similar situation arises if the judgments will be used to construct scenarios for PA. On the other hand numerical expert judgments are needed to quantify various residual uncertainties associated with technical modeling and analysis and PA (Fehringer and Coplan, 1992).

In general, expert judgment can be viewed as a representation of the expert's knowledge at the time of response to a question. The expert's knowledge changes as new information is received. Because a judgment reflects the expert's knowledge and experience, an expert's judgments will differ over time and multiple experts can differ in their judgments. Experts have differing knowledge and experiences that may lead them to differing judgments in the area of interest. This can be used to advantage by soliciting judgments from multiple experts to capture the range appropriate for analysis.

Expert judgment is the product of knowledge-based cognition (Meyer and Booker, 1987). Cognition refers to how information is processed during problem solving and decision making. Knowledge-based cognition is analysis of new or uncertain problems or decision situations supported by an extensive body of knowledge. Experts, by definition, maintain a large body of knowledge, both factual and experiential, about their area of expertise.

All scientific research involves expert judgment in an informal manner at some level. It is applied every time a researcher makes decisions about what phenomenon to study, how to initialize a model, what variables are important to the process, what data are appropriate, how to process and use that data, and what interpretation to apply to the result. Expert judgments are a normal part of the scientific process and, under well-controlled and understood situations, such judgments can be validated or invalidated by the process itself.

Expert judgment is used widely to synthesize and interpret other types of data. It provides information when other sources, such as measurements, observations, experimentation, or simulation, are unavailable or prohibitively costly, or when data from these other sources are sparse, questionable, or only indirectly applicable. Expert judgment can support a wide variety of scientific activities in the HLW repository program, including scenario development, forecasting, model building, parameter estimation, interpretation of results, and determination of compliance with the pertinent regulatory requirements.

A representation of the current state of science or engineering can be provided by expert judgments. Like all other types of data, expert judgment can be misinterpreted, misrepresented, and misused, but thorough documentation helps mitigate this possibility. Expert judgment complements other sources of data, but is not equivalent to calculations based on universally accepted scientific laws or to extensive measured data on the quantities of interest. Expert judgments can augment information when measured data are scarce or lacking and are a snapshot of the state of knowledge of the individual expert. As new data become available, they are incorporated within the existing state of knowledge, possibly changing the judgments of experts in the area.

Expert judgments are used also outside the realm of science. They contribute to policy decisions in business, legal, and political sectors. Fehringer and Coplan (1992) refer to "decision-maker judgment" as needed to address regulatory significance of residual uncertainty in the HLW program. Expert judgment is a basis of decision making across a continuum from science and engineering to art and the humanities. It is used in all facets of human endeavor. As with all forms of data, expert judgments can be misused. One misuse of expert judgment is over-reliance on it as a basis for decision making. Expert judgments may contain considerable uncertainty, which should be documented and evaluated in the decision making process. Expert judgment may be misused as a justification to avoid gathering additional data.

In order for expert judgment to be subject to evaluation, the basis for the judgment must be well documented. Peers need to be able to trace the work from assumptions through interpretation of results and conclusions. There can be no gaps in the documentation so that an evaluation can be based on understanding of the work presented.

1.5 THE ROLE OF EXPERT JUDGMENT

Conceptually, expert judgments fit into programmatic and technically based endeavors in many ways. A general flow of technically based (scientific and engineering) analysis is shown in Figure 1-1. Ideally, the approach for a particular problem would be predetermined and acceptable data would exist to carry out each of the steps. While this is the preferred situation, this is often not the case, especially when the state of the science is advanced. In these cases, uncertainties can exist in many parts of this paradigm. For example, the approach, data sufficiency, appropriate models, and interpretation of the results are but a few of the areas that can be open to conjecture. Many of the areas involved in characterizing and solving complex problems are plagued by model inadequacies and data paucity. The investigator uses expert judgment to select an approach, design a data collection effort, choose among competing models, interpret the results, and draw conclusions. The documentation of the investigator's rationale for these judgments is quite variable and occasionally nonexistent. Depending on the problem with its attendant constraints, the complexity and criticality of an issue and time and resources available for data collection create a situation in which the other methods of scientific resolution are not possible. In these cases, formally derived expert judgments fill the gaps in the technical approach and allow the

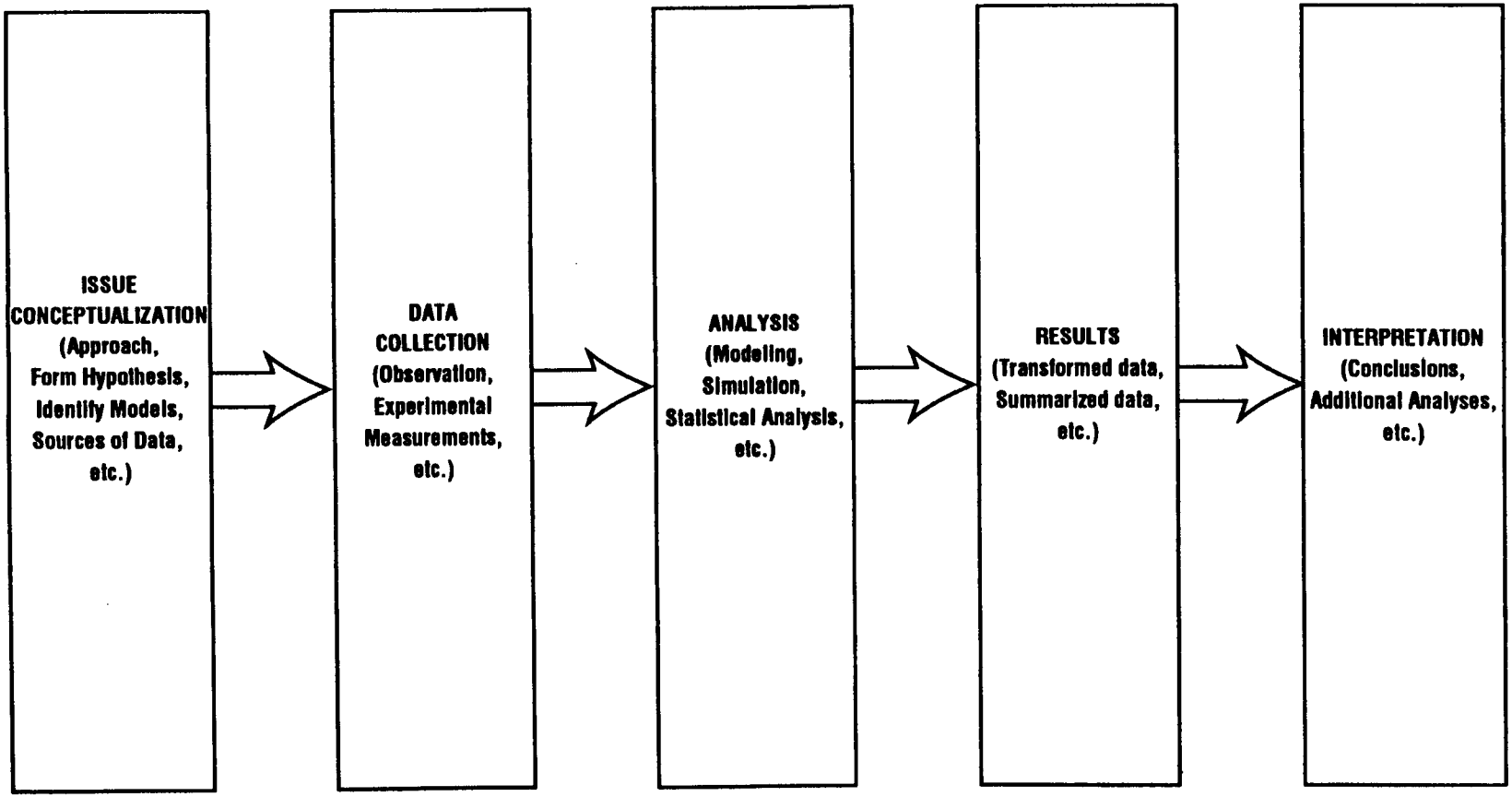

Figure 1-1. Technical problem solving paradigm

process to continue toward a solution.

Expert elicitation fills a similar role in the decision making/scientific process as other data analysis methods, such as modeling, simulation, or statistical analyses. Expert judgments differ from the other methods because the basis for them is generally less specified, that is, the analysis procedure cannot be described mathematically or algorithmically. Expert judgments can be used to generate conclusions or recommendations or they can be subjected to further data analysis, for example, used in modeling a phenomenon. As an analysis and data generation method, expert elicitation should be used only when needed, which is generally when other methods cannot supply data or analyses within time and resource constraints. It is, however, a valuable tool that has application in the HLW program.

1.6 CIRCUMSTANCES FOR COLLECTING EXPERT JUDGMENTS

The uses of expert judgments are pervasive and varied in any technical endeavor. They range from confirming the theoretical underpinnings of a model to driving the model directly by providing parametric bounds on the input. Expert judgment can also be used as a decision making tool for complex policy decisions. In this case, it is a method of summarizing existing data and theory pertaining to the issue and evaluating the risk of various alternatives. Expert judgment is used in all aspects of a professional enterprise.

1.6.1 Potential General Uses of Expert Judgment Associated with Data

In any technically-based program like the HLW repository, two general areas arise for the potential use of expert judgment which are associated with data. The first is in the assessment of what type and amount of data are needed to support the technical work in the program, and the second is in the synthesis, interpolation, and extrapolation of data.

1.6.1.1 Determining What Data to Collect and How to Collect It

Expert judgment can play a major role in determining the value of acquiring new data and can influence the decision of whether or not a specific type or amount of data should be collected and the manner in which the uncertainties should be mitigated if the decision is against collecting an amount or type of data Morgan and Henrion, 1990; Goodwin and Wright, 1991). The current state of knowledge generally serves as a basis to decide what type and amount of information should be collected and how it should be collected. Additional information can be gathered in a variety of ways: (i) collection of site—specific field data, (ii) collection of related generic-site field data, (iii) laboratory experiments, and (iv) modeling. Expert judgment may be important in selecting among the alternatives to obtain more information.

The use of the concept of the expected value of new information (Lindley, 1985a; Morgan and Henrion, 1990; Goodwin and Wright, 1991; Bonano and Thies, 1994) allows the evaluation of the various alternatives for data collection and selection of the one that is preferred, given all the considerations that go into the decision. The expected value of new information is a means to determine how new information will contribute to the resolution of a given issue or problem, and whether the benefits reaped by using the new information will outweigh the cost of obtaining it. This decision analysis problem (i.e., how much and by what means should new/additional information be collected) relies on both quantitative and value judgments to balance the advantages and disadvantages of the available different alternatives

(i.e., collecting a specific type and amount data versus collecting a different type and amount of data, and supplementing with a synthesized set of existing data).

An important decision which can rely on expert judgments is the evaluation of the potential impact on the solution of the problem if it is decided not to collect a given type or amount of data based on feasibility considerations, such as destruction of the specimen (i.e., data gathering which would alter the setting to such an extent as to render a candidate site unusable). Expert judgments can also be used in the identification of alternatives that can mitigate such impact.

One alternative to data collection in this situation is the identification and implementation of physical modifications to the engineered components of the system, more commonly known as engineered alternatives, and experts can become involved in such exercises. For each engineered alternative considered, there are complementary construction decisions to be made. All these decisions affect the repository performance and may involve crucial expert judgments that weigh performance against the costs of implementing the alternative and benefits that can be realized in terms of the demonstration of compliance. Expert judgments used in these decisions are likely to be both quantitative judgments about uncertainties and value judgments. Some experts may provide quantitative information regarding the expected performance of each alternative, while others may be involved in the design and implementation of the process to evaluate the alternatives.

The situation may arise when experts are used to develop information about a given technical issue that cannot be reasonably addressed in any other way. If expert judgments are sought on the distribution of the numerical value of X, the question posed could be "given the available information about the value of X, what is the probability density function (PDF) that best represents the uncertainty about its value?" If expert information about the uncertain parameter X is desired, the question posed could be "what information and experience do the experts have that is directly relevant to the estimation of the value of X?"

Value judgments can be used to aggregate the various impacts for each of the alternatives. Due to the uncertainties regarding those impacts, some of these value judgments can address risk attitudes concerned with those uncertainties and, because there are multiple objectives, some of these value judgments may address tradeoffs among the different objectives.

1.6.1.2 Extending the Utility of Existing Data

Expert judgment, although generally informally, is prevalent in all aspects of data collection, processing, and use. Situations can occur where the utility of an existing data set can be significantly enhanced by the use of expert judgment in the selection of appropriate technical mean (processing algorithms and statistical techniques) for interpolation or extrapolation. Expert judgment can also be used in certain circumstances to directly synthesize, interpolate, or extrapolate a sparsely populated data set.

Two such circumstances are the cases of destructive testing and forecasting far into the future. In the first case, continued gathering of data will likely lead to significant damage and alteration of the valuable specimen, rendering it useless for its intended application (e.g., see section 5.2.1.3 for a discussion of the potential damage to a repository site if numerous boreholes are drilled to characterize the geology completely). In this case, expert judgment can be used to either recommend statistical techniques to link existing data points or to directly synthesize and interpolate between data points to supplement the data set needed for analysis.

The second case requires forecasting of situations and technical performance into the future far beyond where modeling or other technical approaches have been validated. Expert judgment can be used to synthesize existing data with historic and causal bases to extend the data into the future.

In both cases described, it is incumbent on the scientists and engineers to document the process used and the specific data obtained by expert judgment. Using expert judgment to synthesize, interpolate, and extrapolate data which is otherwise unobtainable is a valuable and appropriate use of expert judgment. But this use of experts to supplement data should be used sparingly and not as an excuse to proceed with analysis when difficulties with cost and schedules arise.

1.6.2 Possible Uses of Expert Judgments in a Nuclear Waste Program

Potential areas have been identified for which expert elicitation might be used in the HLW program (Bonano and Apostolakis, 1991). These areas, which will be discussed in more detail in Chapter 5, are scenario development, forecasting, model development, parameter estimation, and information gathering.

1.6.2.1 Scenario Development

Identification of plausible events and processes is an area dominated by expert judgment. No method has been demonstrated for arriving at a unique exhaustive list or for objective evaluation of such a list. Classification of the identified events also requires subjective expert judgment. The categories of classification must be selected and the events and processes assigned within the classification scheme. Cranwell et al. (1990) outline a methodology for selecting and screening scenarios for possible future states. Scenario selection and screening involves a number of phases: (i) initial identification of plausible events and processes, (ii) classification of events and processes, (iii) initial screening to remove unimportant events or processes, (iv) combining of events and processes into scenarios, and (v) screening of scenarios to arrive at a final set for analysis. This can be used to provide estimates on new, rare, complex, or otherwise poorly understood phenomena.

Kahn and Wiener (1967) describe the process of scenario analysis which is extremely expert judgment dependent. Scenario analysis is concerned with the generation of sequences of events for the purpose of focusing attention on causal processes and decision points. This creates a focus on the step-by-step process of achieving some final event or situation and the alternatives that exist at each step for entities to facilitate or hinder the process. The analyst identifies a set of long-term trends that are relevant to the issue under investigation and extrapolates them into the future, using any theoretical or empirical knowledge that might influence the extrapolation. The result is what is called the surprise-free scenario. Alternative scenarios or canonical variations are defined by varying key parameters in the surprise-free scenario. The probability of the various scenarios are not explicitly taken into account. In fact, the surprise-free scenario may be judged to have a very low probability, no greater or less than any variation. The focus is on elucidating basic trends rather than attempting to predict future events.

1.6.2.2 Forecasting

Forecasts are generally represented as probabilities and can be collected through probability elicitation. Probability elicitation is a special, but perhaps most common, case of expert elicitation that focuses on collecting subjective probabilities or probability distributions. In this view, a probability is

interpreted as an individual's degree of belief based on knowledge (i.e., an educated guess) that an event will occur. This differs from situations in which the relative likelihood of outcomes (e.g., the roll of a die) can be determined or in which there is empirical evidence describing the relative frequency of events (e.g., the frequency of accidents at an installation). Formal elicitation techniques have been devised to collect and interpret subjective probability estimates. A conclusion of a workshop sponsored by the Nuclear Energy Agency in 1987 (Nuclear Energy Agency, 1987) is that analysis of HLW repositories will be characterized uncertainties that are quantified as probabilities representing degrees of belief based on knowledge.

Subjective probability distributions reflect the expert's knowledge about possible values of an uncertain quantity or random variable. The median of the distribution serves as a useful location parameter; the value of the variable is just as likely to fall above the median as below it. The spread of the distribution indicates the expert's certainty associated with the variable, greater spread indicating less certainty and vice versa. Once elicited, subjective probability follows the usual mathematical rules of probability theory. (Ramsey, 1931; Savage, 1954; Kyburg and Smokler, 1964).

1.6.2.3 Model Development

Building a model of a phenomenon requires extensive expert judgment. Expert judgment is used to select and interpret data, to specify the model including all simplifications and assumptions, and to validate the models.

The objectives to be achieved by developing the models need to be clearly specified. These objectives need to be related to the fundamental objectives of the project. Expert judgment can be used to answer questions about how model development will contribute to better understanding and better decision making. Once the objectives are specified, general types of alternative models worth developing should be specified. They should be screened to generate a list of the most useful. The screening criteria should be well specified using judgments of experts. After models are selected, they are developed.

Experts must select data to use in modeling that are obtainable and reflect the processes of interest. The experts should select the sources of data, then screen out unimportant sources and retain the most relevant ones. Once the data are selected and screened, experts must interpret them. The experts make inferences from the data that form the basis of the conceptual models. The data may be qualitative, quantitative, or some combination thereof. The expert can be used to interpret different sources of qualitative data or to integrate quantitative and qualitative data to make a judgment.

1.6.2.4 Parameter Estimation

Parameters are coefficients or constants of models and processes that describe or control the behavior of a model. Parameter estimation involves selecting appropriate numerical values for parameters and quantifying their uncertainty. First, important parameters must be identified and then the uncertainty in their values quantified.

The identification of parameters governing models and processes involves substantial expert judgment. The experts suggest how the important parameters should affect the events and processes of interest. Once parameters are identified, their importance within a model can be determined by a sensitivity analysis.

It is frequently necessary to quantify the uncertainty in the values of the input parameters for the models used. The uncertainty can be expressed in a number of ways. One way is to estimate the mean and variance of a distribution of possible values. Another is to estimate the range of possible values and to assess a PDF covering that range. This method is conventionally used in PA analyses because it provides a rather complete description of the uncertainty.

1.6.2.5 Information Gathering

Additional data can be gathered from experts to supplement the current state of knowledge and to improve problem solving and decision making. With any information gathering problem, the key is to precisely specify the objectives of the exercise. The objectives of the activity need to be clearly specified and reasonable alternatives proposed for the type of information that can be obtained. The alternatives are then screened to identify competitive options. For each of the competitive options, it may be necessary to estimate the information that can possibly be learned. An analysis can indicate the relative desirability of the options based on assumptions specified by project personnel.

Formal expert elicitation can be used to specify the goals and objectives of laboratory and field experiments. Different objectives drive the collection of different types of information. Formal expert elicitation can also be used as a method of characterizing expert problem solving processes. This type of expert elicitation has been used extensively in the area of artificial intelligence and is important when expertise is rare and the expert's problem solving strategies are needed to improve current practices, to train others, or to create software systems that provide expert advice. In this case, the normative expert encourages the expert to verbalize every thought during the problem solving process. The techniques and strategies used by the expert can be inferred from this record.

1.7 THE VALIDITY OF SUBJECTIVE DATA

Expert judgment can be considered a type of data. Expert judgment includes the expert's answer to a question and all supporting analyses used by the expert to reach the conclusion (Meyer and Booker, 1991).

Expert judgment is often considered soft data in comparison to data that has been measured or obtained from observations or instruments. Ericsson and Simon (1984) argue that data can be regarded as soft to the extent that they incorporate the assumptions and interpretations of observers. Therefore, the softness of the data is not a fixed attribute of the source, but something the researcher can exercise control over. The researchers can strive to keep their interpretations of the data separate from the data itself and strive to make a minimal number of assumptions, but in practice this is not often achieved.

Some people question whether gathering and analyzing expert judgment can be considered scientific. In the sense that it follows the basic tenets of scientific inquiry—observation, hypothesis formation, and experimentation—expert elicitation can be considered to be scientific. However, definitions of science often include the notion of reproducibility of results to some arbitrary level of fidelity. One of the underlying principles of formal expert elicitation is that each expert is internally consistent. That is, given the same information, each expert should be able to reproduce his/her judgments, and related judgments should be consistent with each other. Expert judgments should be updated when new information becomes available, but the update should be effected carefully to ensure consistency; for example, when updating the probability of occurrence of a given event due to new information, internal

consistency means that the expert needs to simultaneously update the complement of the event. Formal expert elicitation increases the likelihood of internal consistency by identifying and forcing the resolution of potential inconsistencies.

Even so, format expert judgment can be collected through procedures that are based on sound experimental results from the social sciences. Variability in judgments may be introduced into the process because each expert brings his or her own education and experience to the table. Thus, even when presented with identical data, different experts may reach different conclusions. This does not necessarily invalidate the process; it simply provides a range of possible responses based on the expertise existing in the field as a whole. Also, expert judgment is usually sought in cases where there is an absence of standards and well-developed theories, which can lead to variability of responses.

Expert judgments are both valid in their own right and comparable to other data. All data are imperfect representations of reality. The validity of expert judgment data, like any data, can vary based on the procedures used to collect it. So-called "hard" data, such as data taken from instruments, cannot be considered to be perfect because of problems such as random noise, equipment malfunction, operator interference, data selection, or data interpretation. The validity of all data varies. The validity of expert judgment depends heavily on the quality of the expert's cognitive representation of the domain and ability to express knowledge. The elicitation of expert judgment is a form of data collection that can be scrutinized; the use of the judgments can, and should, also be scrutinized. However, there are different sets of rules for doing so relative to "hard" data.

1.8 SUMMARY

Expert judgments contribute to many facets of scientific and technical, as well as programmatic, endeavors. As such, their validity and reliability are critical issues in the application of expert judgment in the regulatory and public policy arena. The more scrutable the source of the expert judgment the more acceptable the application will be. Areas in which expert judgments are used, and should be used, need to be identified and techniques for the use of expert judgments need to be formalized.

Expert elicitation is often used to obtain expert judgments of uncertain quantities and probabilities of potential events in order to assess risks. As such, it plays a role in scientific and engineering activities. It can also be used in a variety of ways, from identifying and selecting preferred technical approaches to obtaining supplemental data for a wide variety of substantive issues. Expert judgments can be quantitative or value judgments; both types have played, and will continue to play, a significant role in the HLW repository program. Expert judgment should not be viewed as a substitute for experimentation, modeling, and observation. Instead, it can complement and supplement these activities under appropriate conditions.

The following chapters discuss when, why, and how expert judgments can be used both in general and specifically in relation to the HLW program. They also lay out issues that should be considered in the collection of expert judgments and describe the implications of how those issues are resolved. The report contains an evaluation section for distinguishing the quality of expert judgments based on the quality of the elicitation procedure.

2 CONSIDERATIONS IN ELICITING EXPERT JUDGMENTS

2.1 EXPERT JUDGMENT ELICITATION

Expert elicitation, the process used to collect expert judgment, varies along a continuum from informal to formal. Formal procedures actively promote reducing of bias, providing consistency and thorough documentation for the expert judgments. These procedures have evolved over the course of the last 30 years. Formal methods are useful but have some drawbacks in practice. This chapter describes the history of expert elicitation, issues in the uses of formal expert elicitation, methods of eliciting expert judgments, and a study that applied a formal expert elicitation to obtain data relevant to HLW repository performance.

The difference between formal and informal expert elicitation revolves around the selection and decomposition of the problem, selection of experts, training and debiasing, conduct of the elicitation sessions, and documentation of the process. Expert judgment is part of all aspects of the scientific, technical, and decision making processes, yet frequently its use is unidentified, implicit, or assumed. That practice is pervasive and simply a well accepted part of the enterprise (e.g., a researcher's selection of a model exercises his or her judgment). However, when expert judgments are used to make high risk or critical decisions, the source and justification for those judgments needs to be more rigorously evaluated and subject to the scrutiny of interested parties. To those ends, procedures have emerged that differ in their degree of documentation and preparation of the experts to make the required judgments. These procedures vary from the informal (e.g., informal expert panels) to the formal (e.g., a formal procedure to elicit probabilistic judgments).

The decision to use a formal elicitation process should be carefully considered to ensure that its benefits outweigh its costs. The following circumstances suggest that a formal elicitation procedure is appropriate (Bonano et al., 1990).

- Importance of the Issues. An issue that is likely to receive close scrutiny and criticism should employ formal elicitation. These types of issues are often public policy decisions that affect the general public's health and safety. Due to the nature of the issue, the procedures used to collect any and all data need to be precisely specified. A formal elicitation enhances the transparency of the procedure and improves communications, making it useful when the issues addressed are important.

- Complexity of the Issues. The more complex the issues or the team assembled to address them, the more necessary the formal elicitation process becomes. Formal elicitation explicitly describes the issues and procedures and allows for superior coordination of a group process.

- Otherwise Unobtainable Data. Expert judgment can be used to supplement existing data or summarize the state-of-knowledge until data can be collected. Use of formal expert elicitation to provide needed data is often indicated where data cannot be obtained by other preferred means (e.g., collecting of field data or laboratory experiments). Examples of situations for which the only feasible data source is often expert judgments are forecasting far into the future and destructive testing (see section 1.6.1.2).

- Level of Documentation Required. The documentation generated by a formal process is more complete and consistent and therefore defensible because there is greater attention to the procedure, assumptions, and findings than in an informal elicitation process. For issues that are important or can have potentially serious consequences, this level of documentation may be crucial for acceptance of the outcome of the study.

2.2 PEER REVIEW

Much of scientific and engineering development is subjected to the normal peer review process of critical evaluation by colleagues in various venues. A peer review is a documented, critical review to evaluate the acceptability and adequacy of the research, performed by peers who are independent of the work being reviewed. A peer review can be conducted by obtaining input separately from a number of peers or by convening a panel to conduct the review. The intent of convening a panel might be, for instance, to accelerate the natural peer review process associated with scholarly publication and presentations that can take months or years for feedback to occur. Also, discussions among the panel members can generate useful information not available from a set of independent reviews. The most common peer review process (i.e., pre-publication technical review of a scientific article) typically uses informal expert judgment to evaluate scientific methods and results. However, in principle, the nature of peer review is sufficiently flexible so that its rigor and formality is commensurate with the study being reviewed. For example, the review of technical documents and reports in the HLW program can be considered a semi-formal peer review under strict quality assurance (QA) requirements. Peer reviews can also be conducted using a formal process to review the solution of problems of high importance. Formal peer review has the same basic attributes of the formal expert elicitation process.

The peers are recognized experts in the domain of interest as evidenced by their scientific qualifications. The peers comment on the validity of the assumptions, the appropriateness and limitations of the methodology and procedures, the accuracy of the calculations, the validity of the conclusions, and the uncertainty of the results and consequences of the work. They may also offer alternative explanations of the results and comment on the adequacy of the information and data used to obtain them.

The peer review process requires the expert judgments of peers. However, it is important to note that peer review as an expert judgment process is different from the formal elicitation of expert judgments in the context of this report. The reference to expert judgment herein denotes judgments, opinions, or information provided by subject matter experts to contribute to the solution of a given problem. On the other hand, peer review refers to judgments by subject matter experts to evaluate the solution to the problem. In this context, expert judgments can, and should, be the subject of peer review. The differences between the elicitation of expert judgment and independent peer review notwithstanding, both processes contribute in a positive way to enhancing the quality of the solution of a problem.

2.3 FORMAL EXPERT ELICITATION

Formal expert elicitation refers to a structured procedure designed to gather judgments about an issue or problem from experts. Topics of focus of elicitations include programmatic issues, probability encoding, scenario development, and model selection. The elicitation procedure generally has the participation of three parties: normative experts, generalists, and experts. Generalists are usually members of the project team with a good grasp on the issue that is being addressed in the elicitation, and with a good understanding of the overall problem and the manner in which the judgments will be used. Experts,

often called the specialists or panelists because it is their judgment that is elicited formally, are individuals with substantive and recognized knowledge on the subject matter of interest. In other words, experts should have a proven and accepted track record of research and applications in the required technical field or discipline. Normative experts are individuals with the necessary expertise and experience in the elicitation of expert judgments. Normative experts have a suitable background in decision theory, probability theory, and psychology. Of particular importance is the selection of normative experts that are well versed in the identification and compensation of biases, and in the consideration of dependence among experts when using multiple experts. The elicitation procedure includes methods to select and train the experts as well as a documented approach to gathering the appropriate knowledge.

Generalists interact with experts during training and could assist normative experts during the elicitation. Generalists, jointly with the normative experts, are usually the first individuals to examine the judgments provided by the experts and provide feedback to the elicitation regarding the judgments. Generalists, oftentimes referred to as analysts, can play a key role in the aggregation and use of the judgments, such as in the study described by Hora and Iman (1989). Finally, generalists can also be used as experts and their judgments elicited, such as in the study by Merkhofer and Runchal (1989). Experts are the individuals whose judgments are sought in the elicitation process. The normative expert with the help of the generalist elicits judgments from each expert according to this procedure and carefully documents the expert's rationale.

For large programs or for specific projects with many disciplines represented, there is a need for the individual investigators to share the results of their expert judgments so that issues such as modeling consistency and data accuracy/compatibility can be addressed across the entire effort. Formal expert elicitation is designed to meet this need.

Compared to an informal expert elicitation, formal expert elicitation increases the scrutability of the judgments and enhances the communication of the results. The normative experts carefully document the procedure used to obtain the judgments and the rationale for each judgment, making the process readily available for external evaluation. Formal elicitations also improve the calibration of probability judgments (Lichtenstein et al., 1982; Fischhoff, 1982). The training incorporated in the formal process ensures that the participants are aware of the cognitive and motivational biases that can influence expert judgments and teaches them how to try to avoid them. Accuracy is improved because the problem definition is clear and the issues addressed are explicitly stated. Finally, formal expert elicitation increases the consistency of procedures across studies, facilitating study comparisons.

There are drawbacks to using the formal expert elicitation process—chief among them is the relative cost. Expert elicitation is resource and time consuming, requiring many staff hours from both the expert and the normative expert. It also takes a considerable amount of time to establish and implement a formal process. Informal elicitation generally does not require organizing and scheduling participants for multiple sessions, all of which involve substantial effort. Also, compared to informal elicitation, formal elicitation may make the process somewhat more cumbersome and less flexible. Formal methods promote accurate, consistent, and efficient collection and processing of expert judgments. Increased reliance on expert judgment in a study indicates the need for a formal process. The cost of the formal process can be spread across multiple judgments, making it less expensive.

2.3.1 History of Formal Elicitation

The notion that the speculations of experts could be a significant input in a structured decision process dates to the establishment of RAND corporation in the 1940s. At this time, structured expert opinion was conveyed to decision makers through either scenario analysis or the Delphi method. Both were developed at RAND corporation.

The Delphi method was developed originally at RAND (Dalkey, 1967; 1969) in the early 1950s and has undergone many variations. The goal of a Delphi method is to achieve a high degree of consensus between highly knowledgeable experts on the issues in question. The process begins when the assessment team defines a set of issues and selects knowledgeable experts who are the respondents. The respondents typically do not know who the others are and all responses remain anonymous. A questionnaire is sent to the respondents and their answers are analyzed. The raw scores and analyzed results are returned to the respondents, who have the opportunity to revise their initial predictions. Respondents whose answers remain outside the interquartile range for a given item are asked to give arguments for their prediction. The revised predictions are processed again and arguments for the outliers summarized. This information is sent back to the respondents and the process is iterated. A Delphi exercise typically involves three or four iterations. Generally, the responses on the final iteration show a smaller spread than the responses on the first iteration.

The development of Bayesian methods in the 1950s and 1960s led to work on probability elicitation, especially exploring techniques for encoding of an expert's uncertainty for specific events (Schlaifer, 1959). This work helped solidify the benefits of a formal process to assess probabilities.

The next step in the development (Spetzler and Stael von Holstein, 1975) of the expert elicitation process was the Stanford/SRI assessment protocol in the 1960s and 1970s. This protocol has been extremely influential in formal expert elicitation and is the forerunner of the currently accepted approach to formal expert elicitation. The basic Stanford/SRI interview process consists of five phases: motivating, structuring, conditioning, encoding, and verifying. During the motivating stage, the analyst describes and justifies the reason for the expert elicitation to the expert. At this point, the analyst also determines whether or not the expert has a motivational bias that would prohibit participation.

The structuring phase of the elicitation involves arriving at an unambiguous definition of the quantity or quantities to be assessed and stating them in the form in which the expert will most likely be able to provide reliable judgments. It is important to determine whether there are any conditioning factors that may influence the value of the quantity and to reach agreement on what the conditioning values examined in the elicitation will be. During this process, the form of the expert's judgment should be determined.

During the conditioning phase, the objective is to get the expert conditioned to make the judgments and avoid cognitive biases. The expert is asked to explain how he or she makes the probability judgments, including what data sets and other information are to be referenced and how they will use this information. The expert is encouraged to consider other ways of examining the problem and the analyst looks carefully for possible sources of cognitive bias.

The fourth stage of the protocol is the encoding stage. At this point, the analyst and expert work together to produce probability estimates. During this phase, the analyst may discover that there is a need

to return to one of the previous phases, particularly to the structuring phase. The objective of the final verifying phase of the protocol is to provide feedback to the expert about the judgments to see if they correctly reflect the expert's beliefs. The results can be plotted and the expert can make corrections.

The basic Stanford/SRI assessment protocol was refined in the 1980s and 1990 by applications in probabilistic risk assessment (Humphreys, 1988; Otway and von Winterfeldt, 1992; Whitfield and Wallston, 1989; EPRI, 1986. Winkler et al., 1994) and multi-attribute decision making (U.S. Department of Energy, 1986; Keeney, 1992; Keeney and von Winterfeldt, 1989; 1991) culminating in the description of the steps in a formal expert elicitation delineated by Bonano et al. (1990) and DeWispelare et al. (1993).

2.4 STANDARD SETTING AND INFORMAL PANELS

The nature and complexity of issues that need to be dealt with in radioactive waste management present tremendous challenges that, more often than not, cannot be resolved using conventional scientific approaches and axioms. A radioactive waste disposal system is so complex that all the data needed to predict its long-term performance with certainty are not going to be available; instead, decisions regarding the performance of a disposal system will be made based on quantitative and qualitative assessments employing the information that is available. The incompleteness of the information base makes forecasting such performance an underdetermined problem; that is, there will be multiple interpretations that are consistent with the available information. In addition to the scientific and technical difficulties, the management and disposal of radioactive waste has considerable economic and social implications that also need to be factored into the solution of the problem.

One common approach used, both nationally and internationally, is to convene groups or panels to address a given issue or problem. Often, the objective of such panels is to provide a forum for the exchange of ideas about the resolution of the issue. However, many of these panels have also been used: (i) to reach consensus on the resolution of an issue or problem, or (ii) to review approaches being used or suggested to resolve the issue or problem. While the information gained from these panels can be very useful, extreme care needs to be exercised as to the manner in which technical staff uses this information. Regardless of the excellent credentials and reputation of the members of these panels, the judgments (e.g., opinions, recommendations, or suggestions) that the members of a panel offer individually or collectively could have significant impacts on the results of analyses of the performance of the repository system. It should be recognized that many of these judgments are provided using less than a formal elicitation process.

The International Atomic Energy Agency (IAEA) and the Nuclear Energy Agency (NEA) of the Organization for Economic Cooperation and Development (OECD) have established many groups consisting of technical individuals directly involved in radioactive waste management programs in member countries. For example, the IAEA in the late 1970s established a group, the charter of which was to develop a comprehensive list of initiating features, events, and processes as a starting point for scenario development. The NEA has sponsored a large number of groups, such as the Scenario Working Group, the Probabilistic Safety Assessment Group (PSAG), and the Human Intrusion Working Group, among others. The Scenario Working Group reviewed and formed opinions about the different approaches for handling the uncertainty associated with predicting the future states of the repository system. The PSAG conducted a number of intercomparison exercises for a number of PA computer codes used by OECD

member countries. The Human Intrusion Working Group examined technical and policy issues associated with the consideration of human intrusion into radioactive waste repositories.

Organizations in other countries, most notably the Swedish Nuclear (SKI), have formed international groups, similar to the PSAG, to tackle technical issues associated with prediction of repository system performance. SKI, for example, since the early 1980s has sponsored the INTRACOIN (transport code intercomparison), the HYDROCOIN (hydrologic code intercomparison), and INTRAVAL (international transport validation) studies and the GEOVAL conference. Other international code comparison studies are BIOMOVS (biosphere model validation) and DECOVALEX (coupled code validation). The insights gained from participating in these groups is very useful as a means to build confidence in the computational models used in PA and other technical analyses. However, it should be recognized that, while the intercomparison exercises can make positive contributions to confidence building, these exercises can also easily lead to a false sense of confidence that can be very problematic later.

Areas such as confidence building exercising of computational models and establishing lists (e.g., IAEA's list of initiating features, events, and processes for scenario development; IAEA, 1981), which are relied on as de facto validated standards and references, are useful to technical staffs. However, the documentation of many judgements contained in these national and international groups and panels is often incomplete. These judgements are often provided in open forums employing less than a formal elicitation process. The basis of these judgements should be examined before they are used.

Caution should be exercised in basing technical and programmatic approaches and analyses on national and international groups and panels because of generalizations (often espoused as validated truth) which are frequently derived from limited technical work and undocumented informal expert elicitations.

2.5 ISSUES IN THE FORMAL ELICITATION OF EXPERT JUDGMENTS

The decision regarding the use of expert judgments and the approach selected for the elicitation of the judgments are likely to get as much scrutiny as the judgments themselves. Criteria that could be used to decide whether the use of judgments is justified and whether or not a formal process is needed to obtain the judgments are discussed in Chapter 4. In this section, some key issues associated with the design of the elicitation process are discussed. These issues are:

- Defining the issue, problem, or question of interest
- Problem decomposition
- Selection of experts
- Single versus multiple experts
- Information provided to, and use of available information by, experts
- Form of judgments
- Elicitation training and calibration
- Elicitation of judgments
- Dependence of experts
- Debiasing
- Documentation

Review of the relevant expert elicitation literature suggests that these issues are the most important ones; however, this does not necessarily means that they are the only ones that need to be addressed. The resolution of these issues and, consequently, the design of the elicitation process, is based on subjective judgments. Therefore, there is a need to carefully address these issues in a thorough and documented fashion.

It will become evident, from the discussion of these issues, that many of them are closely interrelated; the decoupling of the issues as presented here was an arbitrary decision by the authors to facilitate the discussion. The same issues could have been discussed using a different approach.

2.5.1 Defining the Issue, Problem, or Question

Defining the objectives and goals of the study requires an understanding of how the data collected in the expert elicitation will be used in subsequent analyses. This understanding dictates the overall content goals of the study, the nature of the expertise that will be required, and the assumptions that will be shared. This is a critical step that requires input from generalists who are familiar with the data needs and uses. Without a succinct definition of the issue, problem, or question to be addressed and the assumptions to be shared by the experts, designing the elicitation is impossible; the issue statement drives the choice of experts, the information provided to them, and the form of the judgments that will be required. For example, one problem encountered by DeWispelare et al. (1993) when attempting to behaviorally aggregate predictions of precipitation in the Yucca Mountain region (YM) was disagreement on the current level of precipitation in the region. The experts had assumed different current values, which led to considerable disagreement in their attempt to aggregate their estimates.

Early in the process of issue definition, the focus should be on broadening the issues considered. This ensures that as many issues have been identified as possible, and that those not included in the final issue statement are rejected based on sound consideration. Once a list of issues is developed it should be screened so that it includes the most probable and the most important issues that enhance the likelihood of achieving the project objectives and goals. The criteria used to select the issues should be explicit and should be reviewed by the experts at the initial meeting of the entire group. A clear statement of the issues and assumptions should be presented to the experts and refined based on their feedback. The project team generally assembles the initial statement by consulting the users of the data. The expert may make a number of changes but any changes must be iteratively evaluated by the project team to assure that the issues and assumptions conform to the need for the data in subsequent analyses.

The definition of the issues should be precise; clarity of the issues is critical for the elicitation design. The issues can range from general to specific and from simple to complex. For example, a general question might be "what are the primary climatological controls operating at YM?" A more specific question might be "what is the likely average annual precipitation amount at 3,000 years after the present (AP) in the vicinity of YM?" The generalists and normative experts should propose initial conceptual models and scenarios for the processes of interest that will be reviewed by the experts at the first meeting.

This process also involves formalizing the quantities that will be elicited and stating all assumptions that may influence the judgments. Winkler et al. (1978) state that all questions asked of the experts should be about observable or at least theoretically observable quantities.

2.5.2 Problem Decomposition

Conventionally, complex problems are broken up into smaller and simpler components to facilitate the solution of the problem. The tenet of problem decomposition is that the solution of the smaller components is more tractable than that of the entire problem. Problem decomposition related to the elicitation of expert judgments is advocated by many (e.g., Bonano et al., 1990; Chhibber et al., 1992; Hora and Iman, 1989; among others) as the vehicle to increase the likelihood that the judgments are focused on issues that the experts may be more familiar with. Ravinder et al. (1988) provide a theoretical argument for the superiority of decomposition-based assessments. Morgan and Henrion (1990, p. 116) state that "It has become an article of faith in the decision analysis community that disaggregation of an elicitation problem holds the potential for significantly improved performance.... ." However, as will be discussed later, these authors describe a study, the results of which challenge this general belief. In principle, once judgments have been obtained for the different components of the problem, the judgments can be aggregated to arrive at judgments applicable to the entire problem. Problem decomposition offers both advantages and disadvantages.

Among the advantages of problem decomposition in the elicitation of expert judgments (Armstrong, 1985; Chhibber et al., 1992) are: (i) experts deal with smaller, simpler issues; (ii) judgments can be better focused; (iii) users of the judgments are able to examine and evaluate judgments better due to improved documentation; and (iv) it facilitates refinement of judgments in light of new or additional information. Armstrong (1985) claims that decomposition is conducive to improved judgments in situations involving high uncertainty. Chhibber et al. (1992) suggest that decomposition offers a compensation mechanism by which errors made in one component of the system can be compensated by errors made in another component. These authors, however, caution that such reliance on chance as the means to improve the accuracy of the judgments is ill advised.

Morgan and Henrion (1990) describe a study aimed at elucidating the impact of decomposition (or disaggregation, as these authors refer to decomposition) on the accuracy of judgments offered by a group of individuals. Their results show that a poor decomposition could have detrimental effects on the accuracy of the judgments. Some basic issues which arise in problem decomposition (Chhibber et al., 1992) are: (i) the level of complexity of the components or parts of the problem, (ii) the thought or reasoning process used to decompose the problem, and (iii) the interdependencies between the different components.

Before deciding on the level of complexity desired for the components of the problem, it may be advisable to conduct a cost-benefit analysis that attempts to assess, quantitatively or qualitatively, the cost associated with increased accuracy of the judgments. It is possible that a point of diminishing return is reached, beyond which large increases in cost due to increased decomposition into smaller parts only yields a small improvement in accuracy from one level of decomposition to another.

Decomposing a problem is likely to make a difference if aggregation of judgments offered by different experts is needed. This was indeed the case in the NRC reactor safety study, better known as NUREG-1150 (Hora and Iman, 1989). A group of experts was assembled to offer judgments that were used to develop cumulative distribution functions (CDFs) for uncertain parameters for input to a probabilistic risk assessment. Each expert was allowed to use his or her own decomposition of the problem. As a result of that decision, each expert's decomposition was unique, and this forced the use of a variety of aggregation techniques to arrive at the desired distribution functions.

It may be advisable for the project team to decompose the problem, present it to the experts, and refine the decomposition based on feedback from the experts. This would eliminate the possibility of multiple decompositions. However, the danger in such a strategy is that the project team may not have the necessary expertise on the subject matter to arrive at a meaningful decomposition. This leads to the third issue: capturing and accounting for possible interdependencies among components. It is important that critical couplings are not severed in decomposition; that is, critical couplings between different aspects of the problem should be contained within a given component. In reality, some couplings may be severed in decomposition. Therefore, such severed couplings should be accounted for in: (i) the judgments for the different components, and (ii) in the aggregation process. Careful attention should be paid to ensure that: (i) if the severing of potentially couplings cannot be avoided, the judgments recognize and account for the severed couplings to facilitate the aggregation of the components and the associated judgments; and (ii) the reduction of biases introduced by the decomposition.

2.5.3 Selection of Experts

Selection of the experts is one of the key aspects of the elicitation process. The credibility of an elicitation exercise, to a large extent, depends on the credibility of the experts (Morgan and Henrion, 1990). Three types of experts are typically needed in elicitations (Chhibber et al., 1992): (i) generalists, (ii) experts, and (iii) normative experts. Each of these types of experts plays a different role in the elicitation, as discussed in section 2.3.

In addition to substantive demonstrated knowledge in the subject matter of interest, experts should also exhibit several other attributes. Some examples are: (i) independence; (ii) lack of motivational, personal, or organizational biases; and (iii) willingness and ability to provide judgments in a clear and traceable manner. Independence refers to an individual's ability to discard explicit or implicit influences by other experts, whether these experts are other individuals or a specific piece of information. That is, an expert should be able to analyze the available information and articulate judgments that are based on his/her thought processes and ideas as opposed to depending on another individual's thought process or ideas. Lack of motivational, personal, or organizational biases refers to the selection of experts that do not stand to benefit in any way from the outcome of the elicitation. It is preferable to select experts that are devoid of any such biases. However, care should be exercised to ensure that the best qualified individuals in a given subject are not disqualified due to potential biases. A good normative expert should be able to recognize such biases and will ensure that they are articulated and compensated for in the elicitation. Finally, experts must be willing to participate and provide judgments that will be explicitly attributed to them. The experts should be willing and able to clearly articulate not only the judgments, but also their reasoning, assumptions, and analysis approach.

Typically, more than one expert will be used in an elicitation. When this is the case, care should be taken to ensure than there is "diversity" among the experts; that is, the various approaches, philosophies, insights, and experience of the discipline should be represented in the experts to provide robustness in the elicited judgments. This is extremely important in building confidence and credibility of the elicitation outcomes.

Experts, in principle, are not the only ones that should be devoid of biases and conflicts of interest; generalists and normative experts should abide by these criteria as well. Normative experts can influence the judgments by the manner in which these are elicited, and generalists in the manner in which the judgments are analyzed and used.

Bonano et al. (1990) as well as Chhibber et al. (1992) provide guidelines for the selection of the three types of experts discussed here, and the interested reader is encouraged to review that literature for more details.

2.5.4 Single Versus Multiple Experts

Chhibber et al. (1992) state that the use of multiple experts lends to the perception of credibility because of the belief of safety in numbers. However, such a perception is not universal. Some (e.g., Armstrong, 1985) consider the elicitation of judgments from a group of experts as being sound, whereas others (e.g., Clemen and Winkler, 1985), while agreeing with soundness of judgments from groups of experts, have reservations about increased confidence on judgments from a group of experts that can be highly dependent.

The idea that a group of experts is more credible than a single expert arises from the perception that a group introduces diversity of opinion. The strongest reason for the use of a group of experts is the increased likelihood of capturing the real uncertainty about the subject for which the judgments are being elicited (Bonano et al., 1990; Chhibber et al., 1992; Cambridge Decision Analysts Limited, 1992; Hora and Iman, 1989; among others). A secondary reason for using a group of experts is that biases introduced by one expert can be compensated by opposite biases introduced by other experts in the group. However, combination of judgments from a group of experts seldom accounts for the biases and the elimination of the biases are often left to chance (Chhibber et al., 1992).

The selection process consists, in order of implementation, of three basic stages: (i) decision regarding the use of a single expert, one group of experts, or several groups of experts; (ii) decision on criteria that will be used to select the experts; and (iii) implementation of the selection criteria. As stated earlier, because of the general belief that a group or groups of experts provide better judgments, little attention has been paid to the decision regarding whether a single expert or multiple experts should be used. By and large, the default position is that multiple experts, assembled in one or more groups, will be used. Given that the elicitation of expert judgments using a formal process is resource-intensive (i.e., expensive and time consuming) and that the amount of resources required increases (not necessarily linearly) with the number of experts used, it seems prudent to pay careful attention to the decision regarding the number of experts needed for the elicitation. It is possible that in some cases, the use of a single expert may be more appropriate than the use of multiple experts. A cost-benefit analysis should be performed to support the decision on the use of one versus multiple experts; Roberds (1990), for example, discusses a series of issues on this subject that may very well provide a framework for this cost-benefit analysis.

Several issues arise when it is decided to use a group of experts as opposed as a single expert Clemen and Winkler, (1985). First the suggested number of experts that can provide most of the expertise in a given subject matter is between three to five. Second is that the technical stature and reputation of each expert in the group should be comparable. This is particularly important because one wishes to avoid the situation in which the more reputable members of the group unduly dominate the results of the elicitation, specially if group sessions are held. Third is that, as stated earlier, the experts should be independent; however, familiarity with each other's work is essential. Fourth is the mix of disciplines needed to be represented by the group this, of course, is problem dependent. For some problems, such as the development of a PDF of hydraulic conductivity (Merkhofer and Runchal, 1989), the group

consists of experts in a single discipline (e.g., hydrologists). In other problems, an interdisciplinary group is advisable (e.g., see Hora et al., 1991).

One of the most critical issues that needs to be dealt with when using a group of experts is the aggregation of the judgments. Aggregation of judgments will be discussed in more detail later. However, suffice it to say that aggregation needs to consider not only the biases and possible dependencies among the experts, but also should account for the different ways in which the different experts or groups of experts view and analyze the problem. This problem was evident in the study by Hora et al. (1991) in which the modes and probabilities of human intrusion into the Waste Isolation Pilot Project (WIPP) in southeastern New Mexico were elicited. Different approaches used by the different groups of experts in that elicitation made it very difficult to aggregate the judgments later so that they could be used in the PA calculations. On the other hand, the study by DeWispelare et al. (1993) ensured, by the manner in which the judgments were obtained from each of the experts, that at least the judgments could be aggregated mechanically.

2.5.5 Information Provided to the Experts

The project team that commissions the elicitation should assemble and provide the experts with information it deems necessary and sufficient to help the experts to arrive at their judgments. Depending on the type of, and the manner in which, the information is provided, it can influence the judgments. Already, an earlier study—that of Hora et al. (1991)—was mentioned in which critics believe that the information provided to the experts had considerable influence in their estimates of the probability of human intrusion. Another example of presented information influencing judgments is presented by Bonano and Apostolakis (1991). In that study, a group of experts was asked to provide judgments regarding the recurrence of a characteristic earthquake at a hypothetical repository site; the purpose of the study was to demonstrate how expert judgments can be combined with other sources of information. Bonano and Apostolakis found that the experts were unduly influenced by a single report provided as reference material to the extent that the investigators were forced to redesign the study so that the report became the expert and the group of individuals simply evaluated the credibility of the expert. Yet another example in which experts complained about the information provided is the elicitation conducted as part of the United Kingdom Department of Environment's (UKDoE) Dry Run 3 (see Zimmerman et al., 1992). In that study, experts were apparently confused by a large amount of extraneous information for some variables, but expressed desire for more information on other variables.

Not enough information may not allow the experts to provide judgments with confidence, whereas too much information—some of which may be irrelevant—could confuse and mislead the experts. Some (e.g., Kahneman and Tversky, 1973; Chhibber et al., 1992) argue that providing too much information does not necessarily result in improved accuracy in the judgments. It is also reasonable to expect that, if the individuals providing the judgments are indeed experts, they should be able to distinguish and discriminate between useful and irrelevant information provided to them. However, criticisms levied against studies, such as that of Hora et al. (1991), suggest differently. It is possible that the experts themselves can suggest relevant supplements to the set of information to prepare themselves on the elicitation topic.

Care must be exercised in selecting the type and amount of information that will be provided to the experts. Fortunately, there are methodologies available that could be adapted to aid the project team in this selection. Recognizing that the information to be provided is likely to be uncertain, use of one of

these methodologies could allow the project team to perform an analysis that would assess the impact that the information may have on improving the judgments. Several authors, among them Lindley (1985a) and Goodwin and Wright (1991), discuss methodologies for estimating the expected value of uncertain or imperfect information prior to gathering the information.

In order to maximize the utility of data and processed information, the elicitation team should be available to answer questions and provide clarification. This support to bring reality to the data rapidly can be assisted by experiments, demonstrations, and even visitation to the location of interest. DeWispelare et al., (1993) found a site visit to be very beneficial to the experts not only in providing probability assessments for future climate at YM, but also in the confirmation of the assessments and their basis following the elicitation.

2.5.5.1 Use of Available Information by Experts

The fact that expert judgments are being sought is an indication that there is either a lack of major gaps in the available information about the questions, issues, or problems being addressed and, more importantly, that obtaining such information or closing critical gaps in it is highly unlikely. Therefore, a critical issue that arises is the type of information that is presented to the experts and the manner in which the experts are expected to use the information, which is the subject of this section. It is likely that the information provided to the experts applies to circumstances similar to the problem of interest, but not identical. Thus, it is important to ensure that the experts are able to discern both the similarities and differences between the site (or conditions) for which the information is applicable and the site (or conditions) of interest, and to factor these in the interpretation and analysis of the information.

Experts are likely to have preconceived notions about the issue or problem they are addressing. These notions are most often the result of many years of involvement in research in the specific area in which the judgments are sought, which is the primary reason why specific individuals are recruited to serve as experts. It is also likely that these notions will highly influence the experts' judgments. There is a tendency to readily focus on results for problems of a similar nature without, as stated earlier, paying close attention to the differences and how these need to be factored in arriving at the judgments relevant to the problem at hand. Slovic (1991) concluded that most individuals have great difficulty in accepting disconfirming information, and that generally they tend to discard information that does not fit within their reasoning paradigms. Morgan and Henrion (1990) suggest that experts need to be presented with, and "forced" to consider, information that challenges or disconfirms their conventional wisdom. These investigators state that experts are likely to exhibit the overconfidence bias because they fail to examine information that will support a point of view different than theirs. Morgan and Henrion (1990) argue that when experts explicitly consider information that contradicts their judgments, there is a general tendency to reduce overconfidence biases and improve the judgments. This suggests that normative experts have the daunting task of needing to challenge the experts to consider disconfirming information. This become particularly important when new information becomes available that, in principle, should result in a revision to prior judgments.

2.5.6 Form of Judgments

There are several types of expert judgments; for example, quantitative, qualitative, and value judgments, and anonymous and known judgments, among others. The type of the judgments, to a large extent, drives the entire elicitation process: (i) formal versus informal elicitation, (ii) decomposition of

the problem, (iii) selection of aggregation methods, (iv) selection of experts, (v) training of experts, and (vi) conduct of the elicitations.

One key issue is whether the judgments should be anonymous or associated with each expert. The elicitation of anonymous judgments has the marked advantage that individual experts can be shielded from strong pressures that may influence their opinions, and are likely to feel more comfortable with openly expressing their true views. There is a feeling that protecting the experts from outside pressures is beneficial (Morgan and Henrion, 1990; Chhibber et al., 1992). However, the use of anonymous judgments has some serious disadvantages. First, it was stated earlier that the credibility of the elicitation to a large extent depends on the experts used and, therefore, the use of anonymous judgments can be detrimental in this aspect. One way of enhancing the credibility of the elicitation is by associating each judgment with a specific individual. Second, the use of anonymous judgments is inconsistent with the licensing process, in which specific experts may be required to defend their judgments. Third, it hinders the review of the judgments because these are not likely to be associated with a given reasoning process. Through an analogy with the judicial review of scientific rulemaking (e.g., McGarity, 1984), it is reasonable to expect that the resolution of opposing views among experts of equal stature will depend on each expert's ability to articulate convincing arguments.

If the judgments are quantitative in nature and they represent subjective probabilities, then care must be exercised as to the selection of the form for expressing the associated uncertainty. Depending on the variable (i.e., discrete versus continuous variables), the probability can be of different forms. There are different forms of expressing the uncertainty, such as probability distributions or ranges of values. There may be "community accepted" or hypothesized specific distribution shapes (e.g., uniform or normal) for the elicitation situation. The project team must decide on the validity of these distribution shapes, and whether they will be enforced during the elicitation.

Similar to quantitative judgments, qualitative judgment can take different forms. A common form is pair-wise comparisons; that is, experts are asked to express preference for one event over another. Experts may be asked to provide their judgments as to whether one mode of human intrusion is more likely than another, or whether one phenomenon is more important that another when constructing a conceptual model (e.g., Thorne 1992; 1993).

2.5.7 Elicitation Training and Calibration

A reason for training of the experts prior to the conduct of an elicitation is to familiarize them with the problem at hand, the reason for obtaining judgments, and the manner in which the judgments are to be used. Another reason is to train the experts in elicitation techniques and allow them to become accustomed to providing their opinions, the reasoning and assumptions underlying those opinions, and to recognize and avoid—to the extent practicable—biases that can influence the opinions. If the purpose of the elicitation is to obtain probability judgments, then one key aspect of training should be the familiarization of the experts with probability encoding. Many individuals with substantive knowledge in a subject area often find it difficult to quantify uncertainty in terms of probabilities (Roberds, 1990). There are other reasons for training. One of them is to discuss and agree on a common framework to provide the judgments, in effect, to calibrate the elicitation team. Bonano and Apostolakis (1991) and Chhibber et al. (1992) refer to this framework as the model of the world. Generalists and normative experts play a key role in the development of the model of the world. This can be translated to mean that training could help ensure that the experts in a group or multiple groups are providing answers to the

same questions. It also means that the experts are exposed to the same understanding of the issue of interest.

One important aspect of training is the conduct of practice sessions for the purpose of assisting the experts to become familiar and comfortable with offering judgments, and the underlying reasoning and assumptions. Boyd and Regulinski (1979) mention that, in order to appear knowledgeable, an expert may attempt to suppress expressing uncertainty that he or she believes is actually present. It is particularly important to ensure that the judgments represent what is truly known and not known (i.e., express the real current state of knowledge). When experts are asked to express degrees of belief in terms of subjective probabilities, they need to become used to the techniques that will allow them to perform this task. Practice sessions provide the means for the experts to practice the application of these techniques.

If the experts are providing quantitative judgments (e.g., the PDF for the numerical value of a parameter), the user of the judgments could determine, in principle, the goodness of the judgments through a comparison with known values of the parameter. However, such an approach, by and large, is inconsistent with the need to acquire the judgments, for the latter may be necessary due to lack of data. Therefore, expert calibration is not likely in the strictest sense of the term. One possible approach suggested by both Cooke (1991) and Chhibber and Apostolakis (1993), is the use of surrogate variables. In this case, the user of the judgments selects variables similar in nature to the ones the experts will address, but the values of which only the user knows. The experts are then asked to provide judgments about the surrogate variables that can be compared to the known values. The capability of the experts to predict the values of the surrogate variables correctly can be used to determine their ability as experts. However, the issue is whether or not meaningful surrogate variables can be identified, for there is lack of experience with HLW disposal systems.

Finally, experts should be trained to recognize and, to the extent practicable, overcome the different types of biases. These biases can be partitioned into two main types; cognitive, and motivational. The cognitive biases are more commonly known as: (i) availability, (ii) anchoring, (iii) representativeness, (iv) ignoring base rates, (v) nonregressive predictions, (vi) overconfidence, and (vii) confirmation (Tversky and Kahneman 1982b; Morgan and Henrion, 1990; Cambridge Decision Analysts Limited, 1992; Cooke, 1991; among others). These biases are individually discussed in section 2.5.10.2.

Dealing with biases should not be restricted to the experts; motivational biases that can be introduced by both generalists and normative experts should also be dealt with. Generalists can bias the elicitation by the manner in which the problem is posed and/or decomposed, and by the amount and type of information that is presented to the experts. Kahneman and Tversky (1973) state that experts "... respond differently when given no specific evidence than when given worthless specific evidence ..." In the former case, experts tend to use prior probabilities properly, whereas in the latter, prior probabilities are likely to be ignored (Kahneman and Tversky, 1973). Perhaps the strongest criticism of the elicitation exercise conducted by Hora et al. (1991) has been that the project team presented information that may have influenced the experts and caused unrealistically low probabilities of human intrusion by drilling. Some of the WIPP oversight groups, like the Environmental Evaluation Group, argue that the information presented to the experts was not consistent with drilling rates for the region as contained in Bureau of Land Management records. Normative experts can bias the outcome of an elicitation by the nature of questions asked and by the manner in which they are asked.

2.5.8 Elicitation of Judgments

Occasionally, the tendency is for a formal elicitation of expert judgments to be very structured and rigid. This rigidity can be both advantageous and detrimental. It can be advantageous in the sense that, when using more than one expert, the judgments are likely to all be expressed in similar, if not identical, manner. This, in turn, should facilitate the processing and use of the results from the elicitation as input to computational analyses. The lack of rigidity and structure in the approach may make it difficult to consolidate the results in a manner that will render them useful for the subsequent analysis. Such was the case in the elicitation of drilling rates for human intrusions into the WIPP (Hora et al., 1991) discussed above. In that elicitation exercise, four different groups of experts were used to obtain estimates of drilling rates for intrusions that will intersect the disposal facility. Each group was allowed to choose the method to estimate the probability of intrusion and hitting a waste container. Due to a lack of a common set of rules among the four groups, the probabilities could not be combined to arrive at useful drilling rates for the consequence analysis of human intrusion (Hora et al., 1991). However, it can be detrimental because too rigid an approach may not allow for sufficient flexibility in the way the different experts interpret and process information to capture important and necessary diversity of opinion. Instead, a complex procedure had to be designed to use the probability estimates (Trauth et al., 1993). Proper balance must be achieved in designing expert judgment elicitations so that the process has a definite framework to ensure the utility of the results, but at the same time, it should allow for enough flexibility to capture the diversity of opinion that the experts may represent.

2.5.9 Dependence of Experts

The use of multiple experts introduces questions about the dependence among the experts. Expert dependency becomes more critical when using a group of experts with similar educational backgrounds and experience (Chhibber and Apostolakis, 1993) because this increases the likelihood that they will invoke similar assumptions and approaches to arrive at their judgments. Clemen and Winkler (1985) concluded that dependence between experts could be detrimental to the posterior confidence on the judgments. Dependence among experts has the net effect of reducing the effective number of experts. Let's say that in a group of five experts, four of them are highly positively correlated (correlation coefficient >0.8); this means that, in practice, there are approximately only two truly independent experts because the four highly dependent experts closely act as a single expert. The effect of correlation is a significant loss in the diversity of opinion originally sought when convening the group.

2.5.10 Debiasing

A bias is a systematic tendency to ignore relevant factors and/or take into account irrelevant factors to the problem at hand. There are two general classes of bias: motivational and cognitive. Motivational biases occur because an expert has a vested interest in an issue and consciously or unconsciously distorts his judgment. Cognitive biases occur because of a failure to process, aggregate, or integrate the available data and information (Tversky et al., 1982). Motivational biases can generally be avoided, or at least reduced, by a careful expert selection process. A training session can teach the experts how to recognize conspicuous or implicit motivational biases and can familiarize the analyst with the expert's information processing characteristics. This allows the analysts an opportunity to recognize when implicit motivational biases influence the judgments so that they can take appropriate action.

Cognitive biases may be less difficult to alter. Cognitive biases in probability judgments include overconfidence, anchoring, availability, representativeness, ignorance of base rates, nonregressive predictions, and the confirmatory bias. According to Evans (1989) the major cause of bias in human reasoning and judgment lies in the factors that induce people to process information in a selective manner. Selective processing can occur during the formation of a mental representation of the information provided or during the subsequent processing of the information. People often selectively ignore relevant features of the problem or alternatively take into account irrelevant features, leading to biased judgments.

Selective processing of information is far from unintelligent; it is, in fact, a perfectly reasonable solution for reducing the enormous amount of information confronted by the problem solver. Some researchers, for example, Newell and Simon (1972), have described the appropriate reduction of information by applying debiasing as a defining characteristic of intelligent behavior. Selection is fundamental to intelligence. However, mistakes in selection occur and, when systematic, lead to bias in judgments. The following sections describe the major biases in making probability judgments.

2.5.10.1 Motivational Bias

Motivational bias is assumed to be driven by our human needs, such as for approval. Motivational bias occurs when: (i) experts do not report their actual solutions or thought processes, (ii) the interviewer or analyst purposefully misinterprets the experts' knowledge.

Experts may change the descriptions of their solutions due to economic incentives or personal interest or because of concerns over the social acceptability of their response. This last case can manifest itself in an expert responding to questions in a way that they think others, either present or not, would find acceptable.

Another potential source of bias is the interviewer and analyst from the elicitation team. The interviewer may ask leading questions that cause the expert to misrepresent an answer. The analyst can also error in the interpretation of the expert's information. Humans selectively perceive and interpret incoming information to support their existing beliefs. This can lead to biased analysis. This is a particular sticking point of knowledge elicitation because the analyst learns from the expert through the filter of their own interpretation.

Expert judgment can also be altered by misrepresentation. Expert data can be misrepresented by coding it into the requested response mode or when it is modeled for analysis or knowledge-based system development. A model may assume particular properties of the data, such as its distribution, that may not be valid. Or the analyst may have to aggregate multiple and differing expert judgments to provide one input value and do so in an undocumented arbitrary way without the benefit of the elicitation team.

2.5.10.2 Cognitive Biases

Availability

When asked to estimate some quantity, people tend to base their estimates on the ease with which members of the category can be retrieved from memory. Thus, subjects estimate the frequency of an event based on the number of examples that they can recall. For example, when asked to estimate the probability of death from various causes, subjects typically overestimate the risks of well publicized

causes (e.g., botulism, snake bite) and underestimate more common causes (e.g., stomach cancer, heart disease). Also, well publicized causes are more likely to be recalled (Slovic et al., 1982).

Anchoring

When asked to estimate a probability, people fix on an initial value and then adjust or correct this value as they consider the problem. Frequently the adjustment to the initial value is insufficient. In one experiment, subjects were asked what percent of the member nations of the United Nations were African. The experimenter then generated a number between 1 and 100 by spinning a wheel. The median estimates of the percentage of African countries was 25 and 45 for groups with anchor numbers 10 and 65, respectively (Tversky and Kahneman, 1982a). Thus, even a seemingly randomly generated number can serve as an anchor and influence subsequent judgments. Focusing on the extremes of a distribution, rather than the central part, can help reduce the tendency to anchor on a central value and reduce the overconfidence bias as well (Winkler et al, 1992).

Representativeness

In judging the likelihood that a specific example belongs to a general class, people expect the fine details of the object to reflect the larger class. For example, Kahneman and Tversky (1972) asked people to judge the likelihood of strings of independent tosses of a fair coin. People judged the string of tosses HTHTTH to be more likely than either HHHTTT or HTHTHT because the first appears to reflect a random coin toss. Although all three strings are equally likely, the first looks more random. Use of the representativeness heuristic also occurs when people pay too much attention to specific details and ignore background information such as base rates.

Ignoring Base Rates

Base rate refers to the frequency of an event in the population of events. This is a bias in which people ignore information about the relative likelihood of instances in a population. The best known example is the cabs problem (Kahneman and Tversky, 1972). Subjects read that two cab companies run in a city, one that uses blue cabs and one that uses green cabs. The blue cab company has 85 percent of the business in the city, while the green cabs service only 15 percent. A cab is involved in a hit and run accident, and an eyewitness identifies it as a green cab. Testing the witness reveals that, under similar viewing circumstances, the witness correctly identifies the cab color 80 percent of the time and is incorrect the other 20 percent of the time. The subject must judge whether it is more likely that the cab is green or blue. Most subjects say the likelihood that the cab is green is higher when in fact a blue cab is more likely. This and many other studies have demonstrated that subjects will ignore or severely discount base rate information.

Nonregressive Predictions

Relationships between variables are unreliable and vary over time. A nonregressive prediction assumes that the relationship between variables is reliable. This type of bias is very common in modeling where algorithms are assumed over a wide regime when they may have been tested and shown to be valid over only a narrow range. This is also true for parameter values which are robustly characterized when there is little data to support this. Since many predictions are temporarily based, historic relationships are often extended into the future when in actuality there is little or no basis for this in terms of underlying relationships.

Overconfidence

Overconfidence occurs when people supply judgments that express less uncertainty than their knowledge justifies. For example, studies have been conducted that ask general knowledge or almanac questions. The subject needs to provide an answer and a probability that the answer is correct (Lichtenstein et al., 1982). These studies show that the subjects provide probabilities that are significantly more extreme than the corresponding relative frequencies, showing more confidence in their answer than is justified.

Confirmation Bias

People have a fundamental tendency to seek information consistent with their current beliefs, theories or hypotheses, and to avoid collecting potentially falsifying information. Wason (1960) asked subjects to discover a simple rule guiding the formation of number sequences. Subject were told that the triple "2 4 6" conformed to the rule. The rule was that any ascending sequence was correct. Subjects formed triples and the experimenter gave feedback. The majority of subjects provided at least one wrong guess for the rule and a substantial number failed to discover. The interesting outcome is that in general subjects tested their hypotheses with only positive examples and not with examples designed to falsify their currently held hypothesis.

2.5.10.3 Approaches to Debiasing

Researchers have suggested a number of approaches to debiasing experts, including procedural, training, task, and decision aid approaches. This section describes these alternatives.

A Procedural Approach

Meyer and Booker (1991) outline a six-point approach to handling bias in expert elicitation.

- Anticipate which biases are likely to occur
- Redesign the elicitation to make it less prone to the anticipated biases
- Make the experts aware of the potential intrusion of particular biases and familiarize them with the elicitation procedures
- Monitor the elicitation for the occurrence of bias
- Adjust, in real time, to counter the occurrence of bias
- Analyze the data for the occurrence of bias

These steps are useful in order to become aware of bias, prevent its occurrence, avoid criticism of the data collected in the elicitation, or to analyze the data for the presence of bias. However, approaches to handling bias are as much an art as a science. Researchers cannot compare an expert's answers to any objective standard, so it is impossible to completely rule out the influence of bias. Biases are very difficult to detect and counter.

In the first step of this program, project personnel should attempt to determine which biases are likely to emerge in the planned elicitation. This step should be performed in any elicitation process. If the potential for the manifestation is present, the elicitation should be redesigned to avert the bias. Also, the experts should be informed about the potential for bias and participate in debiasing exercises. This part of the elicitation training is very helpful in reducing bias. During the elicitation, the project team should be sensitive to the possibility of introducing bias and monitor the elicitation for signs of bias. If these signs occur, the team should adjust the elicitation in real time to counter the bias. Finally, the team should review the data for signs of bias.

Training Approaches

Training to reduce possible bias varies in effectiveness depending on the nature of the bias and the type of training approach. Evans (1989) argues that training is effective only on those biases in which people have an inadequate understanding of the appropriate rule to apply. When people generally understand and know the rule but do not use it, training is less effective.

Studies of training in probability assessment have shown generally positive results, indicating that prior to training people have an inadequate grasp of statistical rules. For example, Fong et al. (1986) report that both rule-based and example-based training was effective in facilitating the use of the statistical law of large numbers (i.e., the larger the sample size, the more closely the characteristics of the sample approximate the population characteristics) in problem solving. Attendance at formal statistics classes also yielded marked improvement in problem solving.

Also, Lichtenstein and Fischhoff (1980) found considerable improvement in calibration after a single session of 200 discrete judgments with comprehensive feedback. This improvement generalized to other discrete choice tasks differing in difficulty, content, and number of alternatives but did not generalize to assessing continuous distributions.

The type of training technique is also an important consideration. Experience-based training rather than instructional training may be more effective in developing the implicit thought processes necessary for a full application of appropriate rules. For example, Reber (1976) asked subjects to learn strings of letters that were produced by a finite state grammar. Reber showed that subjects learned the rules that produced the strings better when they were told simply to memorize the strings rather than explicitly asked to discover the rules for letter orders. Subjects in the explicit learning condition formulated a number of inaccurate rules. Berry and Broadbent (1984) found similar results for experience-based versus instructional training for tasks in which subjects learn and apply implicit rules.

The type of feedback is also important in reducing bias. Hogarth (1987) argues that task structure feedback about the relationships between available information and the tasks is more effective than pure outcome feedback. Lichtenstein and Fischhoff (1980) report that a personal discussion of the results may be more effective than a written summary because the former is more difficult to dismiss.

Another consideration in training is the specificity of the approach. Transfer of training between one domain and another is minimal. Lichtenstein and Fischhoff in their 1980 study found substantial improvement in calibration due to training but found little transfer to question answering in a different domain. Fong et al. (1986) found that the benefits of training transferred to a new problem domain but only immediately; a delayed test did not show any benefit of training on a second task although the

benefit in the first task remained. Evans (1989) recommends the use of domain-specific, experience-based training to reduce biases in reasoning.

Task Approaches

Altering the task environment can reduce biased reasoning. This is particularly true in circumstances in which people fail to apply a principle of which they have at least some understanding. For example, in studies of the representativeness heuristic Kahneman and Tversky (1972) found that subjects were insensitive to sample size in making statistical judgments. However, Evans and Dusoir (1977) replicated the study and found that a significant majority of subjects could solve the problem when the wording was simplified. And Evans and Dusoir (1975) found almost 100 percent responses when asking subjects whether a statistical estimate based on a small or large sample size was more likely to produce an accurate result. Thus, when people have a general understanding of a principle, simplicity and clarity of presentation is important in reducing potential errors in judgment.

Another method of altering the task environment that has an effect on accurate judgments is encouraging subjects to attend to negative or disconfirming information. People fail to make use of disconfirming information; subjects actively attempt to verify rather than falsify their hypotheses, a confirmation bias. In probabilistic estimation this promotes overconfidence (Einhorn and Hogarth, 1975). When subjects are forced to consider disconfirming information, overconfidence is significantly reduced. Or when subjects are encouraged to form positive or neutral representations of features that would normally be viewed negatively, they avoid the confirmation bias (Tweney et al., 1980).

Another feature of the task environment that leads to debiasing is decomposition of the problem. For example, Armstrong et al. (1975) found that subjects' estimates of general knowledge quantities were more accurate if based on decompositions than if made directly. The decompositions used in this study were made by the experimenters. This effect presumes that the subjects know more about the judgments in the decompositions than the overall quantity. Henrion et al. (1989) studied the impact of decomposition on subjects answering almanac questions. Assessments based on decompositions were no better calibrated than aggregated assessments. This might have been due to the fact that the subjects had no better ability to estimate the decomposed quantities than the overall quantity. However, the direction of the bias was shifted from systematic underestimation to systematic overestimation of the values. Thus, the structuring of the problem can have a dramatic effect on the outcome.

Problem formulation can have a significant effect on bias. Good reasoning is facilitated by successful identification of analogies between the problem at hand and previous situations that have been learned through experience (Evans, 1989). Abstract or arbitrary terms inhibit good reasoning, and concrete examples are of great importance in learning how to reason and reasoning effectively.

Interactive Decision Aids

Many researchers have suggested that computer systems should be designed to augment rather than replace human judgment. These systems could support those functions in which humans are notoriously deficient. Evans (1989) suggests that decision aids could perform the following functions: (i) provide factual information to inform the user of aspects of the decision; (ii) perform computational procedures to assist the user in calculating the consequences of possible decisions; (iii) help the user to structure the problem and represent its essential characteristics; and (iv) use decision theory to recommend optimal decisions. A system of this kind aids in debiasing decision making in a number of respects.

Providing relevant and accurate information helps avoid errors due to ignorance or insufficient access to data. Computational procedures replacing intuitive judgments reduce bias by ensuring the appropriate weighting of factors such as base rate information. Helping the user represent and structure the problem reduces the likelihood of error due to limitations in the user's working memory capacity or poor organization and facilitates decomposition as an effective decision strategy. Finally, the use of normative decision theory provides a rational basis for choice.

2.5.11 Elicitation Documentation

Precise and complete documentation is pivotal to the success or the elicitation. Complete documentation should include a discussion of all steps in the elicitation procedure. Each step should be described and the results presented. The results should be presented for each expert and then for any combined analyses.

Documentation of the elicitation session by the elicitation team is critical to the utility of the data produced. Documentation is the vehicle through which the quality of expert judgment is evaluated. Having the experts produce a written basis for their judgments is very useful. Checks for consistency and logic flow during the elicitation are fundamental ways of alerting the elicitation team of the possibility of bias being exhibited. An expert's own documented stance for judgments is a primary way for the elicitation team to check for consistency and deviations in the logic flow int he bases of the judgments. DeWispelare et al. (1993) also found this pre-elicitation documented basis to not only assist in logic and consistency checks during and after the elicitation, but also was very valuable in focusing the questioning and discussions during the elicitations.

2.6 AN EXAMPLE ELICITATION

DeWispelare et al. (1993) used the following eleven steps to implement a formal expert elicitation procedure:

(i) Determine the objectives and goals of the study
(ii) Recruit the experts
(iii) Identify the issues and information needs
(iv) Provide initial data to the experts
(v) Conduct the elicitation training session
(vi) Discuss and refine the issues
(vii) Provide a multi-week study period
(viii) Conduct the elicitations
(ix) Provide post-elicitation feedback to the experts
(x) Aggregate the experts' judgments (if required)
(xi) Document the process

The goal of this study was to evaluate the climate in the YM vicinity over the course of the next 10,000 years. To meet this goal a panel of five experts was assembled. The experts were selected from nominations of professional societies. Experts that were willing, able, and had no conflict of interest were included in the final list of nominees. This list was circulated to the experts and they rated each other on their suitability to address the issue under investigation.

Once the panel was established, the elicitation team at the Center for Nuclear Waste Regulatory Analyses (CNWRA) provided the initial data for review and assembled the group for a training and debiasing session. Following a study and research period, the experts participated in individual elicitation sessions. They then reviewed the results of these sessions and met once again to attempt to behaviorally aggregate their individual judgments. The entire process was documented rigorously.

2.6.1 Lessons Learned

Several lessons were garnered from the climate application (DeWispelare et al., 1993) that bear listing. All of these lessons relate to issues discussed in this report.

- The quality of the experts is directly related to consistency and facility of the elicitations. The generalists all commented that the clarity and logic provided in the judgments and supporting rationale by the experts was due to their overall expertise.

- Defensible process for selection of the experts is feasible. To quiet the voice of frequent criticism for lack of independence or bias in the experts, a process like the peer nomination and selection process utilized is possible even with a relatively tight time schedule.

- Debiasing training of the experts is essential to a smooth elicitation. As in the climate elicitation, most experts have had only limited or no experience at producing consistent subjective probability distributions. All of the experts agreed with the normative experts and generalists that this training was essential to the process.

- A mechanical aggregation of the experts' judgments was easier to implement than behavioral aggregation. When it is necessary to aggregate the individual judgments of the experts after they have been elicited, it is efficient to utilize a mechanical aggregation scheme. This scheme can be easily documented to provide traceability. As expected, the elicitation team had considerable difficulty in its attempt to form a behaviorally based aggregation because of the differences in the individual judgments and the experts' conviction towards their judgments.

- A site visit by the experts is valuable to calibrate the interpretation of data and research done in preparation for the elicitation. The experts indicated that the experience gained through the site visit was notable in accelerating the preparation for the elicitation, as well as aiding in the post-elicitation validation process.

- Individual documentation is critical to a successful elicitation. The documentation of note consisted of two parts. A short paper to form the basis of the judgment by each experts served as the reference for understanding the reasoning expressed in the elicitation. The second part consisted of documentation of each elicitation session by the elicitation team to insure that the reasoning utilized by each expert was understood and expressed consistently. Video taping of each elicitation session served as a basis for the team to use in checking session notes and as a permanent record of each session.

2.7 SUMMARY

The resolution of each issue in conducting a formal expert elicitation influences the outcome of the process, and the resolution of many of the issues relies on subjective judgments of the project or assessment team. The issues of expert selection, training, and documentation are arguably the most critical. The selection of the experts needs to be formal and as fair as possible. For example, having academic and professional society nominees rate each other in an atmosphere of confidentiality increases the credibility of the judgments obtained. The credentials of the experts establishes credibility of the elicitation, and their ability to communicate their reasoning is a primary determinant of the quality of the results. Training the experts to avoid bias and effectively communicate their judgments and the basis for them is essential in acquiring the best possible information. Finally, comprehensive documentation of the elicitation process and results enables a detailed review of the outcome, such as that can be provided by a peer review.

3 METHODS OF COLLECTING AND PROCESSING EXPERT JUDGMENTS

3.1 ELICITATION TECHNIQUES

Once a decision to use expert judgment is made, the techniques used to elicit the judgments need to be considered. This chapter discusses methods of collecting expert judgment, additional issues of aggregating opinions, updating of the judgments as new information becomes available, and combining of elicited judgments with other types of data.

The technique used to obtain an expert's opinion depends on the type of information required. The following techniques can be used to structure the expert elicitation and obtain expert judgments. People make judgments by identifying the options or events to be judged, screening the options and events, decomposing the judgmental task into smaller, more manageable pieces, and quantifying judgments about the options and events. There are techniques for aiding the expert during expert elicitation in each of these tasks.

3.1.1 Identification Techniques

Identification techniques aid in activities like scenario generation and development of conceptual models through stimulating the thought process of the experts. In scenario generation, the emphasis is on stretching the bounds of event generation. In conceptual model generation, the emphasis is on generating model alternatives.

Among the identification techniques are the following: (i) forward and backward induction, (ii) value-driven event generation, and (iii) analog-driven event generation. Forward and backward induction consist of listing all possible and conceivable events that, for example, could occur in a HLW repository. In forward induction, the listing starts with initiating events and fans out linking events that could occur in thousands of years. In backward induction, the final states are starting points and the expert works backward to generate sequences of events that could result in each final state. In each case, branches that are extremely unlikely, redundant, or impossible are pruned from the event tree.

Value-driven generation establishes the objectives of a given situation and what events or scenarios would follow. This technique helps identify events and scenarios that are desirable to achieve and those that are potentially troublesome.

In an analogy, the expert tries to report events and scenarios that are applicable to a situation that has aspects in common with the situation under investigation. Then the expert indicates whether any of those events or scenarios are likely to occur in the current situation. For example, an expert might attempt to generate events that could go wrong in a mine containing lethal gases and then determine if any of those are applicable to the HLW repository. A discussion of recent experiences with identification, development, and screening of scenarios in HLW programs is presented in section 5.2.

3.1.2 Screening Techniques

The first step in screening any set of issues or events is to identify the attributes with which to screen alternatives. After the attributes have been identified, target levels and constraints on the values of the attributes need to be established. Alternatives are screened out that do not meet the target levels and constraints. Attributes for screening scenarios might included scientific acceptance (due to parametric values), physically realizable and consistent, predictive ability, simplicity, and cost.

3.1.3 Decomposition Techniques

Problem decomposition is widely used to simplify a complex problem into components that are more manageable and more easily solved. Problem decomposition also is an important tool in expert elicitation (Raiffa, 1968; Armstrong et al., 1975). Problem decomposition in elicitation refers to breaking down issues to provide for easier and less complex assessments that can be recombined into a probability density or utility function.

Among the major techniques for decomposing factual problems are: fault trees and event trees. Fault trees focus on a possible failure in a system and trace back the possible component causes of the failure. In fault tree analysis, the components are assigned probabilities of failure from which an overall failure probability for the system can be found. While fault trees end in a single failure event and trace its possible causes, event trees begin with an initiating event and draw out its possible consequences. The event tree lays out the sequence such that the probability of successive events are conditional on their predecessors.

3.1.4 Techniques for Quantifying Probability Judgments

Bonano et al. (1990) categorize techniques for quantifying probability judgments according to whether the judgment is about a discrete event or continuous quantity and whether the judgment is a magnitude judgment of an event or a judgment of indifference between gambles, creating four classes of procedures. These techniques are used during the individual elicitation process to help the expert generate probability estimates for the variables of interest.

3.1.4.1 Quantification of Uncertainties Using Probabilities

Uncertainties affecting the prediction of the long-term behavior of HLW repositories should be quantified to the extent practicable. This task is typically accomplished using probabilities. There are two main classifications of probability (e.g., Morgan and Henrion, 1990; Neapolitan, 1990): (i) classical probability, (ii) relative frequency probability, and (iii) subjective probability. In the following, we summarize these three types of probability and indicate that situations will arise where these probability types are appropriate for the quantification of uncertainties associated with the analysis of HLW repositories.

Classical Probability

Classical probability is based on the assumption of equally likely outcomes of an event (flip a fair coin, spin a fair roulette wheel, etc.). These probabilities can be determined theoretically *a priori* as one over the number of possible outcomes (1/2 for a two-sided coin, 1/6 for a six-sided die, etc). While

of interest in games of chance, classical probabilities have little utility in a complex situation like a HLW repository.

Relative Frequency Probability

The relative frequency or, as it is more commonly known, the frequentist probability, is based on the number of times a given outcome or result repeats itself in a large number of identical experiments. Occasionally, a theoretical numerical value can be assigned to the probability *a priori*, the so called limiting frequency, (e.g., Neapolitan, 1990). This theoretical limiting frequency is conceptual, requiring an infinite number of trials. In practice, we have finite sample sizes. The mathematical definition of the frequentist probability of event E occurring, p(E), as the result of repeating an identical experiment, is (von Mises, 1957):

$$p(E) = \lim_{n \to \infty} \frac{S^n(E)}{n} , \qquad (3\text{-}1)$$

where

n = the number of times the experiment is repeated

$S^n(E)$ = the number of times event E occurs.

The most common example employed to explain the frequentist probability is the coin toss experiment. The typical coin in the United States has one side commonly known as heads and another side known as tails. The example is as follows (the numbers used in the example were chosen arbitrarily and do not represent the results of an actual experiment): if one tosses an unbiased coin in the air 10 times, the heads side will turn up 4 times when the coin lands on the floor; 46 times for 100 trials; 487 times for 1,000 trials; 4,953 for 10,000 trials; 49,967 for 100,000 trials; and so on. The trend is quite obvious: the frequency (or number of times) that heads come up asymptotically approaches 1/2 the number of trials as the number of trials increases; thus, 1/2 is called the limiting frequency of heads turning up when a coin is tossed. In the frequentist approach, the probability of heads turning up in a coin toss experiment is 1/2 (0.5) (Morgan and Henrion, 1990; Neapolitan, 1990). The probability of tails turning up is also 1/2 (i.e., 1-probability of heads).

For example, frequentist probabilities are widely used in the insurance industry and represent the analysis of actuarial data. Insurance premiums are determined based on the probability that the circumstances against which protection is sought will actually occur and the company underwriting the insurance will have to pay out the benefits (i.e., money). In the United States, automobile insurance premiums (or rates) for unmarried males under the age of 25 are higher than for any other group because the frequency with which members of this group are involved in automobile accidents is higher than for any other group. Thus, unmarried males under the age of 25 have the highest probability of being in an automobile accident compared to any other population group in the United States, and insurance companies incur in a higher risk when they provide automobile insurance for members of this group; hence, the companies charge higher premiums for these individuals.

In summary, frequentist probability is not based on a single event but rather on the repetition, theoretically, of a situation an infinite number of times and the likelihood or propensity for a specific consequence of the event to occur over these repeated trials.

Subjective Probability

Subjective probability represents an individual's educated guess that, after considering all currently available information, a specific event will occur (de Finetti, 1974; Morgan and Henrion, 1990; Neapolitan, 1990). The dependence of this type of probability on the available information cannot be overemphasized because as the information, or state of knowledge, changes so can the value of the probability. As the state of information or knowledge changes as a result of new data or evidence, the value of the probability can be updated.

Different individuals can arrive at different values of the probability of an event because they may have different information about the same event or may interpret the same information differently. However, subjective probabilities should not be assigned in a totally arbitrary or capricious manner. These probabilities must satisfy the axioms of probability theory (Morgan and Henrion, 1990). That is, if an individual assigns x as the probability that event E will occur, he or she must assign the value 1-x to the probability of its complement (E will not occur). The probability of occurrence of a set of mutually exclusive events must be the sum of the probabilities of the individual events and this sum shall be equal to 1. Therefore, if new information on an event leads an individual to update the probability of occurrence of that event, he or she must simultaneously update the probability of the complement of the event for the individual must maintain self-coherence in the assignment and updating of probabilities.

Use of Subjective Probabilities in the High-Level Nuclear Waste Repository Program

For certain analyses in the HLW repository program, frequentist probabilities are appropriately used. Site exploratory field or exploratory data gathering and analysis is a case where relative frequency-based probability is well suited to characterize the uncertainties with events or technical descriptions.

Frequentist probabilities rely on numerous repetition of the same experiment, event, or conditions, a situation that cannot be reasonably expected to apply to all aspects associated with a HLW repository. A HLW repository is a one-of-a-kind system; thus, frequentist probabilities may not be suitable for quantifying all of the uncertainties affecting the prediction of the system's performance. Many investigators (e.g., Apostolakis, 1990; Smith, 1990; Morgan and Henrion, 1990; Neapolitan, 1990) advocate the use of subjective probabilities for the quantification of uncertainty related to one-of-a-kind complex problems or systems, such as may be found in a HLW repository. Furthermore, it is well accepted that, by and large, the uncertainties affecting the prediction of performance of the repository system are due to lack of knowledge, and subjective probabilities can be updated as more information becomes available via Bayes' Theorem. Bonano and Apostolakis (1991) demonstrate that Bayes' Theorem is flexible enough to accommodate new information from various sources, such as new data from experimental investigations, modeling studies, and expert judgments.

Finally, subjective probability can be used to quantify a wide variety of uncertainties. The uncertainty can be depicted in many ways, including ranges of values, histograms, PDFs, and CDFs, among others (Finkel, 1990). Therefore, it readily lends itself to uncertainty analysis and sensitivity analysis.

It can be concluded that the variety of uncertainty analyses involved in the HLW repository program will appropriately utilize both types of probabilities as circumstances dictate.

3.1.4.2 Judgments About Discrete Events

Techniques in this category are used to assess probabilities for a finite number of mutually exhaustive and exclusive events. The probabilities of the events should sum to one. Techniques that fall into this category are direct probability and direct odds. The direct probability technique simply asks the expert to provide a probability estimate for each event. For more than two events, the expert should provide probabilities for all events under consideration, then check the sum and adjust the estimates as necessary.

The direct odds techniques allows the expert to make judgments for pairs of events. In this case, the expert estimates the relative odds of one event over another. Probabilities can be calculated from the results of the paired comparisons.

Indifference between gambles can also be used to quantify judgments for discrete events. Among these techniques are the reference gamble technique and the certainty equivalent technique. In the reference gamble technique, the expert selects one of two gambles. In the first gamble, the expert wins a stated prize if the event occurs and loses if it does not. The second gamble is one in which the expert receives the prize with a known probability p if the event occurs. This gamble is manipulated until the value of p is such that the expert is indifferent between the two gambles. That value p is the probability assigned to the likelihood of the event.

In the certainty equivalent technique the expert compares one gamble (winning if the event occurs and losing if it does not) and one sure amount. The expert states a certain amount of money for which he would be indifferent between taking the gamble and simply receiving the money. The certain amount is the certainty equivalent of the gamble at the point at which the expert is indifferent between receiving it and taking the gamble. From this certainty equivalent, the expert's probability for the event of interest can be determined.

3.1.4.3 Judgments About Continuous Uncertain Quantities

Fractile and interval techniques, among the possible techniques in this category, are used to estimate continuous numerical quantities. The fractile technique is widely used to construct the CDF of an uncertain quantity. A z-fractile is the magnitude x_z of the quantity x such that there is a probability z that the true magnitude falls below x_z and a 1-z probability that it falls above it. Thus, the cumulative distribution shows the fractiles plotted against the magnitude of the uncertain quantity.

The interval technique is also used to produce a CDF describing the expert's uncertainty in the magnitude of a quantity. In this technique, the expert assigns probabilities to intervals of the uncertain quantity for the probability that the true magnitude will fall between the upper and lower bound of the interval. Alternatively, the expert can estimate the probabilities that the true value falls into the open intervals below and about the preselected interval. The results of either method can be plotted to obtain a CDF.

No matter which technique is used to quantify the probabilities, consistency checking results through alternative representations or use of another technique is essential. For continuous techniques like fractile and interval, constructing the PDF directly is also an option. Whether a CDF or PDF is produced directly, converting to the other in having the expert validate the converted distribution is an appropriate consistency check. Automated software programs to rapidly draw and identify distributions and

mechanical devices like the probability wheel are but a few of the means available to assist in assessing probabilities or rapidly checking for their consistency.

3.2 AGGREGATING THE JUDGMENTS OF THE EXPERTS

Many expert elicitations produce information in the form of probability distributions to conveniently represent the uncertainty in the judgments. The beginning assumption is that the experts have ironed out differences in definitions, all agree on exactly what is to be forecast or assessed, and as much as possible has been done to eliminate individual cognitive and motivational biases. In this case, of course, it is still possible for reasonable experts to disagree for a multitude of reasons, ranging from different analytical methods to differing information sets or different philosophical approaches. Thus, consulting multiple experts may be seen as a subjective version of collecting more data in a survey. Also, because subjective information is often viewed as being softer than hard scientific data, it seems appropriate to consult multiple experts in an attempt to increase the confidence in the information obtained.

These motivations are reasonable; the fundamental principle that underlies the use of multiple experts is that a set of experts can provide more information than a single expert. The nature of expert judgment, though, requires great care in how multiple judgments are used. Cooke (1991) spells out a number of issues that must be considered in using expert judgment: variability among experts, dependence, reliability (reproducibility), and calibration. All of these come into play when dealing with multiple experts, but the most crucial for our purposes are calibration of the individual experts, which can be characterized as the quality (e.g., bias and precision) of individual expert judgments, and dependence among experts.

Any procedure for using multiple experts should yield information that is useful in the decision context. The objective is to extract as much accurate and appropriate information as possible from the experts and to evaluate various expert-use approaches on this basis. How can information about risks be captured, communicated and used in a coherent way? The available information may include the individual judgments, of course. It may also include background information about the experts' analytical methods, philosophical bases, data used, and so on. In addition, there may be information available regarding the quality of the experts' judgments, including accuracy and dependence.

Information about quality of the experts' opinions may come from the experts themselves or from other sources. In some cases, there may be a record of past judgments and actual outcomes that can be used for verification, although this situation is rare in risk analysis. Another important source is the team of individuals who structure and perform the risk assessment. This assessment team, being composed of individuals knowledgeable about expert judgment and knowledge elicitation, as well as the content area, can provide assessments of the quality of the experts' judgments. Thus, the experts provide judgments about the risk in question, and the assessment team, being experts on experts, can provide judgments about the quality of the original expert judgment. Although it sounds as though it might become an infinite regress of always having to judge the quality of the previous assessments, in practice there is no need to go beyond some straightforward assessments of the quality of the original experts' judgments.

The need to incorporate information about the experts' judgments runs throughout the literature on aggregating expert opinion. Early axiomatic models for aggregating probabilities derived opinion

pools, weighted averages of experts' probabilities (see Genest and Zidek, 1986). Though procedures for assessing the weights were typically not specified, authors indicated that these weights were at the discretion of the person making the aggregation and should reflect the relative quality of the experts' judgments (Genest and McConway, 1990). Since the early work of Reid (1968) and Bates and Granger (1969), the work on combining forecasts has focused on deriving combining weights that reflect qualities of the forecasts as reflected in past data (Clemen, 1989). Beginning with Morris (1974), a Bayesian paradigm for aggregating experts' probabilistic judgments has explicitly required the assessment of a multivariate likelihood function that encodes information about the quality of the experts. Examples include Winkler (1981), Lindley (1983; 1985b), and Clemen and Winkler (1993).

The examples above all represent analytical models for combining expert judgment. These are called *mechanical aggregation methods*, because they consist of processes or analytical models that operate on the individual judgments to produce a single forecast, probability, or risk assessment. In contrast, *behavioral approaches* attempt to generate agreement among the experts by having them interact in some way. Among the most well known of these are Delphi (Linstone and Turoff, 1975) and nominal group technique (Delbecq et al., 1975). In contrast to mechanical aggregation, some behavioral methods aim to produce *consensus*, a genuine agreement among the experts. Behavioral methods also rely implicitly on a determination of relative accuracy and dependence among the experts. As information is shared, it is anticipated that the better arguments and information will be more important in formulating the group's consensus, and that redundant information will be discounted. In contrast to the mechanical methods, though, attention to relative accuracy and dependence is not explicit in behavioral approaches, hence the ability to scrutinize and justify results from behavioral combining procedures may be reduced.

In many cases, decision makers want to avoid the difficulty of comparing experts; making judgments of experts' relative accuracy is difficult, and judgments of dependence are even more so. Often judgments will be combined by means of a simple average, giving all of the experts equal weight. *It is important to realize that doing so is tantamount to a judgment that the experts are indistinguishable.* In many cases this may be reasonable; the experts may indeed be equivalent in the eyes of the decision maker or the assessment team, or it may be politically infeasible to treat the experts in any other way. Even when the experts are treated this way, though, they may still be dependent, and the degree of the dependence can and should have an important impact on the nature of the combined judgment, regardless of the specific aggregation method. Clemen and Winkler (1985) show that dependence, even among indistinguishable (or exchangeable) experts, can have a substantial effect on the amount of information in the combined probability distribution, as measured by the spread of that distribution.

The next section discusses available behavioral methods for combining expert judgments. Considerable empirical work has been done to evaluate the performance of these methods, and pertinent studies are reviewed. The following section turns to mechanical methods. Several approaches for mechanical aggregation have been proposed; the section presents a number of these models. A substantial amount of empirical work has been accomplished to determine the performance of mechanical aggregation methods, and relevant work is reviewed. The objective of maximizing the amount of available information provides a framework for comparing the various methods that are available.

3.2.1 Behavioral Aggregation Approaches

In a behavioral combination approach, experts interact in some way. Such interaction may involve group meetings with face-to-face interaction among the experts, but may involve other kinds of

interaction as well, including interaction via computer or simply the sharing of information or probabilities. When interaction is face-to-face, it can involve the assessment of probabilities or forecasts or just the discussion of relevant issues and ideas without any formal judgmental assessment. Furthermore, these activities can be interspersed with judgmental assessments and sequenced in various ways. Although a few fairly specific approaches to behavioral aggregation have been around for some time, sticking to those approaches is not necessary; researchers can tailor a behavioral aggregation method for a particular application.

The simplest and perhaps most purely behavioral approach is to bring the experts together and ask them to come up with a single forecast, probability, or probability distribution that represents their combined judgments. They can discuss the issues at length and debate about what the group judgments should be. Presumably each expert brings some relevant information to the table, and the discussion provides for the sharing of that information. In an ideal world, this sharing of information would lead to agreement on the forecasts or probabilities of interest, and under certain assumptions it has been shown that agreement should be reached (e.g., Aumann, 1976). In practice, though, the experts generally will not agree. Any group judgments that come out of the discussion will have required the experts to negotiate and compromise in some way. In this sense, the group judgment may not be a perfect reflection of any one expert's thoughts, only an opinion that each expert can go along with for the sake of the group appearing to speak with one mind.

Potential problems associated with group interaction have been pointed out in the literature (e.g., Janis and Mann, 1977). The interaction may not be even-handed. Some individuals tend to dominate discussions while others are reluctant to express their views. In some groups, new ideas may be discouraged, and the group may focus on a limited set of information. The phenomenon of group polarization may occur, in which a group tends to adopt a position more extreme than its information warrants (Plous, 1993). Hogarth (1977) notes, though, that there is no reason interaction processes must be dysfunctional. One way to deal with such problems is to have experienced analysts serve as facilitators for the group. The facilitators can encourage open exchanges and try to get everyone in the group involved in the discussion or can go beyond that to structure the process, drawing out information systematically and leading the group through the judgmental process carefully. This facilitation/structuring is in the spirit of decision conferencing (Phillips, 1984; Phillips and Phillips, 1990), which has been designed with decision making in mind but can be used equally well for group probability assessment (Phillips, 1987).

One of the oldest approaches to structuring group judgments, the Delphi method, requires indirect interaction (Dalkey, 1969; Linstone and Turoff, 1975; Parenté and Anderson-Parenté, 1987). There are different variations of Delphi, but experts typically make individual judgments, and all judgments are then shared anonymously (often with some summary measures). The sharing can be done by mail or some other form of distribution; with computing advances it makes sense to send the information over a computer network. Each expert is then given the opportunity to revise his or her judgments, and the process can be repeated if desired. Again, in an ideal world agreement would be reached after a few rounds. Typically this is not the case, and the combining problem still remains. When the iterative Delphi process concludes (often after just one revision), the resulting expert judgments typically are combined mechanically.

Another method that has been proposed is the Nominal Group Technique (Delbecq et al. 1975). In this procedure, experts are brought together and first record their judgments (and supporting comments, if appropriate) individually, then present them to the group (again individually). The next step

is group discussion with the assistance of a facilitator, followed by the experts reconsidering their individual judgments. Since the resulting sets of expert judgments may not agree completely, a mechanical aggregation procedure may be used in the final step to arrive at a combined judgment. The Nominal Group Technique shares some features with Delphi (the assessment of initial and revised expert judgments individually) and some with decision conferencing (face-to-face interaction with discussion led by facilitators). Lock (1987) proposed an approach similar to the Nominal Group Technique with added stress on defining the task carefully and encouraging multiple advocacy to legitimize alternative viewpoints.

The methods described above have the objective of providing a single forecast, probability, or probability distribution as a result of the interaction. As indicated, it may be necessary in some cases to report final individual judgments (probably highly dependent after the interaction) along with a mechanically combined judgment. It is anticipated that the discussion among the experts will include issues of common information and relative quality of experts' information bases. Delphi and nominal group technique are structured in such a way that some such discussion is likely to occur in the course of the discussion.

A recently proposed aggregation method described by Kaplan (1992) is designed to account explicitly for the information available to the group. In Kaplan's approach, a facilitator/analyst first leads the experts through a discussion of the available information. The objective of this discussion is to determine the consensus body of evidence upon which the judgment should be made. When consensus is reached regarding the relevant information, the analyst proposes a probability distribution for the event, quantity, or parameter of interest, conditioned on the consensus body of evidence. At this point, the analyst must obtain assurance from the experts that the evidence has been interpreted correctly in arriving at the probability distribution. Because different experts may interpret the evidence in different ways, the group judgment may differ from the individual experts' judgments and may result from something like a negotiation process among the experts.

3.2.1.1 Accuracy of Behavioral Approaches: Experimental Results

Perhaps the best known result from the behavioral group-judgment literature is group polarization, the tendency of groups to adopt more extreme positions than would individual members. "Groupthink" (Janis, 1982) is an extreme example of this phenomenon. According to Plous (1993), hundreds of studies of group polarization have been performed over the years, with the consistent conclusion that after discussion, a group will typically advocate a more extreme course of action than they would if acting individually or without discussion.

The results on group polarization would appear to pose some danger for using behavioral judgment-aggregation methods. However, it is important to realize that the results on group polarization apply primarily to unstructured group discussions. It has been shown that group polarization can be deterred by such measures as delaying commitment of the group, spreading power among members, seeking additional information, and encouraging conflict among members (see Park, 1990). In addition, polarization has been studied primarily in the context of group decision making as opposed to group judgments. When looking at group judgments, polarization would suggest that groups might tend to be overconfident about their conclusions or might tend arrive at a consensus that displays more bias than individual judgments would.

A number of studies have examined the accuracy of group judgments. In a review, Hastie (1986) considers quantity estimation (comparable to point forecasting), problem solving, and answering almanac questions. Over all three areas, groups tend to perform better than the average individual, but the best individual in a group often outperforms the group as a whole. Looking only at quantity estimation (most pertinent for risk-assessment studies), the conclusion was that the groups were only slightly (1/8 of a standard deviation) more accurate than individuals on average. More recently, Sniezek and Henry (1989, 1990) have produced experimental evidence that the group's advantage in quantitative estimation may be somewhat greater than was reported by Hastie.

A related finding is that group accuracy often depends on the rules used by the group to arrive at a single judgment. Sniezek (1989) compared five types of group-aggregation methods. Four were behavioral-aggregation methods, and the fifth was a simple average of individual judgments. All four of the behavioral methods were more accurate than the average, but of those four, the best results by far were obtained by the dictator rule, in which the group selects on the basis of discussion a single individual whose opinion the group adopts. In an interesting twist, though, the chosen spokesperson always modified his or her opinion to be closer to the group average, thereby decreasing group accuracy slightly. In this case, the subjects again were college students and the task was sales forecasting.

One of Sniezek's five group techniques was Delphi, and her results were similar to earlier studies on the accuracy of Delphi (e.g., Dalkey, 1969; Dalkey and Brown, 1971). These studies showed that forecasts of individuals converged and that the Delphi technique performed slightly better than a similar procedure with face-to-face open interaction groups. Using bankers as subjects, Brockhoff (1975) found the same results for almanac questions but discovered that for economic forecasting the face-to-face interaction provided better forecasts than Delphi. More recent work on Delphi has produced mixed results; see Hastie (1986) and Parenté and Anderson-Parenté (1987). The conclusion appears to be that Delphi has no distinct advantage over other behavioral combination methods.

3.2.1.2 Behavioral Approaches and Information: Summary

In summary, carefully executed behavioral aggregation methods are capable of incorporating individual judgments into a single forecast or probability distribution that may be more accurate (i.e., more informative) than any of the individual judgments. Following the theme of the chapter, though, it is clear that the quality of the experts' information is not always explicitly considered. Two exceptions are Sniezek's dictator approach and Kaplan's approach (Kaplan, 1992). In the dictator approach, group members choose a spokesperson, presumably at least in part on the basis of superior knowledge. In Kaplan's approach, an attempt is made to clarify the information base explicitly. In neither, though, are formal judgments about the quality of the information both made and used explicitly in the formation of the group judgment. For example, the dictator is chosen on the basis of unstructured discussion and may then make an unstructured adjustment of his or her judgment based on other members' judgments. Kaplan's analyst simply proposes a consensus forecast or distribution, and adjusts it as necessary to gain approval of the experts.

Finally, behavioral aggregation methods do not generally aim to provide more than the single group judgment. The individual judgments may contain information, though, and should be reported as well. Consider two situations: one set of experts differs widely in their opinions, but another set has very similar individual opinions. Even though a behavioral aggregation may yield the same combined judgment, a decision maker might consider the two cases to be very different. In the former case, it might be important to explore the sources of the disagreement more fully, or it may be appropriate to take a

sensitivity-analysis approach to the use of the judgments. In the second case, the decision maker may be quite satisfied with the group judgment but may also want to consider the possibility that the experts may be highly dependent.

3.2.2 Mechanical Aggregation Approaches

Mechanical aggregation methods for expert probabilities have been in existence and studied for many years. This section reviews some of the mechanical aggregation methods. Much of the work has focused on the aggregation of probabilities of a single event (e.g., the occurrence of an earthquake on a specific fault), and it is briefly reviewed here. Although in principle it is possible to extend such models directly to the aggregation of probability distributions for continuous random variables, the tendency has been to create special models specifically for the aggregation of probability distributions and densities. Because risk analysis often deals with entire probability distributions, the main focus of this chapter is on recent developments in the mechanical aggregation of distributions.

3.2.2.1 Axiomatic Approaches

Early work on mechanical aggregation of probabilities focused on axiom-based aggregation formulas. In these studies, the strategy was to postulate certain properties that the combined distribution should follow, and then derive the functional form of the combined distribution as a result. French (1985) and Genest and Zidek (1986) provide critical reviews of this literature, and our summary here draws heavily on these sources.

An appealing approach to the aggregation of probabilities is the *linear opinion pool*, so named by Stone (1961), and dating back to Laplace (Bacharach, 1979):

$$p(q) = \sum_{i=1}^{n} \left[w_i p_i(q) \right] \tag{3-2}$$

where $p_i(q)$ represents Expert i's probability distribution for unknown q, $p(q)$ represents the combined probability, and the w_i are such that $p(q)$ is a probability distribution. A similar formula exists for pooling density functions. The linear opinion pool clearly is no more than a weighted linear combination of the probabilities, and as such it is easily understood and calculated.

The linear opinion pool satisfies a number of seemingly reasonable axioms. Most importantly, though, it has been shown to be the only combination scheme that satisfies the marginalization property (MP). Suppose q is a vector of uncertain quantities, and the decision maker is interested in combining probabilities of just one element of the vector, q_j. According to MP, the combined probability is the same whether one combines the experts' marginal distributions of q_j or combines the experts' multivariate distributions of the vector q and then calculates the marginal distribution of q_j. (Winkler et al., 1978).

The weights in Eq. (3-2) clearly can be used to represent, in some sense, the relative quality of the different experts. In the simplest case, the experts are viewed as equivalent, and Eq. (3-2) becomes the simple arithmetic average (i.e., each $w_i = 1/n$). The determination of the weights is a subjective matter, and numerous interpretations can be given to the weights (Genest, 1984; Genest and McConway, 1990).

Another typical combination approach uses a geometric average and is sometimes called a logarithmic opinion pool. In this case, the combined probability is of the form

$$p(q) = k\prod_{i=1}^{n}[p_i(q)]^{w_i} \tag{3-3}$$

where k is a normalizing constant and the weights w_i satisfy some restrictions to assure that $p(q)$ is a probability distribution. Typically, the weights are restricted to sum to 1. For the case of the expert's judgments being viewed equally weighted, each $w_i = 1/n$.

Equation (3-3) satisfies the principle of external Bayesianity (EB). Suppose a decision maker has consulted the experts, has calculated $p(q)$, but has subsequently learned some new information relevant to q. Two choices are available. One is to use the information first to update the experts' probability distributions $p_i(q)$ and then combine them. The other is to use the information to update the combined $p(q)$ directly. A formula satisfies EB if the result is the same in each case.

Neither Eqs. (3-2) nor (3-3) is entirely satisfactory. Difficulties with the axioms themselves are discussed by French (1985) and Genest and Zidek (1986). Lindley (1985b) gives an example of the failure of both axioms in a straightforward example, with the interpretations that MP ignores important information and that the EB requires that the form of the pooling function not change. In addition, French (1985) points out that *impossibility theorems* exist (along the lines of Arrow's (1951) classic work on social choice theory) whereby a combining rule cannot satisfy at once a set of seemingly compelling desiderata. Moreover, despite the work of Genest and McConway (1990), no foundationally-based method for determining the weights in Eqs. (3-2) and (3-3) is available. Finally, there is no obvious way with Eqs. (3-2) or (3-3) to account for correlation among the experts. French (1985), Lindley (1985b), and Genest and Zidek (1986) all conclude that for the typical risk analysis situation, in which a group of experts must provide information for a decision maker, a Bayesian updating scheme is the most appropriate method.

3.2.2.2 Information Issues and Likelihood Assessment

Before presenting some of the aggregation models, it is appropriate to consider some of the information issues with regard to the Bayesian aggregation models. It is clear that construction of the likelihood function is of paramount importance in this procedure. As mentioned in Sections 3.3 and 3.4, this likelihood function must incorporate information about the quality of the individual expert judgments as well as information about dependence. In contrast to the behavioral aggregation approaches, the Bayesian approach requires that this information be explicitly considered and incorporated into the aggregated distribution in a formal and systematic way.

When data are not available for estimating parameters in the likelihood function, judgments about expert quality and dependence must be made subjectively. The most appropriate source for these judgments is the assessment team, the group of individuals who guide the experts through the assessment process. Composed of individuals knowledgeable about both substantive issues as well as the elicitation process, the assessment team is in the perfect position to make the necessary judgments regarding the experts' information.

3.2.2.3 Comparisons Among Mechanical Methods

Some evidence is available regarding the relative performance of various mechanical aggregation methods. In an early study, Stael von Holstein (1972) examined probabilities for stock market prices. Simple averages of probabilities revealed that the average stock market expert did better than any subject, and only 29 percent of the subjects beat the average banker despite the fact that eight of the ten bankers were in the lowest 30 percent of subjects in terms of average scores. In addition to the simple average, weighted averages with several weighting schemes were tried. Most of these mechanical methods performed similarly, with weights based on rankings of past performance slightly better than the rest.

In the study by Seaver (1978) described earlier, simple and weighted averages of the individual probabilities were considered. The performance of the different combining methods was similar, and Seaver's conclusion was that simple combination procedures, such as an equally weighted average, produces combined probabilities that perform as well as those from more complex aggregation models. Clemen and Winkler (1987) reported similar results in aggregating precipitation probability forecasts.

In a follow-up study comparing mechanical aggregation methods, Clemen and Winkler (1990) studied the combination of precipitation forecasts using a wider variety of mechanical methods. Their conclusions showed that the best method was one of the more complex methods that was able to account for dependence among the forecasts. Although they did not explicitly consider a simple average, they did include a weighted average that resulted in weights for the two forecasts that were not widely different. This weighted forecast performed almost as well as the more complex scheme.

Winkler and Poses (1993) report on the combination of experts' probabilities in a medical setting. For each patient in an intensive care unit, four individuals (an intern, a critical care fellow, a critical care attending, and a primary attending physician) assessed probabilities of survival. All possible combinations (simple averages) of these four probabilities were evaluated. The best combination turned out to be an average of probabilities from the two physicians who were simultaneously the most experienced and the least similar, with one being an expert in critical care and the other having the most knowledge about the individual patient.

All of these results are consistent with the general message that has been derived from the vast empirical literature on the combination of point forecasts. The message is that, in general, simpler aggregation methods perform better than more complex methods. Clemen (1989) discusses this literature. In some of these studies, taking into account the quality of the expert information, especially regarding relative precision of forecasts, turned out to be valuable.

All of the above studies focus on the combination of point forecasts or event probabilities, and mechanical methods studied have been either averages or something more complex in which combination weights were based on past data. Does the result that simpler methods work better than more complex methods carry over to the aggregation of probability distributions, especially when the quality of the expert opinions must be judged subjectively? Little specific evidence appears to be available on this topic. Clemen et al. (in press) reported that Winkler's (1981) normal model and the more complex conditional-distributions model (Clemen and Winkler, 1993) performed at about the same level. These results are consistent with a number of other studies on the value of decomposing judgments into smaller and more manageable assessment tasks. Wright et al. (1988), though, found little difference between holistic and decomposed probability assessments; on the other hand, Hora et al. (1993) provide empirical support for decomposition in probability assessment. With regard to decomposition and the assessment of point

estimates, Armstrong et al. (1975) and MacGregor et al. (1988) found that decomposition was valuable in improving the accuracy of those estimates. Morgan and Henrion (1990) review the empirical support for decomposition in probability assessment, and Bunn and Wright (1991) do the same for forecasting tasks in general. A tentative conclusion is that, for situations in which the aggregation must be made on the basis of subjective judgments, appropriate decomposition of those judgments into reasonable tasks for the assessment team may lead to better performance.

3.2.2.4 Accuracy of Mechanical Aggregation

Mechanical aggregation methods have a compelling theoretical basis and explicitly incorporate information about the quality of expert information. The process required to perform mechanical aggregation provides an explicit audit trail of the necessary steps and judgments; after the fact, it is straightforward to describe and justify each step of the elicitation and aggregation process. While this is all appealing, what evidence exists regarding the accuracy of mechanical aggregation methods? First, let's ask whether one should bother with the formal aggregation model and required judgments. Perhaps it suffices to look at the individual expert judgments and aggregate them intuitively, directly assessing a probability distribution for the quantity or event in light of the experts' information. The next section also looks at the comparison of mechanical and behavioral combination methods in terms of accuracy. A limited amount of evidence is available pertaining directly to each of these two accuracy issues. Finally, it looks at evidence regarding the relative performance of different kinds of mechanical combinations.

3.2.2.5 Mechanical Versus Intuitive Aggregation

Hogarth (1987, Chapter 3) discusses the difficulty humans have in combining information from different data sources. Although his discussion covers the use of all kinds of information, his arguments apply to the aggregation of expert opinion. Among other phenomena, Hogarth shows how individuals tend to ignore dependence among information sources, and he relates this to Kahneman and Tversky's (1972) representativeness heuristic. In a broad sense, Hogarth's discussion is supported by psychological experimentation showing that expert judgments tend to be less accurate than statistical models based on criteria that the experts themselves claim to use. Dawes et al. (1989) provide a review of this literature.

Although little evidence is available comparing mechanical and intuitive combinations of probability judgments, a recent study by Maines (1994) has examined the intuitive combination of point forecasts. In this study, the results are consistent with Hogarth's discussion.

Clemen et al. (in press) also study the aggregation of point forecasts. However, they use Winkler's (1981) normal model and Clemen and Winkler's (1993) conditional-distributions model, and they compare the probability distributions derived from these models with intuitively assessed probability distributions. Thus, the comparison is between an intuitively assessed posterior distribution and posterior distributions based on mechanical aggregation methods. Although their sample size is small, the results suggest that the mechanical methods are somewhat better in terms of performance on estimating answers to almanac type of questions than the intuitive assessment, and the authors speculate that this is due to the structured nature of the assessments required in the mechanical-aggregation models.

3.2.2.6 Mechanical Approaches and Information: Summary

In contrast to behavioral aggregation methods, mechanical methods have the capability of formalizing and using information about the quality of the experts' opinions. In data-based aggregation

models, this information is developed using data on performance of the experts in previous situations. In subjectively based methods, information about the quality of the expert judgments must be assessed; it can be argued that the assessment team is in the best position to make these judgments.

The empirical work to date has focused typically on point forecasts or single-event probabilities, and the mechanical methods have typically been either data-based or simple averages. In these studies, the simpler approaches have generally performed well relative to the more complex methods. In these cases, it would appear that the additional information about the quality of the individual forecasts, while appealing in principle, does not always result in improved performance. These studies do not, unfortunately, directly address the precise issue that needs to be addressed. For the purpose of the typical risk analysis in which probability distributions are to be combined, but no past data are available on the basis of which to perform the aggregation, the real question is, "What is the best way to combine the judgments?" No empirical study has been done to address this question directly. However, work has shown the empirical value of decomposition in probability assessment. These results suggests that structured methods for assessing expert judgment quality can be of value in mechanical aggregation procedures.

Finally, mechanical aggregation methods by their nature provide documentation on all phases of the aggregation process. Individual expert judgments are normally reported, as are the judgments by the assessment team regarding expert judgment quality. In this regard, an important dimension of value is the ease of understanding and making the judgments. In this respect, both Winkler's (1981) normal model and the Jouini and Clemen (1994) copula model appear to have an advantage. These models require only straightforward calibration assessments of the individual expert opinions and an assessment of dependence among those opinions.

3.2.3 Mechanical Versus Behavioral Aggregation

Most of the research comparing mechanical and behavioral aggregation has focused on comparisons with a simple average of forecasts or probabilities rather than with more complicated mechanical combination methods. Results from these comparisons have been mixed. For example, Rohrbaugh (1979) found that forecasts of college students' grade-point averages were more accurate in Delphi groups and groups with face-to-face interaction than when obtained by taking simple averages of the individual group members' forecasts. Hastie (1986), Hill (1982), and Sniezek (1989) reached similar conclusions as described above. However Lawrence et al. (1986) found that mechanical combination was superior to behavioral combination in their forecasting study, and Flores and White (1989) conducted an experiment in which mechanical and behavioral combinations performed similarly. Goodman (1972) had college students assess likelihood ratios individually and in groups, and the behavioral combination was slightly better than the mechanical combination.

In an extensive study comparing mechanical and behavioral combination procedures, Seaver (1978) used student subjects and had them assess discrete and continuous probability distributions for almanac questions. Seaver used different conditions: individual assessment, Delphi, the Nominal Group Technique, free-form discussion, and two other variants differing in rules for information exchange and discussion. In addition to the behavioral procedures, he investigated simple and weighted averages of the individual probabilities. Seaver concluded that interaction among the assessors yielded no overall improvement in the quality of the assessed probabilities, although it did lead to the subjects being more satisfied with the aggregation results.

3.3 UPDATING SUBJECTIVE PROBABILITY DISTRIBUTIONS

This section briefly discusses the problem of updating a quantitative expert judgment after new information has been obtained or after the experts have learned more. Although Bayes' Theorem provides an updating scheme for an individual expert, the problem is more difficult when a decision maker consults multiple experts. Theoretically this is not difficult, but in practice it requires careful statistical modeling of all of the information sources. It is important to note at the outset that the discussion in this section is somewhat speculative. Compared to the problem of combining information, issues of updating combined distributions and of combining expert judgments with data have received relatively little study.

3.3.1 Updating Expert Judgments Based on New Information

As stated earlier, subjective probabilities represent an individual's degree of belief based on knowledge (educated guess) with respect to either an event occurring or the numerical value of a parameter. Subjective probabilities are dependent on the information available to the expert, and as such, provide a snapshot of the state of knowledge at a given time. Therefore, as new information becomes available and the state of knowledge changes, so should the subjective probabilities.

However, it should be noted that the updating of subjective probabilities needs to be done within the axioms of probability theory. For a set of mutually exclusive events, the sum of the corresponding probabilities must be unity. Therefore, if p(A) represents the subjective probability about the occurrence of an event A, then 1-p(A) is the probability of its complement (i.e., A does not occur). When the state of knowledge about A changes due to the availability of new information, the probability of A is expected to change. As the probability of A changes, so does the probability of its complement. This mean that if an expert updates his/her probability about event A, then he or she must maintain self-coherence and update the probability of all other events in the set, for failure to do so will violate the basic tenets of probability theory.

An alternative to directly updatingprobability distributions (basically repeating the expert judgment elicitation) is to use Bayes' Theorem.

Bayes' Theorem provides the framework for updating probabilities in light of new information. To illustrate how Bayes' Theorem could be applied, consider the example of conceptual model uncertainty. Scant data at a candidate HLW repository site renders the modeling problem to be underdetermined and, consequently, several conceptualizations of the site behavior are possible that are consistent with the available data. These conceptualizations are manifested as multiple conceptual models, and several groups of investigators (e.g., Chhibber et al., 1991; Heger and Hill, 1993) have suggested that a probability can be assigned to each model representing the degree of belief based on knowledge that the conceptual model is valid. Let's assume, for the sake of presenting an illustrative example, that the notion of assigning a probability to a conceptual model is acceptable.

Assuming that (i) there are three possible conceptual models and, (ii) given the available information, one feels that the three models are equally likely, then the prior probability for each conceptual model $p(M_i)=1/3$ (for $i=1,2,3$). The next logical step should be to collect data that would allow discrimination between the models. The new data, if relevant, would strengthen one of the conceptual models at the expense of the others, and the posterior probabilities would not be all equal; instead, one of the models will end up with a posterior probability greater than 1/3 and the other two with

posterior probabilities less than 1/3. In principle, this process can be repeated several times until one of the conceptual models emerges as the strongest one represented by a fairly high posterior probability, while the other two end up with very low posterior probabilities.

The process just described can now be put in the context of Bayes' Theorem. The expert judgments are initially used to estimate the prior probability for each possible conceptual model. New data are used by the experts, in principle, to construct a new likelihood function for each conceptual model. Substituting the prior probability and the new likelihood function for each conceptual model into Bayes' Theorem a posterior probability is obtained for each model. The process can be repeated several times until the analysts feel that the problem has been resolved and a single conceptual model has survived as the preferred one. For each new iteration of the process, the posterior probability from the previous iteration becomes the new prior probability, and a new likelihood function is constructed to obtain a new posterior probability. It should be noted that, to be consistent with probability theory, each iteration needs to independently determine a new posterior probability for each of the conceptual models, and the sum of these probabilities must be unity.

The same basic approach can be used to update expert judgments used to estimate the PDF of a parameter X. In this case, probabilities are replaced with PDFs for X. New data are used by the experts to construct the likelihood function $p(E|X)$ where E represents the new data. The likelihood function and the prior PDF are substituted into Bayes' Theorem and the posterior PDF of X conditional on the new data E, $p(X|E)$is calculated.

It should be mentioned that in practice constructing the likelihood function is not trivial, and this is where much of the effort is expended when applying Bayes' Theorem (Eslinger and Sagar, 1989; Bonano and Apostolakis, 1991; Chhibber et al., 1992).

3.3.2 Bayesian Approaches

A Bayesian view of the general information-aggregation problem began with Winkler's (1968) work. In many ways, Winkler provided a Bayesian framework for thinking about the combination of information and ways to assess differential weights. This work was followed by Morris (1974, 1977), who established a clear Bayesian paradigm for aggregating information from experts. The notion is straightforward. If n experts provide information g_1, ..., g_n to a decision maker regarding some event or quantity of interest q, then the decision maker should use Bayes' Theorem to update a prior distribution $p(q)$:

$$p^* = p(q|g_1,\cdots,g_n) = \frac{p(q)\ L\ (g_1,\cdots,g_n|q)}{p(g_1,\cdots,g_n)} \qquad (3\text{-}4)$$

This general principle can be applied to the aggregation of any kind of information, ranging from the combination of point forecasts or estimates to the combination of individual probabilities and probability distributions. Resting on the solid ground of probability theory, including requirements of coherence as described by de Finetti (1937) and Savage (1954), Morris's Bayesian paradigm provides a compelling framework for constructing aggregation models. Over the 1970s and 1980s, attention has shifted from the axiomatic approach to the development of Bayesian combination models.

At the same time that it is compelling, the Bayesian approach is also frustratingly difficult to apply. The problem is that the decision maker must construct the likelihood function $L(g_1,\ldots,g_n \mid q)$. This function amounts to a probabilistic model for the information $g_1,\ldots,g_n$, and as such it must capture the interrelationships among q and $g_1,\ldots,g_n$. In particular, it must account for the precision and bias of the individual g_is, and it must also be able to model dependence among the g_is. For example, in the case of a point forecast, precision of g_i refers to the accuracy with which Expert i forecasts q. Bias is the extent to which the forecast tends to fall consistently above or below q. Dependence might be thought of as the extent to which the forecast errors for different experts are interrelated. For example, if Expert i overestimates q, will Expert j tend to do the same?

In developing his model, Morris (1974, 1977) attended specifically to the aggregation of probabilities and probability distributions. Because of the difficulty of creating an appropriate likelihood function, considerable effort has gone into the creation of off-the-shelf models for aggregating single probabilities (e.g., Lindley, 1985b; Clemen and Winkler, 1987), and probability distributions (Winkler, 1981; Lindley, 1983; Mendel and Sheridan, 1989). Many of these methods rely on some version of a multivariate-normal model for aggregating information. All rely on past data as a basis for estimating parameters in an appropriate likelihood function.

Although use of past data is valuable and often an important part of the scientific process, it often is the case in risk analysis that past data are not available for parameter estimation in aggregation models. Some effort in the literature has been made to accommodate risk analysis in this respect. For example, Genest and Schervish (1985) and West (1994) study Bayesian aggregation of probabilities when the decision maker is unable to specify more than a few moments of the distribution. More recently, Clemen and Winkler (1993) and Jouini and Clemen (1994) have developed approaches explicitly designed to facilitate the subjective assessment of the likelihood function. Both methods are highly flexible in their ability to accommodate arbitrary distributions, thereby avoiding the constraints of the multivariate-normal model. Clemen and Winkler's model is meant for aggregating point forecasts specifically, while Jouini and Clemen's approach can aggregate both point estimates and probability distributions.

3.4 COMBINING EXPERT JUDGMENTS WITH OTHER SOURCES OF INFORMATION

One of the reasons for using expert judgments may be to augment data because it may not be practical to obtain all the data that would be required to fully characterize a variable of interest. For example, Eslinger and Sagar (1989) and Merkhofer and Runchal (1989) discuss the use of expert judgments to supplement data about hydrologic parameters at a HLW disposal site. One of the concerns is that an inordinately large number of boreholes are necessary to characterize the hydrology of a site given its heterogeneity and the associated spatial variability of hydrologic parameters. However, drilling a large number of boreholes to collect the data is not only impractical in terms of resources needed, but, more importantly, could compromise the integrity of the site in terms of retarding the migration of radionuclides through the geologic medium. In this case, limited data on the value of the hydrologic parameters will be available, and that data will likely be supplemented with expert judgments (see Eslinger and Sagar, 1989; Merkhofer and Runchal, 1989).

3.4.1 Bayes' Theorem for Combining Expert Judgment and Data

It is well documented in the literature that probabilities quantifying the uncertainty of an event, parameter, scenario, etc. based on expert judgments are a representation of the degree of belief based on knowledge about the occurrence of the event or scenario, or the numerical value of the parameter (e.g., Lindley, 1985b; Eslinger and Sagar, 1989; Merkhofer and Runchal, 1989; Apostolakis, 1990; Morgan and Henrion, 1990; Smith, 1990). It is also accepted that probabilities representing degrees of belief, also known as subjective probabilities, are conditional on the available information; that is, as the knowledge base about a given event, scenario, or parameter changes, the subjective probabilities should be updated to incorporate the new or additional information.

In the scientific community, Bayes' Theorem is accepted as the framework for: (i) combining expert judgments with other sources of information, and (ii) updating subjective probabilities when new or additional information becomes available. Bayes' Theorem can be written as

$$p(X \mid E) = \frac{p(E \mid X)p(X)}{K}, \tag{3-5}$$

where

$p(X)$	=	prior PDF about the event, scenario, or parameter X
$p(X \mid E)$	=	posterior PDF about X, given the new evidence E
$p(E \mid X)$	=	likelihood function that evidence E would be obtained given X
K	=	proportionality constant to ensure that area under the posterior PDF is unity

To illustrate the manner in which Bayes' Theorem can be used to combine expert judgments with available data, let's consider the following situation. The project team has constructed the prior PDF, $p(X)$, of parameter X using the available hard data and realizes that this PDF does not capture the current state of knowledge or uncertainty about X. Therefore, the project team wishes to obtain expert judgments to supplement the hard data. The judgments obtained from the experts provide evidence relevant to the numerical value of X based on an analysis of the hard data available (Kaplan, 1992). Using, for example, the approach outlined by Kaplan (1992), this evidence can then be used to construct the likelihood function $p(E \mid X)$. Since, in principle, $p(X)$ and $p(E \mid X)$ are both known, the expert judgments can be combined with the available hard data by substituting the prior PDF and the likelihood function into Bayes' Theorem. The result is the posterior PDF $p(E \mid X)$ that represents the current state of knowledge about the numerical value of X, the posterior PDF $p(E \mid X)$, where E represents the expert judgments.

Statistically modeling the relationships among the data and expert judgments is a very complicated problem. Moreover, every situation requires special modeling treatment. Consequently, there is no stream of literature that deals with the aggregation of judgment and data specifically. In particular, it is noteworthy that a recent report by the National Research Council (1992) on combining information mentioned the problem, but provided no specific examples and only three general references.

From a practical point of view, the decision maker has the same choices outlined in the previous section. One of these, of course, is to proceed with a behavioral consensus exercise, bring the experts together, provide them with the data, and ask them to produce a consensus distribution for the parameter of interest. Again, all of the advantages and disadvantages of the behavioral procedures are relevant. In

particular, if the only output is the consensus distribution, the decision maker receives a somewhat sparse information set that includes no individual distributions and no information about the quality of the experts' judgments, the quality of the data, or the relationships among them.

An alternative is to proceed with a statistical approach along the lines of Eq. (3-4). Although the specifics will vary from one situation to the next, the approach can be demonstrated by means of an example. Reckhow (1988) studied the response of brook trout populations to the acidification of selected North American lakes. The data included readings of the presence or absence of trout, the calcium level, and the acidity (pH) in each of 45 lakes. In addition, an expert made predictive judgments of the probability of the presence of trout for each of 20 lakes characterized as (calcium, pH) pairs. The model was a logistic regression model for the presence or absence of fish, with the independent variables being ln(Ca) and pH. Assuming that the data and the judgments are independent, a posterior distribution for the regression parameters can be found using procedures in Zellner (1971) and Winkler et al. (1978). If the data and the expert judgments are dependent (for example, if the expert had seen the data or results based on the data), then this dependence should be modeled. Winkler and Clemen (1989) show how this can be accomplished.

In this example, the combination of data and expert judgments follows a procedure similar to ridge regression in econometrics. Lindley and Smith (1972) discuss a Bayesian interpretation of ridge regression, indicating that the procedure is tantamount to a Bayesian model in which the independent variables are considered equivalent in terms of their effects on the response variable. Clemen and Winkler (1986) discuss their forecast-aggregation model in these terms, and the concept can be extended to include shrinkage estimators in general (National Research Council, 1992).

In some cases, it is practical to include expert judgment directly in a statistical model. For example, Blattenberger and Fowles (1994) show in the context of avalanche forecasting how to use expert judgment in tandem with additional data in order to develop an improved forecast of avalanche danger.

As the above discussion indicates, there are a number of results and procedures available that may be useful for combining data and judgment in formal ways. The problem, however, is that the specifics of each situation will dictate how the combination should be done in that particular instance. General principles that can be stated are that statistical models should be formulated carefully to account for all of the appropriate stochastic characteristics of the system (including dependence where necessary) and that, once the modeling has been done, Bayes' Theorem provides a vehicle for deriving the required posterior distribution. The required modeling could involve judgments regarding the relevance of the experts' judgments, the quality of the data, and the relationships among them. By going through a formal modeling procedure, these issues must be faced explicitly and information related to these issues can be readily delivered to the decision maker.

Apostolakis et al. (1991) used Bayes' Theorem to combine different sources of information as an approach to estimate the probability of occurrence for scenarios to be considered in the PA of HLW repositories. For example, for a climate change scenario they combined expert judgments obtained from the interpretation of paleoclimatic data with the results from the application of global climate models. Redus (1994) proposes combining data derived from models, field experiments, and expert judgments with Bayes' Theorem to produce parameter values for waste system characterization.

3.5 SUMMARY

The selection of the appropriate elicitation technique for a given type of judgment is of critical importance. The elicitation technique should possess at least the following attributes: (i) it should be able to yield the judgments in a manner that will facilitate their subsequent use in the originally intended fashion, (ii) its mechanics should be easily understood by the experts, and (iii) it should be easy to implement and to explain to outsiders. The conduct of the elicitation sessions themselves is extremely important and should be well planned and practiced ahead of time.

When the situation in a technical analysis requires aggregation of the judgments of multiple experts, the assessment team faces the choice of behavioral or mechanical methods. The team is often responsible for selecting the method. While behavioral methods may be expedient, mechanical algorithms have the distinct advantage of traceability, that is, the individual judgments are documented prior to the aggregation. Bayes' Theorem is an available technique of both updating previously elicited expert judgments as well as combining expert judgments with other types of data.

4 EVALUATING THE QUALITY OF EXPERT JUDGMENTS

4.1 ISSUES IN THE QUALITY OF EXPERT ELICITATIONS

Two areas are of particular importance in evaluating the quality of the judgments obtained through expert elicitation: evaluating the procedure used and evaluating the quality of probability judgments. The first area involves a systematic evaluation of the overall elicitation process itself. Each step in the process can be evaluated with specific criteria, which build a case for the credibility elicitation. The second area involves evaluating the quality of probability judgments because of the preponderance of expert judgments that are obtained in the form of subjective probabilities.

4.2 EVALUATING THE FORMAL EXPERT ELICITATION PROCESS

Unfortunately, the literature on elicitation and use of expert judgments is consistent in pointing out that each situation is unique and, therefore, the approach utilized to elicit the judgments should be situation specific (see Chhibber et al., 1992). Morgan and Henrion (1990, p. 158) state that "... there is no single correct protocol for expert elicitations and ... different designs may be better in different specific contexts ..." and that "... the process of elicitation should not be approached as a routine procedure amenable to cookbook solutions." It seems that often the suggested elicitation approach and the techniques and methods employed to elicit the judgments depends not only on the specific question/issue the experts are to address, but also on the preferences of the individual(s) conducting the elicitation (normative expert) and the individual(s) that will use the judgments (Morgan and Henrion, 1990). Some prefer to use event trees to organize the judgments (e.g., McGuire, 1990; 1992), others like probability networks (e.g., Heger and Hill, 1993), and yet others advocate the use of influence diagrams (e.g., Hong and Apostolakis, 1993).

When using multiple experts, some elicit the judgments of individual experts and attempt to combine these latter (e.g., DeWispelare et al., 1993), others hold group sessions to arrive at consensus judgments (e.g., Dalrymple, 1990; Stephens and Goodwin, 1990), and others use a combination of both approaches (e.g., Thorne, 1992; 1993). Different protocols have been proposed and used to elicit the judgments (see Cambridge Decision Analysts Limited, 1992 for a summary), and yet others have suggested that "expert knowledge or information," as opposed to expert judgments or opinions, should be the objective of the elicitations (e.g., Kaplan, 1992).

This state of affairs notwithstanding, it is still possible to provide some general guidelines regarding the elicitation and use of expert judgments. It should also be possible to develop general guidelines that will allow those reviewing the elicitation and use of expert judgments to assess the quality and credibility of the judgments; this assessment could be conducted by attempting to answer the following basic questions:

- How was the decision made to use expert judgments?
- How were the expert(s) selected?
- How were the expert(s) trained?
- How was the elicitation conducted?

- How were the judgments documented?

- How were the expert judgments combined, if applicable?

- How were the judgments used?

Each of these basic questions can be further decomposed into a set of more detailed questions or criteria that should help a reviewer reach a conclusion on the quality and credibility of the judgments. Some examples of these detailed questions or criteria are presented in the following sections.

4.2.1 How Was the Decision Made to Use Expert Judgments?

Determining whether or not the use of expert judgments is justified can be as important as determining the quality and credibility of the judgments themselves. It cannot be overemphasized that judgments should be considered in those cases where it has been determined that collection of other specific, directly relevant data is not practical. Therefore, the circumstances that lead to the decision to use expert judgments need to be soundly and defensibly justified. Furthermore, because a formal elicitation is more resource intensive (i.e., takes longer and is more expensive) than an informal one, project staff may opt for the latter. In such cases, the basis for that decision needs to also be clearly articulated and documented. At one point in Miguel de Cervantes' novel *Don Quixote*, Don Quixote tells his sidekick, Sancho Panza, to dress him slowly because he was in a hurry; that is, Don Quixote wanted to make sure that he was dressed properly and that no important piece of clothing was left out because he could not afford to have to return home. Similarly, situations in which potentially credible judgments are discarded because the degree of formalism used to acquire them was not adequate, given the importance and complexity of the question/issue the experts addressed, should be avoided. Care must be exercised to ensure that the degree of formalism employed to elicit and use the judgments is commensurate with the importance of the issue. In order to determine whether or not a sound justification exists for the use of expert elicitation a reviewer could pose the following questions:

- What effort was made to justify the use of expert elicitation relative to data collection, modeling, and/or other types of analyses that could yield the information sought? That is, was expert judgment used because the information was otherwise unobtainable by other means?

- How were the issues that the experts were asked to address defined?

- What effort was made to determine whether the elicitation of the judgments warranted a formal process as opposed to an informal one (i.e., comparison of cost and benefits of the formal versus the informal elicitation of the judgments)?

- Was the rationale for a formal or an informal elicitation of the expert judgments documented?

4.2.2 How Were the Experts Selected?

There is little doubt that, to a large extent, the credibility and acceptability of expert judgments will rest on the credibility of the experts and their acceptability as such. Issues that should be addressed regarding the selection of experts have been discussed extensively in previously published works (e.g., Bonano et al., 1990; Chhibber et al., 1992; Rechard et al., 1993; DeWispelare et al., 1993). By and large, the fundamental criterion for expert selection is that the experts have a demonstrated substantive knowledge of the subject matter of interest. Morgan and Henrion (1990, p. 109) could not have put this more clearly when they stated that "... the most important thing is to find someone with ... substantive expertise. If your expert does not actually know anything about the topic in question, you will never extract anything useful no matter how well calibrated he or she is, and no matter what elicitation technique you use." One could draw a parallel from the judicial review of scientific issues applied to the elicitation and use of expert judgments. In the judicial review, judges, when attempting to decide a close case, pay a great deal of attention to the "bona fides" of the different parties involved (McGarity, 1984). There is little reason to believe that when different experts express opposite points of view in the regulatory process, the situation will be different than in the judicial system.

The quality of the expert selection approach could be ascertained by asking the following questions:

- Were specific selection criteria established and used to identify the experts?
- Was the rationale for the criteria presented?
- Was the rationale for using a single expert versus a group of experts presented?
- Were the selected experts considered the best in their field for the specific types of judgments sought?
- What criteria were used to ensure that the selected experts were, to the extent practicable, free of conflict of interest?
- What criteria were used to ensure that the selected experts capture the appropriate diversity of opinion?
- What criteria were used to ensure the independence of the selected experts?
- Was the expert selection process documented?

4.2.3 How Was Pre-Elicitation Conducted/Experts Trained?

Expert training is a key aspect of any expert judgment elicitation exercise. Training is needed for a variety of reasons. Bonano et al. (1990) gave the following reasons:

(i) Allow the experts to become familiar with the elicitation of judgments and encourage them to provide their opinions

(ii) Allow the experts to practice providing judgments

(iii) Educate the experts in the identification of potential biases and in the use of techniques to deal with the biases

Several other authors, namely Morgan and Henrion (1990), Cooke (1991), and Chhibber et al. (1992), have suggested that training is also needed to:

(i) Assess the calibratability of the experts

(ii) Discuss and clarify the issue that the experts are asked to address

(iii) Present the experts with a common set of information for them to examine and use in the analysis supporting their judgments

The nature and quality of the training given the experts prior to the elicitation can be assessed by answers to the following questions:

- Was training given to the experts to ensure that they were prepared for articulating their judgments as well as the reasons, assumptions, and approaches they used to arrive at the judgments? If so, what was the nature and scope of the training?
- Did the training ensure that the issue the experts were asked to address was clearly defined and understood?
- Were the experts trained in the different elicitation techniques that could be used to obtain their judgments (e.g., probability encoding, utility elicitation, etc.)?
- Did the training include the identification of possible biases and the different techniques that could be employed to minimize the impact of such biases?
- Did the training include practice at elicitation?
- Was any attempt made to calibrate the experts? If so, what calibration approach was used?
- What information, if any, was provided to the expert(s) for him/her (them) to use in arriving at their judgments?
- Was the training documented?

4.2.4 How Was the Elicitation Conducted?

The actual elicitation of the judgments is often the focus of the process. That is, all pre-elicitation activities were aimed at preparing for the elicitation itself. As stated earlier, the elicitation needs to be tailored to the specific question or issue at hand, the type of judgments required (e.g., identification of events and phenomena, subjective probabilities, probability distributions, etc.), the resources available for the elicitation, the availability of the experts, and the idiosyncrasies of the experts,

among others. These factors influence the protocols, methods, and techniques employed in the elicitation. Regardless of the manner in which the elicitation is conducted, reviewers should be able to discern not only the judgments themselves, but also the reasons, assumptions, approaches, and information that each of the experts used. Of particular interest is ensuring that the format of the elicitation does not unduly influence the nature of the judgments provided (e.g., facilitators of group sessions need to ensure that format does not lead to strong personalities in the group dominating consensus judgments).

The manner in which the actual elicitation of the judgments was conducted can be evaluated with the following questions:

- What approach was used to carry out the elicitation(s)? For example, for elicitations involving more than one expert, were the judgments obtained from individual experts, group sessions, or both?
- Were the experts allowed to use information other than that provided to them, if applicable, to arrive at their judgments? If so, did the experts provide a written record of any additional information they used?
- Did the elicitation involve an elicitator properly trained for obtaining the judgments (i.e., trained normative expert)?
- Were the elicitation techniques used appropriate for the task and at the forefront of the state-of-the-art?
- What, if any, post-elicitation activities took place (e.g., feedback provided to the experts)? Did the experts have the opportunity to revise their judgments, and if so, how were these revised?
- How was the elicitation session documented?

4.2.5 How Were the Expert Judgments Documented?

A critical aspect of the use of expert judgments is the documentation of the judgments and the ability to subject these to thorough scrutiny. Not only the judgments themselves should be documented, but perhaps more importantly, documentation should also be provided for (i) the information base used by the experts to arrive at their judgments, (ii) the assumptions and rationales employed by the experts, and (iii) the methods they employed to synthesize the information and their assumptions. Therefore, in order to determine the adequacy of the judgments, a reviewer should be able to find answers to the following questions:

- Is a concise basis for each expert's judgments on record? Does this record indicate the logic, rationals, and thought process used as a basis for the judgments obtained?
- Is it possible to retrieve the basic judgments as well as the information base, assumptions, rationales, and analysis methods used by the experts?

- Can specific judgments be identified with each of the experts (i.e., is the identity of the experts providing the judgments known)?

- Is a permanent record of the elicitation available for future examination?

- Is a detailed record of the entire elicitation procedure used available?

4.2.6 How Were the Expert Judgments Combined?

One of the critical aspects of components of the elicitation of expert judgments, particularly when these are being used to support, directly or indirectly, the demonstration of compliance with regulatory requirements, is the manner in which judgments from a group of experts are combined. The combination of judgments from multiple experts introduces several issues; among these are (i) the tendency for combination to eliminate potentially important diversity of opinion, (ii) the selection of a particular combination method above another, (iii) the preservation of the individual judgments, and (iv) the representativeness of the combined results. It is important to ensure that the combination technique used is appropriate, given (i) the type of judgments elicited and (ii) the expected use of the judgments.

At present there is a considerable amount of debate regarding the selection of a combination method or technique. Recent literature on the subject has suggested many different approaches for combining judgments (e.g., Morgan and Henrion, 1990; Cooke, 1991; Cambridge Decision Analysts Limited, 1992; Chhibber et al., 1992; Heger and Hill, 1993; Thorne, 1992; 1993; Chhibber and Apostolakis, 1993). These approaches range from simple to sophisticated analytical ones, to complex behavioral interactions. Notwithstanding the plethora of suggestions made by different authors, some of which we have listed here, no suggestion seems as sound as that offered by Raiffa (1968). He stated that the selection of a combination method can have a significant impact on the results of the analysis in which the judgments are used. Therefore, he suggested that the sensitivity of the overall analysis results to the combination method used should be examined prior to committing to any one method. This will allow the selection of a specific combination method to be justified.

The aggregation technique used to follow the individual elicitations can be elucidated through the following questions:

- What technique was used to aggregate the individual judgments of the experts?

- If a behavioral aggregation technique was used, is the procedure and specific interaction among the experts documented (e.g., a three-round Delphi with anonymity?).

- If a mechanical technique was used, is the algorithm completely described (i.e., equation and weights for the individual opinions of the experts)?

- Is the aggregation technique documented to the point which allows repeatability?

Regardless of the combination method used, in problems such as the assessment of regulatory compliance of a HLW repository, the judgments from individual experts should be preserved (Bonano et al., 1990). This will allow the regulators and other interested parties to conduct their own assessment on the impact that the combination of judgments can have on the results of the analysis.

4.2.7 How Were the Expert Judgments Used?

Finally, one would like to assess how the expert judgments were used. This assessment should address at least two issues: (i) the relationship between the judgments and other sources of information that serve as input to subsequent calculations, and (ii) the manner in which the analyst(s) manipulated the judgments for use in the calculations. A recent example of the latter is the extensive manipulation of the probability of human intrusion into the WIPP obtained from expert judgments (Hora et al., 1991; Trauth et al., 1993) that had to be done in order to use the probabilities in the calculations (Hora, 1992). The manipulation was necessitated because the judgments were not obtained in a manner that allowed their direct use in calculations (Hora, 1992).

The major misuse of expert judgments can arise from over-reliance on them. This over-reliance can manifest itself in several ways: (i) use of judgments when obtaining the attendant data was practically possible, (ii) the judgments are dominated by overconfidence (i.e., the judgments do not represent the existing range of uncertainty about the issue addressed), and (iii) the judgments were unduly biased, among others.

Several questions can be asked then to determine the appropriateness of the use of the expert judgments:

- Is the relationship between data, models, and expert judgments clearly discernable?
- Did subsequent calculations clearly identify where expert judgments were used as opposed to other sources of information?
- Were the judgments used to determine the type of data and information that should be collected?
- Were the judgments subjected to peer review before being used?
- What steps were taken to ensure that as new scientific and technical information becomes available, that the judgments are updated, if appropriate, based on this new information?

4.3 THE QUALITY OF PROBABILITY JUDGMENTS

Do experts provide better data than nonexperts? The answer to this question depends on the data that experts are asked to give. The expert can be asked to: (i) make predictions, (ii) provide information on the state of the art in the field, (iii) show how to solve a problem, or (iv) to assess the accuracy of his/her responses. Across a wide variety of tasks experts outperform novices. For example, research has shown that experts are better than nonexperts at solving problems in their field (Johnson, 1980; Johnson and Sathi, 1988; Winkler and Poses, 1993; and Goldberg, 1968, among others).

Experts are also better at providing data on how problems in the field can be solved (Johnson, 1980; Winkler and Poses, 1993). This data encompasses formulating the problem, interpreting it, determining what additional information is needed to solve it, knowing whether and where their data are available, knowing how to solve the problem, and providing the solution.

Experts also differ from nonexperts when assessing their own accuracy or confidence in their answers. The process of assessing or improving the accuracy of the expert judgment is sometimes called calibration. People in general are very poor at assessing the accuracy of their own answers and are usually overconfident. Lichtenstein and Fischhoff (1980) found that those knowledgeable in the field are less prone to overconfidence than those who are not. An individual's calibration improves with knowledge until the individual probability given by the expert is over 0.8. Then, they become less well calibrated because they tend toward underconfidence. It may be that those who are very knowledgeable are more award of the dangers of estimation and thus increasingly tend to underestimate their accuracy.

Despite this, experts do not consistently outperform nonexperts in the area of predictions. In fact, Armstrong (1985) claims that there are no experts in forecasting change. This is the heart of the controversy in using expert elicitation in risk assessment. This section reviews the literature on the quality of probability assessments made by both experts and novices. It is common practice to judge the quality of a forecast retrospectively by comparing the forecast to what actually happens.

4.3.1 Evaluating Elicited Probabilities

The evaluation of a probability estimate involves the correspondence of those probabilities and the actual observations (what actually occurred). If an expert's probability for an event is denoted by r and the corresponding observation by x (where x = 1 if the event occurs and x = 0 if the event does not occur), the correspondence between the probabilities and observations can be investigated by looking at the joint distribution g(r,x) of probabilities and observations. The joint distributions can be factored into conditional and marginal distributions in two ways (Murphy and Winkler, 1987; 1992)

$$g(r,x) = g(x|r)g(r) \tag{4-1}$$

and

$$g(r,x) = g(r|x)g(x) \; . \tag{4-2}$$

These factorizations follow directly from standard probability theory, and the elements of the factorizations provide different information about the probabilities.

- $g(x|r)$ is the conditional distribution of the observations for a given probability value. The mean of this distribution is the relative frequency with which the event occurs when a probability value of r is given. This measure reflects calibration.

- $g(r)$ is the marginal distribution of the probabilities. This indicates how often the expert uses extreme probabilities. This distribution describes the refinement of an expert's probabilities.

- $g(r|x)$ is called a likelihood; for a given x (zero or one), it shows how likely different values of r are. This relates to discrimination.

- $g(x)$ is the base rate, indicating how frequently the event occurs in the data set.

The following sections further describe these indices of the quality of probability estimates.

4.3.1.1 Calibration

Probabilities may be said to be good in the sense that they suitably reflect uncertainty. Calibration refers to the extent to which probabilities assessed for events conform to the relative frequency with which these events occur. For example, the stated probability of precipitation should reflect the relative frequency with which rain occurs. On those days that the weather forecaster announces a 60 percent chance of rain, it should rain about 60 percent of the time. If for every forecast value (10 percent chance of rain, 20 percent chance of rain, etc.) the observed relative frequency of rain corresponds to the probability, then the forecaster is well calibrated. Calibration is associated with the faithfulness of probabilities—how well they predict the relative frequency of events.

One property of well-calibrated probabilities is that a plot of the observed relative frequencies against the probabilities will depict a 45 degree line. A measure of deviation from the 45 degree line, such as the maximum vertical distance from the 45 degree line, can be taken as a measure of calibration.

4.3.1.2 Refinement

The amount of information in a probability or probability distribution is called its refinement. Probabilities close to zero and one have more refinement than those near 0.5. To describe probabilities for continuous quantities, probability density functions that are tightly concentrated have more refinement than those that are spread out. Sometimes the terms precision and sharpness are used to describe refinement.

von Winterfeldt and Edwards (1986) note that good calibration and high refinement are often in conflict. High refinement sometimes may be achieved only at the expense of poor calibration. Conversely, it is possible to be well calibrated but have no refinement. For example, a weather forecaster who gives the same forecast on every day may be well calibrated. If it rains on 20 percent of the days and the prediction each day is for a 20 percent chance of rain, the forecaster is well calibrated but has no refinement. One learns nothing from the forecasts that is not readily available from historical data.

4.3.1.3 Discrimination

Calibration and refinement both focus on the probability values that are provided by the experts. Calibration involves how well these probability values match with empirical reality in the form of relative frequencies, and refinement deals with the closeness of the probabilities to zero or one. In contract, discrimination asks how different probability values discriminate between the occurrence and nonoccurrence of the event, regardless of the actual numerical values themselves. For instance, suppose an expert always assigns probabilities of either 0.60 or 0.40; whenever the probability is 0.60, the event occurs, and whenever the probability if 0.40 the event does not occur. Although the numerical values of 0.60 and 0.40 do not seem very extreme, the expert exhibits perfect discrimination.

4.3.1.4 Qualitative Evaluation of Validity

Characteristics such as calibration, refinement, and discrimination are useful concepts for appraising the quality of probability assessments. Empirical measures of calibration and discrimination are limited, however, to situations where the true outcomes are known. This will often not be the case in the assessment of radioactive waste disposal issues. In this case, the quality of assessed probabilities must be judged by the quality of the experts themselves and the process used to acquire the probabilities (see Sections 4.2.2 and 4.2.4).

The experts providing the probabilities should have possession or access to exceptional knowledge that sets them apart from others. They should be free from motivational bias and conflict of interest, and they should receive training to reduce cognitive biases. When there are multiple scientific viewpoints or approaches to a problem, multiple experts should be used to ensure that diverse viewpoints are included. This will help capture the true range of uncertainty about the question. The quality of probabilities obtained in an expert elicitation also depends in part on the way the questions are asked. Properly structuring and presenting questions helps ensure that the expert is responding to the same question that is being asked. Questions should be asked in a manner free from suggesting or promoting certain answers. In general, a formal well-structured elicitation process is important (e.g., Morgan and Henrion, 1990; Keeney and von Winterfeldt, 1991; DeWispelare et al., 1993).

4.3.2 Experts Providing Probabilities

Cognitive science has documented a clear advantage for experts compared to novices in a wide variety of tasks. However, in studies of decision making under uncertainty experts have not fared as well as expected. Goldberg (1968) found that clinical psychologist and psychiatrists were more accurate in interpreting the Minnesota Multiphasic Personality Inventory to diagnosis psychosis. They were accurate 65 percent of the time, while undergraduate students were accurate 58 percent of the time.

Also of note are comparisons of expert performance to simple statistical models in which expert performance is compared to the performance of a simple linear regression model. In almost every study, the model's predictions are more accurate than the expert judge. For example, Einhorn (1972) compared the ability of physicians to predict the severity of cases of Hodgkin's disease to the ability of a simple linear regression model. The model performed modestly well statistically while the physicians performed no better than chance. Linear regressions have predicted success in graduate school (Dawes, 1979) and security prices (Wright, 1979) more accurately than human experts. The behavioral decision literature presents a dismal view of expert prediction. Experts are often only slightly, if at all, better than novice judges, and they almost universally perform worse than simple statistical models. These effects are very robust across domains and task types.

Winkler et al. (1992) reviewed the behavioral decision making literature relative to probability calibration across a variety of disciplines, including engineering, science and risk analysis, weather forecasting, medicine and psychiatry, intelligence and military applications, and business. For example, Mosleh et al. (1987) present an interesting study of judgmental distributions for component repairs obtained from review of power plant operating histories. The judgmental distributions had been obtained for use in nuclear reactor probability risk assessments. The authors conclude that the experts systematically underestimated the actual variation of the mean maintenance duration from one plant or component to another. However, in eleven of the twelve cases examined, the judgmental mean was within

a factor of four of the empirical mean. Nine of the twelve maintenance time means were larger than the corresponding empirical means, perhaps indicating some conservatism. This result compares favorably to the data reported by Kidd (1970), which shows that engineers considerably underestimated the repair times for generators.

Cooke (1991) conducted a calibration study with experts in the atmospheric sciences. The purpose of the study was to develop subjective probability distributions for dispersion and deposition coefficients. As part of the study, the experts were asked to provide medians and 0.05 and 0.95 fractiles for quantities that were the realizations from experiments involving either dispersion or deposition measurements. Thus, the study involved the unusual and desirable combination of having experts responding to questions requiring their specific expertise and, at the same time, having known answers so that the quality of the probability distributions could be directly judged.

Each of the eleven atmospheric dispersion experts was asked to respond to 36 questions for which the true value was known. An x^2 statistic was computed for each of the experts and, at approximately the 0.01 level of significance, it was possible to reject the hypothesis of perfect calibration for each of the eleven experts. Cooke cautions, however, that the observations are not independent and thus the effective sample size in the x^2 test is actually much smaller than the 36 questions would indicate. This dependence will cause the x^2 statistic to tend to larger values and thus lead to the conclusion of miscalibration more often than warranted. In contrast, within the group of four experts responding to 24 questions about deposition experiments, three of the four experts were shown to be well calibrated.

Physicians have also been expert subjects in studies of probability assessment. Wallsten and Budescu (1983) report on a study by Lusted (1977) involving physicians assigning probabilities to the most important and most likely diagnosis. The diagnoses were then evaluated via an x-ray. Ludke et al. (1977) presented calibration curves for three classes or problems—skull fracture, extremity fractures, and pneumonia. The calibration for extremity fractures is nearly perfect, while probabilities are uniformly overestimated for skull fractures. For pneumonia, low probabilities are overestimated while the converse is true for large probabilities. DeSmet et al. (1979) also present calibration data for physicians that also suggest over estimation of probabilities associated with head traumas.

However in researching a variety of probability attributes, including discrimation, refinement, and calibration, Winkler and Poses (1993) found that expertise was related to better prediction by physicians. They examined the probabilities of patient survival given by physicians in an intensive care unit. The physicians were classified into four levels of experience by title. While all four groups of physicians were reasonably well calibrated, those with the most experience and expertise performed better overall in terms of average scores. The more experienced physicians demonstrated the ability to group together patients with similar survival chances.

Overconfidence when responding to general knowledge questions is pervasive and can be found among professionals as well as students. Cambridge and Shreckengost (1978) found overconfidence among U.S. Central Intelligence Agency analysts, and Hazard and Peterson (1973) found the same bias among students at the Defense Intelligence School. These studies have also revealed a tendency for overconfidence to become severe as the difficulty of the task increases.

Why do experts frequently perform poorly in some predictions? In decision making under uncertainty there is no single correct procedure for making an accurate judgment. Most other tasks offer a correct and validated method of obtaining a correct answer and checking the answer once it is obtained.

Decision making under uncertainty provides no such methodology. At least three hypotheses have been suggested to explain the relatively poor performance of experts (Johnson, 1980):

(i) Experts are fallible. Because of the limitations of information processing capacity, experts may not be able to integrate the predictor variables accurately. In this view, experts are good at selecting and coding information but have a limited ability to combine it.

(ii) Experts use nonlinear rules. In this view, experts may use the same variables as the statistical models, but they combine them differently. Experts report using complex configural rules (i.e., the impact of one variable depends on the value of other variables) in verbal reports of their problem solving behavior. However, the judgments of expert decisions are modeled well using linear models. This does not in itself invalidate the notion of nonlinear decision making. The second finding, that experts are less accurate than the linear models, suggests that, whatever they are doing, it does not improve their performance.

(iii) Experts attend to different variables than do the models. Experts may pay inordinate attention to unusual and infrequent events than the models. In fact these type of events may be too infrequent to be estimated by the regression but may be quite diagnostic.

All of these hypotheses, and many of the previously discussed studies on probability assessment, assume that experts use linear rules to some extent. Thus, the previous results are more a comment on the lack of linearity in expert performance than an indictment against accuracy in the predictions by the experts.

Johnson (1980) reports a pair of studies to evaluate these alternative hypotheses. In the first study, 12 physicians and 2 undergraduate novices reviewed the applications of candidates for internships and residencies. Two of the experts and the two novices reported their thoughts aloud in a verbal protocol while reviewing the applications. Each protocol was segmented into complete thoughts and the thoughts were categorized according to whether they were retrievals, recall, evaluation, scaling, inference, goal, miscomprehension, or comment statements.

The researchers found both qualitative and quantitative differences in the way experts and novices reviewed the applications. The experts required significantly less time to review the applications than the novices. The experts' verbal protocols contained may more goal statements and fewer retrievals (i.e., paraphrased quotation of information in the folders) than the novices. The experts also examined different information than the novices, concentrating on information they felt was diagnostic based on their knowledge of medical education. And they paid particular attention to application-specific information. The author concluded that the experts examined the information in a top-down manner, using their knowledge of medical education to structure the search. This advantage is similar to the advantage of expertise in other domains. Additionally, the processes did not resemble a linear regression model. There seemed to be marked use of one time, case-specific information.

The second study provided additional evidence that the experts are not simply fallible linear regression models. Johnson and Sathi (1988) compared predictions of changes in security prices made by experienced security analysts and by inexperienced MBA students. The securities were described by a set of 22 variables and half of the securities were accompanied by a news item about the company. The names of the securities were not released and the subjects attempted to predict year end prices. The

experts performed more poorly than a simple regression, but improved dramatically when a news item was present. The novices performed most poorly and were not sensitive to the presence of a news item. These results suggest that, while experts make good use of nonlinear cues in their judgments, regression models are better at consistently combining more mundane information that yields good predictions. This distinction is similar to the general tendency of decision makers to underweight base rate information. The experts have a good ability to recognize general patterns and synthesize case-specific information but neglect base rate information. The origins of this neglect is the relatively cumbersome cognitive processing burden of evaluating covariation between variables.

4.3.3 Improving Calibration

The best way to improve calibration is through repeated forecasting with feedback about the accuracy of the predictions. Weather forecasters are the best example of this. Murphy and Winkler (1987) analyzed 17,514 weather forecasts and found nearly perfect calibration. Weather forecasters have several advantages: the task is repetitious, there is an excellent base of information, feedback is provided, and they are rewarded for good performance. High quality forecasts have also been observed among physicians in similar circumstances.

The repetitive nature of weather forecasting and medical diagnosis provide opportunities for practice and feedback. Clearly, in these fields, probability assessors can, and should, perform well. Unfortunately, many of the assessments that must be made in risk analysis and PA are not repetitive, nor can feedback be provided. As Wallsten and Budescu (1983) note, experts involved in assessing probabilities for events with which they are familiar can be very well calibrated. Other studies with experts, however, show that when the task is less familiar, so that there is little opportunity for practice and feedback, experts may succumb to the same limitations as nonexperts.

The majority of those studies in which training has been employed show modest to substantial improvement in calibration. For example, Stael von Holstein (1972) asked 72 participants to provide discrete probability distributions for changes in stock prices. Ten of the participants worked in the stocks and bonds department of a Stockholm Bank, 10 were connected with the stock exchange, 11 were statisticians in academic positions, 13 were business administration faculty, and 28 were students. The target quantities were share prices 2 weeks hence for 12 stocks. The assessments involved assigning probabilities to five classes of price changes. Ten sessions were conducted with feedback on performance provided before each new assessment.

The evaluation of the resulting probability distributions was based on a quadratic scoring rule. Comparisons to log and spherical rules were also made. Stael von Holstein was surprised to find that the bankers performed most poorly. From best to worst, the groups of the participants were statisticians, stock market employees, students, business teachers, and, far behind, bankers. At the beginning of the study, the statisticians gave more spread-out distributions. As the sessions progressed the distributions averaged across all experts became more spread out. Thus, the subjects learned through feedback to spread their distributions, resulting in better calibration.

Training and practice in making assessments should be considered as an important step in improving the quality of assessments. The evidence also indicates that the quality of assessments improves with simpler questions. This suggests that a strategy of decomposition may also help improve the quality of the elicited probabilities. Additionally, the practice, as well as the actual assessment, should utilize base

rate information and feedback disconfirming information or counter examples to improve accuracy (Hoch, 1985; Russo and Schaemaker, 1992).

While all the discussed attributes of assessed probabilities (refinement, discrimination, etc.) are important, the often recognizable and documented flaw in many studies of elicited probabilities is the tendency toward an understatement of uncertainty. Certainly, methods should be considered to counteract this tendency (calibration), as well as other potential problem areas. Asking assessors to consider reasons why an event does not occur as well as why it might occur is an example of an elicitation technique that can be used to improve the quality of probabilities. While in some circumstances experts may provide excellent probabilities, one should not conclude that expertise alone is sufficient to guarantee that probabilities are of high quality. Practice and evaluation seem to be key ingredients in producing high quality probability assessments, and careful design of the overall assessment process is also important.

4.4 SUMMARY

Expert elicitation methods are based on sound scientific principles yet there is considerable difficulty in evaluating the quality of the data obtained. Evaluators thus must rely primarily on an evaluation of the procedures used to collect the data. The beginning of this chapter discussed a number of criteria that can be used to determine whether an expert elicitation is likely to yield unbiased, informed, and useful expert judgments.

Various questions must be asked to determine the quality of the elicitation procedure. The answers to these questions are not necessarily uniform across all conditions. For example, the issue of allowing experts to practice with elicitation techniques in a training session depends on the type of judgment that will be required in the actual elicitation. Experts should practice the procedures for providing probabilistic judgments, but it is not as critical for qualitative judgments. However, most of the questions outlined in this chapter constitute criteria that should be met for a high quality formal expert elicitation.

Increasing the quality of probability judgments requires providing opportunities for training and practice. This will not in itself ensure good calibration, so elicitation techniques should be used that emphasize base rate information and disconfirming information or counterexamples to ensure that the expert estimates uncertainty as accurately as possible.

5 USING EXPERT JUDGMENT

5.1 THE USE OF EXPERT JUDGMENT IN THE HIGH-LEVEL NUCLEAR WASTE REPOSITORY PROGRAM

Expert judgment has been used in the past and will continue to be relied on in the HLW repository program. The NRC is interested in two aspects of this use: (i) the potential use and impact of expert judgments collected by the DOE, and (ii) the applicability of expert judgments to the NRC staff in pre-licensing and licensing activities. This chapter first examines the past use of expert judgments in radioactive waste programs. Next, areas are summarized in which the use of expert judgments by the DOE may have a significant impact and, therefore, should be elicited via a formal process. The discussion that follows is presented following the same template as Bonano et al. (1990). As a matter of fact, significant portions of this chapter are summaries of the information in Bonano et al. (1990); however, the information has been revised or augmented, as necessary, based on relevant studies published since that earlier report.

The discussion then turns to the NRC and the potential utility of expert judgment to the NRC in its role as regulator. A historical look at NRC use of probabilistic risk assessment and other examples are presented and the chapter concludes with a discussion of the likely utilization of expert judgments by the NRC staff.

5.1.1 Past Use of Expert Judgments in Radioactive Waste Management Programs

Expert judgment has been relied on frequently in radioactive waste management programs both within and without the United States. The conduct of a formal expert elicitation has, on occasion, been the venue used to produce the judgments. However, to date, the informal use of expert judgments has been more preponderant. In the following, a few examples of the use of expert judgments, both formal and informal, are highlighted.

(i) **The Nuclear Waste Repository Siting Study.** The DOE used a multi-attribute utility analysis to rank five potential host sites for a HLW repository. Five sites were proposed originally as candidates for the first civilian HLW repository in the United States: Yucca Mountain (Nevada), Richton Dome (Mississippi), Deaf Smith (Texas), Davis Canyon (Utah), and Hanford (Washington). Expert panels were convened for each impact and for each site. For example, expert panels estimated the number of fatalities that could occur by transporting wastes to the various sites. Point estimates and 95 percent confidence intervals were collected for each impact. The impact estimates were weighted and aggregated using a linear, additive utility model, producing a ranking of the five sites (U.S. Department of Energy, 1986).

(ii) **The Earthquakes and Tectonics Expert Judgment Elicitation Project.** The Electric Power Research Institute (EPRI) conducted a formal expert elicitation to evaluate the likelihood of earthquakes in the YM Vicinity over the next 10,000 years. This study assessed the individual probability distributions of seven geologists and seismologists with extensive knowledge of the YM/southern Great Basin area or tectonically analogous areas. The panel met in its entirety in two session: the first for training the group in

elicitation methods and the second for a presentation and discussion of the issues surrounding seismicity in the YM area.

The experts provided individual probability judgments concerning the potential for fault displacement at the site. Although the experts participated in the group sessions to identify and define the issues, they each identified their own significant issues and developed their own models and inputs to the individual elicitations. The individual elicitations collected probability distributions describing each expert's uncertainty about potential sources of fault displacement (Coppersmith et al., 1993).

(iii) **The WIPP Expert Elicitation Project.** Two expert panels were assembled to estimate long-term (10,000-year) radionuclide releases from the WIPP, which is a planned underground repository in southeastern New Mexico. One expert panel estimated the concentrations of specific radionuclides in a repository brine that might be forced up an intruding borehole. The other panel evaluated the retardation of radionuclides in the overlying Culebra Dolomite (Trauth et al., 1992).

(iv) **Dry Run 3.** The UKDoE used formal expert elicitation to provide the basis for assessing the risk of underground disposal of intermediate and low-level nuclear waste (LLW). The experts provided, in group sessions, probability density functions for parameters used in modeling environmental change: porosity, soil characteristics, soil sorption, soil-plant concentration factor, and thermal conductivity of frozen and unfrozen rock. Probability distributions for these parameters were provided by the group of experts as a whole (Dalrymple and Willows, 1992). Expert elicitation was also used to identify and select events, processes, and phenomena to include in computational models (Thorne, 1992; 1993).

(v) **Human Intrusion Study for WIPP.** Sandia National Laboratories (SNL) assembled an expert panel to design permanent markers to deter inadvertent human intrusion into the WIPP. The expert panel identified principles to guide the marking effort, including the following: (i) the message must be truthful and informative, (ii) the message should be presented by multiple means of communication (e.g., pictographs, language), (iii) the marker should be made of materials that have little recycling value, and (iv) there must be an international effort to maintain knowledge of the locations and contents of nuclear waste repositories. The panel also judged the efficacy of the proposed markers. This was evaluated to decrease with time, varying with the mode of intrusion and the technological development of the society (Trauth et al., 1993).

(vi) **SKI Project 90.** The initial PA effort of the Swedish Nuclear Power Inspectorate used an expert panel for various activities, including scenario formulation and classification. Scenarios were developed considering features, events, and processes for areas such as glaciation, faulting, human intrusion, and canister failure modes. Seventeen scenario cases were established (Swedish Nuclear Power Inspectorate, 1991).

(vii) **Initial Total System Performance Assessment for YM.** An initial total system performance assessment for YM was performed by SNL in 1991. Expert elicitation was used to provide a set of hydrologic parameters and associated probability distributions to express the uncertainty in the parameter values. Information on parameters such as

sorption coefficients for various geological media, percolation rates, and likely volcanic dike widths and orientations was elicited as most likely values. Those distributions were included in stochastic simulations to predict groundwater travel values for repository performance (Barnard et al., 1992).

(viii) **The 1993 Total System Performance Assessment for Yucca Mountain.** The 1993 version of the Total System Performance Assessment (TSPA) performed by SNL included an expert elicitation to determine the probability density function for radionuclide sorption and solubility parameters. Two groups of experts, ranging from three to five experts per group, were used to generate the PDFs, which in turn were used as input for the Monte Carlo simulation used in uncertainty analysis (Wilson et al., 1994).

5.2 POTENTIAL USES OF EXPERT JUDGMENTS IN THE HIGH-LEVEL NUCLEAR WASTE PROGRAM

The nature and scope of conducting a technical analysis and PA for a deep radioactive waste repository necessitates that judgments offered by experts be used to estimate the long-term behavior of the system. The spatial and temporal scales over which performance of the disposal system must be examined may preclude the collection of all the data needed to support technical analysis and PA calculations. Therefore, expert judgments are considered necessary in order to supplement the collected data. This is widely recognized both within and without the United States (e.g., Bonano et al., 1990; Dalrymple, 1990; Hora et al., 1991; Thorne, 1992, 1993; Roberds, 1990; Apostolakis, 1990; Apostolakis et al., 1991; Bonano and Apostolakis, 1991; Cambridge Decision Analysts Limited, 1992; Chhibber et al., 1991; 1992; McGuire, 1990; 1992; Trauth et al., 1992, 1993; Rechard et al., 1993; Stephens et al., 1993; DeWispelare et al., 1993; Bonano and Baca; 1994). The issue is not whether the judgments will be used, but rather whether or not they will be obtained and used in an explicit and formal manner that will permit their scrutiny. Bonano et al. (1990) reviewed the use of expert judgments in PA and stated that there is potential for its use in at least five major areas: (i) development and selection of scenarios, (ii) development of conceptual models, (iii) model validation, (iv) quantification of parameter uncertainty, and (v) decisions on data collection.

Technical work in the HLW repository program will use expert elicitation derived judgments to support a number of areas in the program. The use of expert judgment by the DOE is possible in: (i) the preparation of the PA calculations and other technical analyses that will support the license application (LA), and (ii) the analyses and deliberations to interpret the results obtained. The containment requirements previously stipulated in the EPA regulation for HLW disposal dictate that a PA be conducted to determine the ability of the disposal system to isolate the wastes over 10,000 year following closure of the repository. PA calculations aimed at supporting the demonstration of compliance with these requirements must include a prediction of the future states that the disposal system could attain over that period, and expert judgments are pervasive in determining the possible future states of the system.

Expert elicitation can also influence conceptual model development for the HLW disposal site. Conceptual models of fundamental physicochemical principles can rely on expert judgment to decide what information and states need to be included in the model and to validate mathematical models and computer codes that emerge from the conceptual models. Expert elicitation will possibly be used to quantify the uncertainty in the value of input parameters to the models and codes. Expert judgments are also expected to play key roles in the conduct of pre-PA technical analyses, in the design of data

collection strategies, and in strategic engineering decisions aimed at generating the information base necessary for the HLW repository technical analyses. Finally, after assembly and organization of the technical analyses and PA results, expert judgments will likely play a role in the analyses of, and deliberations on, whether or not the results successfully support the demonstration of compliance.

In this section, six generic areas where the use of expert judgments may have a significant impact on the results of technical analyses of PA calculations and on the use of those results to demonstrate (DOE) or determine (NRC) compliance, will be discussed. These areas are:

- Relation of expert judgment to data collection
- Uncertainty about the future states of the repository system
- Conceptual model development
- Model validation
- Parameter uncertainty
- Analysis of results and residual uncertainties

Because, as aforementioned, the use of expert judgments in each of these areas could have significant impact on the outcome of the HLW repository program, where possible, the judgments should be elicited and used employing a formal process. Where available, the use of expert judgments in these areas is supported by briefly discussing past studies involving expert judgments.

5.2.1 Expert Judgments and Data Collection

The use of expert judgments should not be considered a substitute for the collection of data or the conduct of other scientific and technical analyses (Park et al., 1994). The use of expert judgments should, in principle, be restricted to those situations when the collection of so-called "hard" data or other information is not practically possible (see Section 1.6.1.2). It is safe to say that expert judgments are most valuable when data are lacking because the judgments, particularly when offered by multiple experts, can be effectively used to capture what is known and what is not. That is, the expert judgments represent a "snapshot" of the current state of knowledge about a given issue or problem of interest. An important consideration is the determination of when data collection is practically impossible and, to a large extent, such decisions involve a considerable amount of both quantitative and value judgments. Three fundamentally different types of judgments could be involved:

(i) Judgments used (a) decide the type and amount of data to collect and the approach to collect it, and (b) select alternatives to reduce uncertainties if the decision is against the collection of a given type or sufficient amount of data

(ii) Judgments used to generate information on a given issue when the collection of data is not possible

(iii) Judgments used to interpret, synthesize, and extend existing data

5.2.1.1 Use of Expert Judgments Regarding Data Collection and Alternative Courses of Action

Deciding what type and amount of data should be collected and what is the preferred approach to collect it is a decision analysis problem, the solution of which depends in part on expert judgments. Two basic types of judgments are likely to be used: quantitative judgments and value judgments (Bonano et al., 1990; Morgan and Henrion, 1990; Goodwin and Wright, 1991). Quantitative judgments result from the analysis and interpretation of the current state of knowledge. These judgments can be the basis for deciding what type and amount of data should be collected. Quantitative judgments can also influence the decision on how these data ought to be collected; however, value judgments can also play an important role in this latter decision because other factors, such as preserving the isolation integrity of the site, need to be considered as well. Value judgments can be the basis for making the trade-offs that can lead to the selection of the preferred approach to collect the data.

If field data are to be collected at a proposed disposal site, experts may address issues such as the tests to be conducted and the interpretation of collected data. In laboratory experiments, experts may deal with issues such as how representative the experiments are of field conditions, under what conditions the experiments are likely to be valid, how the laboratory data are to be used in conjunction with field data, etc. Finally, if analyses using models are employed to supplement experimental information, experts may address issues such as how the adequacy of the models was established, what key assumptions are contained in the models that cannot be tested, and how to select the values of parameter in the models so that they represent the current state of knowledge about the disposal system, etc.

An important assessment expected to rely on expert judgments is the evaluation of the potential impact on the objective of the evaluation if it is decided not to collect a given type or sufficient amount of data, and in the identification of alternatives that can mitigate such impact. One alternative to data collection, when considerations of such factors as preserving the isolation integrity of a site are important, is the identification and implementation of physical modifications to the engineered components of the system, more commonly known as "engineered alternatives." Experts can become involved in such exercises. Data derived from the laboratory experiments can occasionally substitute for portions of field data, especially under situations where site integrity could be destroyed or the natural process rate is inappropriate as time basis for long-term projections. In these cases, a combination of laboratory experiments and engineered alternatives can mitigate the lack of field data. The formulation of this combination of experiments and engineered alternatives often results from an analysis including both quantitative and value judgments. Technical experts (e.g., individuals in charge of the different experiments as well as individuals cognizant of repository design issues) are providing judgments regarding the possible results from each ongoing or planned experiment, and identifying plausible engineered alternatives and potential combinations of experiments and alternatives. These judgments are being used as input into a computational model that attempts to find optimum combinations of experiments and alternatives. The computational model considers a set of criteria, and judgments are utilized to provide relative importance weights for the criteria.

The analysis of engineered alternatives involves expert judgments in a number of general areas:

- Specification of the objectives of the analysis; for example, performance, cost, schedule, etc.
- Identification and evaluation of alternatives
- Aggregation and analysis of results

- Selection of preferred alternative(s)

5.2.1.2 Use of Information Generated by Experts in Lieu of Data Collection

Modelers often use lumped or representative parameters for convenience. In most cases, there are not physically measurable quantities, so collection of data to directly determine these parameters is not possible. Experts can synthesize available supporting data, and parameter description and function to provide estimates for these convenience parameters.

Experts have been used and will most likely continue to be used to provide information relevant to a given parameter. Goodwin and Wright (1991) as well as Kaplan (1992) distinguish this particular use of experts from the more traditional one, in which the experts are provided with some information and are asked to analyze it to arrive at some opinion about the value of the parameter. The premise behind this use of experts is that the information provided to the experts is relevant to the value of the parameter. The approach advocated by Kaplan (1992) is the use of experts to generate the necessary information relevant to the value of the parameter of interest, which he refers to as the elicitation of expert information and Goodwin and Wright (1991) call the elicitation of expert knowledge. In other words, if one is interested in the probability distribution of a parameter, expert judgments can be used to construct that distribution based on sufficient available information. If the information is not initially available, experts would be used in that case to generate the information. As such, the elicitation of expert information is a precursor to the elicitation of expert judgments in those cases for which the information cannot be obtained by other scientifically or technically acceptable means.

5.2.1.3 Use of Expert Judgment to Interpret, Synthesize and Extend Existing Data

It has been recognized that expert judgments can be used in combination with the available measured data to complement the latter. For example, three-dimensional (3D) geohydrologic data and information are needed for the development and exercise of groundwater flow and radionuclide transport models at a HLW repository site. Ideally, one would like to have as complete a representation of the geohydrologic system as possible; however, this ideal situation will never be realized. First, the cost of obtaining a complete characterization of the geohydrologic system will be prohibitive. Second, and perhaps more important, obtaining large amounts of subsurface information could seriously compromise the integrity of the site, and hence, its ability to isolate the wastes. Realistically, a picture of the geohydrologic system at the repository site is likely to be constructed using a variety of types of data and information. Some of the data or information will be based on actual measurements of key parameters or quantities at the site, and some may come from interpretations of other data or information collected at that site—such as geophysical data—and from interpretations of data and information collected at other sites.

The foundation of the synthesis and interpretation of this latter type of data will be judgments or opinions expressed by subject matter experts after they have had the opportunity to examine and study the different sets of data, have interpreted the data as to their applicability to the geohydrologic system at repository site of interest, and have made interpolation and extrapolation inferences about this site using the relevant data.

The use of expert judgments to interpret, synthesize and extend existing data could serve several purposes. First, it could provide important insights as to whether or not the collection of new/additional data will significantly improve the state of knowledge about the quantity or issue of interest. Second, in

constructing a 3D picture of the geohydrologic system expert judgment could identify specific locations where new/additional data ought to be obtained. Third, expert judgment could be used to assess how the state of knowledge has changed due to the incorporation of newly collected data.

5.2.2 Uncertainty About Future States of the Repository System

The long temporal scales over which the behavior of the repository system and its different components must be evaluated introduces uncertainty about the prediction of the future states that the system can attain. To date, the most common approach to the treatment of the uncertainty about the future states of the system is the postulation and selection of a set of scenarios, with each scenario representing a plausible realization of the future events, phenomena, and conditions (EPCs) that can affect repository performance (Bonano and Baca, 1994). Associated with each scenario is a probability of occurrence representing the likelihood that the scenario will occur during the 10,000-year regulatory period. A scenario is typically defined as a set of EPCs. It should be noted that the use of scenarios, more commonly referred to as the scenario approach, is not the only approach that has been proposed for treating the uncertainty in the future states of the system. Other approaches, such as the environmental simulation approach, have also been developed (Nuclear Energy Agency, 1992).

The difference between the scenario approach and the environmental simulation approach notwithstanding, several observations can be made regarding the use of expert judgments in treating the uncertainty in the future states of the system (Bonano and Baca, 1994):

(i) Expert judgments play a major role in treating uncertainty about the future states of the repository system

(ii) The important roles of scenarios in any technical analyses and in PA, coupled with the complexity of selecting scenarios and the assignment of a probability of occurrence, necessitate a formal and rigorous approach for the elicitation of expert judgments.

There is considerable evidence reported in the literature that supports these observations (e.g., Andersson et al., 1989; Apostolakis et al., 1991; Barr et al., 1993; Cranwell et al., 1990; DeWispelare et al., 1993; Hora et al., 1991; Hunter and Mann, 1989; McGuire, 1990; 1992; Nuclear Energy Agency, 1987, 1989, 1990, 1992, 1993; Stephens and Goodwin, 1990; Thorne, 1992, 1993; Trauth et al., 1993). When initially proposed, those who that advocated the use of the environmental simulation approach argued that it was superior to the scenario approach because it had a dependence on expert judgments (Thompson, 1988). However, recent work by Thorne (1992; 1993) demonstrates that the environmental simulation approach relies on subjective judgments to (i) identify and select EPCs to be included in calculational models, and (ii) assign probabilities of occurrence to the EPCs.

Because the majority of the radioactive waste management programs in the United States and in member countries of the Organization for Economic Cooperation and Development use the scenario approach (Nuclear Energy Agency, 1992; Bonano and Baca, 1994), the discussion that follows below is in the context of the scenario approach. Where applicable, this section discusses the work of Thorne (1992; 1993), which was performed to support the application of the environmental simulation approach. Hopefully, it will become evident that very strong parallels exist between the two approaches insofar as the use of expert judgments in the treatment of the uncertainty in the future states is concerned. Furthermore, regardless of the specific customization of the scenario approach used, the basic framework

is the same among most, if not all users with the exception of the United Kingdom Nirex Ltd. team (Nuclear Energy Agency, 1992; Bonano and Baca, 1994). This framework is the one developed by Cranwell et al. (1990), which consists of the following basic steps:

(i) Identification and Classification of EPCs

(ii) Screening of EPCs

(iii) Construction of Scenarios

(iv) Screening of Scenarios

(v) Estimation of the Probability of Occurrence

To facilitate the presentation of the material, the discussion that follows is structured along these five steps.

5.2.2.1 Identification and Classification of Future Events, Phenomena, and Conditions

The initial listing of plausible EPCs is a creative task that may depend almost exclusively on expert judgment. The main issue here is increasing the likelihood of completeness; that is, attempting to increase the likelihood that no potentially significant EPCs have been left out. Formal elicitation of expert judgments is one means of achieving this goal. Several recent studies demonstrate the use of a group of experts to develop a list of EPCs for use in the development of scenarios and for development models used in the environmental simulation approach. For example, Stephens and Goodwin (1990) describe the use of experts to identify factors, their equivalent term for EPCs, applicable to the disposal of radioactive wastes generated from the operation of CANDU reactors in the plutonic rocks of the Precambrian Shield. To help generate a comprehensive list of factors, the experts classified them based on: (i) type of factor, (ii) component of the disposal system affected by a factor, (iii) origin of the factor, (iv) mode of action of the factor, (v) sub-component of system affected by the factor, and (vi) pathways by which radionuclides can reach humans if the factor were to occur. A total of 270 factors were identified.

The Swedish Nuclear Waste Management Company (SKB) and the Swedish Nuclear Power Inspectorate (SKI) conducted a joint project to agree on the methodology for selecting scenarios associated with the disposal of HLW and spent fuel in crystalline rocks (Andersson et al., 1989). The initial phase of the project was a workshop in which SKB and SKI experts as well as several international scenario experts were assembled to develop a comprehensive list of factors, events, and processes (FEPs), their equivalent of EPCs. The group used a classification approach to facilitate the identification of a comprehensive list of FEPs. Each FEP could be classified in terms of: (i) likelihood of occurrence, (ii) component of disposal system affected, (iii) time of occurrence, and/or (iv) initial cause of the FEP. The group generated an initial list containing 156 FEPs. More recently, Thorne (1992; 1993) used a group of experts to define factors and phenomena to be included in the PA calculations that constituted the U.K. Department of Environment/Her Majesty's Inspectorate of Pollution Dry Run 3. The latter exercise differed from the Canadian and Swedish studies in two ways: (i) Thorne seems to have used a more formal approach in the elicitation of the expert judgments, and (ii) Thorne requested each expert to generate an individual list of factors and phenomena, which he and his staff later combined into a comprehensive list.

No recent study in the United States seems to have been conducted aimed at the development of comprehensive lists of EPCs; rather, United States studies have focused on a specific type of EPC or scenario. Hora et al. (1991) and Trauth et al. (1993) addressed human intrusion into the WIPP, Barr et al. (1993) basaltic volcanism for YM, and DeWispelare et al. (1993) climate change for YM. To the extent applicable, the examples of recent expert judgments studies associated with uncertainty in future states just mentioned will be used in other parts of this section to illustrate specific points being made.

5.2.3 Screening of Future Events, Phenomena, and Conditions

The initial list of EPCs, while not necessarily generic, tends to err on the side of conservatism. Thus, it is customary to shorten it by eliminating those EPCs of little significance to the PA. Experts have been used in identifying and applying screening criteria.

In the Canadian study, the same group of experts that developed the initial list of factors was involved in their screening. Factors that were deemed not to be sufficiently important were eliminated from further consideration. A factor was judged not to be sufficiently important if (i) the decision was unanimous among all members of the group, and (ii) reasons for judging the factor not to be important were documented (Stephens and Goodwin, 1990). Some of the reasons used to eliminate factors were: (i) it did not apply to waste generated from the operation of CANDU reactors, (ii) it did not apply to the disposal in plutonic rocks of the Precambrian Shield, or (iii) it did not apply to packaging in titanium alloy canisters. Criteria contained in the pertinent Canadian regulations were also applied (e.g., factor judged to be significant only after the end of the regulatory period). Of the original 270 factors, 125 were retained and used to construct scenarios.

In the Swedish study, the screening of FEPs was conducted by a subset of the group of experts that developed the original list. A FEP was judged to be insignificant and removed from further consideration for one or more of the criteria that the FEP: (i) had a low probability of occurrence ($< 10^{-8}$), (ii) was of negligible consequence, (iii) was physically unreasonable, (iv) was the responsibility of future generations, and (v) was outside scope of study (Andersson et al., 1989). FEPs that survived the screening were classified into one of four classes that determine how they will be used in the construction of scenarios. The original SKB/SKI study did not proceed further than this step.

In the United Kingdom study by Thorne (1992; 1993), the comprehensive list of factors and phenomena was distributed to the members of the original group of experts. Each expert was instructed to review the list and individually identify those factors in the list that:

(i) Would be significant to PA or that would require further analysis to be able to determine its significance

(ii) Should be excluded from the PA calculations due to negligible impact

(iii) Could be considered by grouping it with others with similar characteristics and potential impact

After results from all the experts were compiled in a single list, the experts were assembled to discuss, modify if necessary, and ratify the final list of factors and phenomena.

5.2.4 Construction of Future Scenarios

Scenarios are constructed from all possible combinations of EPCs remaining after screening. Typically, an event tree is used to generate all possible combinations of EPCs. The procedure is straightforward if the initial list of EPCs is fairly complete and potentially significant ones have not been screened out. While this can, in principle, be done mechanically, expert judgments may be needed to prune first-cut event trees and to check their consistency and completeness.

Several variations and alternatives to event trees have been proposed or used to construct scenarios from the surviving EPCs. These variations tend to rely more on expert judgments than the mechanical event-tree approach. For example, in the Canadian study (Stephens and Goodwin, 1990), the group of experts was used to construct a central scenario that consisted of those factors expected to be always important, occur frequently, or proceed to a significant degree over the length of the regulatory period. Factors that were deemed to be (i) important only occasionally or under special circumstances, (ii) incompatible with the presence of a presumably more important factor in the central scenario, or (iii) amenable to simplification of the analysis were excluded from the central scenario. These factors were called residual factors. Of the 125 factors surviving screening, 117 were included in the central scenario, and the remaining eight residual factors were used to construct 255 alternative scenarios. Andersson et al. (1989) proposed a similar approach. If the Swedish study would have proceeded beyond the screening step, it would have categorize each FEP into one of four classes. Class I would have contained all those FEPs believed to be always active; Class I FEPs were included in the Process System scenario, which is equivalent to the Canadian central scenario. FEPs within each class are combined to construct scenarios.

Recently, Barr and Dunn (1993) proposed a variant to the conventional event-tree approach they call the generalized event tree. In their approach, Barr and Dunn would organize FEPs based on:

(i) Definition of initiating FEP

(ii) Impact of initiating FEP on groundwater system

(iii) Impact of initiating FEP on waste

(iv) Release of waste from engineered-barrier system due to occurrence of FEP

(v) Transport to accessible environment due to occurrence of one or more FEPs

Barr and Dunn would then construct scenarios by starting with an initiating FEP and connecting in a logical and physically plausible manner combinations or sequences of FEPs that would lead to a release of radionuclides to the accessible environment. A scenario would then be a single connecting path through the tree together with sketches of a detailed conceptualization of the system. In principle, there could be multiple scenarios for a single path if multiple conceptualizations are possible. This approach relies considerably on expert judgments in at least two ways: (i) to determine what constitutes a combination or sequence of FEPS and forms a path through the tree, and (ii) to determine the possible conceptualizations associated with each path. The first use of expert judgments in essence constrains the possible number of paths and could be viewed as being similar to scenario screening, which is discussed next, whereas the second is similar to conceptual model development to be discussed in Section 4.2.3.

5.2.5 Screening of Scenarios

One of the drawbacks of the event-tree approach is the construction of a large number of scenarios (Nuclear Energy Agency, 1992); for example, in the Canadian study a total of 255 scenarios were obtained from only 8 residual factors (Stephens and Goodwin, 1990). It is, therefore, customary to screen scenarios so that their number can be reduced to a manageable set. Expert judgments will play an important role in this preliminary screening by developing criteria for screening and applying them. In the aforementioned Canadian study, experts reduced the 255 scenarios to 4 using a variety of criteria.

5.2.6 Assessment of the Probability of Occurrence

Probabilities need to be assigned to scenarios to quantify the likelihoods of scenarios used in PA calculations. Expert judgments play a significant role in assessing probabilities of occurrence for scenarios. Ideally, some historical data exist for a given site on climatic changes, seismic activity, volcanic activity, human intrusion, etc. Expert judgments are used to interpret the data and arrive at probabilities. Realistically, data are likely to be scarce, and expert judgments become the main basis for assessing the probability.

The probability of occurrence of a scenario is a combination of the probabilities of the EPCs included in the scenario. Expert judgments play a major role, not only in determining the probability of the individual EPCs, but also in the way these probabilities are combined to arrive at the probability of the scenario. For example, experts are likely to be used to decide whether a scenario's EPCs occur in a sequence and, if this is so, to determine the sequence, or to determine whether the occurrence of the other EPCs in the scenario are conditional on the occurrence of the initiating one.

The assessments of the probability of EPCs and of scenarios has received much attention in recent years in the United States. Some of the studies conducted are: Apostolakis et al. (1991), Bonano and Apostolakis (1991), DeWispelare et al. (1993), Hora et al. (1991), McGuire (1990; 1992) and Trauth et al. (1993). Three of these studies, DeWispelare et al. (1993), Hora et al. (1991), and Trauth et al. (1993) employed a formal approach to elicit the judgments. DeWispelare et al. (1993) experimented with a variety of combination methods for the probability of climate change in the YM region elicited from the individual experts. They found that mechanical combination methods were easier to implement than behavioral ones; the latter relied on the experts being willing to capitulate some of their positions, which proved to be an unlikely proposition. McGuire et al. (1990; 1992) seem to have elicited judgments from experts in the construction of event trees to arrive at the probability of scenarios in a semi-formal manner; using a multidisciplinary group of experts. The other two studies, Apostolakis et al. (1991) and Bonano and Apostolakis (1991), were more concerned with the investigation of probability assessment techniques, and therefore, did not use a formal process to elicit the expert judgments needed to exercise the different techniques.

Nonetheless, several interesting observations emerged from these studies. For example, Bonano and Apostolakis (1991) found that experts exhibited a strong anchoring bias because they were highly influenced by the information contained in a single report provided to them. As a result, the focus of the study was shifted because the experts anchored on the predictions about the recurrence of a characteristic earthquake in the YM region given in the report. In essence, the experts were treating the report as the expert and they played the role of analysts calibrating the expert. Another interesting finding of that study, which is described in detail in Chapter 5 of Apostolakis et al. (1991), was that the investigators

were able to indirectly calibrate the experts by examining their responses to specific questions that would provide an indication of internal consistency for each expert. Responses from one of the experts were found to be inconsistent, and therefore, that expert's judgments were not included in the model used to calculate the CDF of the recurrence interval of the characteristic earthquake.

5.2.7 Conceptual Model Development

Assumptions and simplifications that can be incorporated into a conceptual model for mathematical simulation of system behavior are made about the behavior of the repository system. Conceptual modeling of a HLW disposal site is based on a combination of the application of fundamental physicochemical principles and data interpretation. Data interpretation and conceptual model development rely almost exclusively on expert judgments.

5.2.8 Data Interpretation

Model development is based on limited, site-specific information about the system geometry, past and active processes, and potentially disrupting future processes and events. Little or no data may be available to determine all of these factors at the proposed repository site. Therefore, experts could select and interpret data from similar sites and relate them to the repository site. Interpretations of scant geologic data may be used to define the system geometry. Experts could infer such things as the geologic continuity between boreholes, the extent and thickness of units, and the extent and character of geologic discontinuities such as faults. The geometry defined by these experts may be based not only on interpolation and extrapolation of the site-specific data, but on data from similar geologic environments. Experts could select and interpret data to decide which processes to consider in assessing the performance of the repository system (e.g., Thorne, 1992; 1993). Not only do the experts have to decide the current dominant processes, but they may decide on future processes that could adversely affect the repository system. This later assessment may require the experts to identify and interpret data from similar systems (i.e., analogs to the future states of the repository).

5.2.9 Conceptual Model Development

Because conceptual models provide the underpinning for the development of the mathematical models and computer codes that will be used in the quantitative estimates of performance measures, these judgments could have a considerable impact on the results of computational analyses. It is quite likely that these judgments, when made by the DOE to support calculations in the LA, will receive considerable scrutiny by the NRC (Park et al., 1994).

The most common approach to conceptual modeling begins with a rough sketch of the model and continues to refine that sketch based on whatever experimental data and other information are available until an adequate first-cut model is produced. Typically, this is done by using one expert's judgment and interpretations of experimental data and other information. To make conceptual modeling more comprehensive and to encourage considerations of alternative models as well as scrutability of the experts reasoning, Bonano and Cranwell (1988) suggested an approach for formalizing the use of expert judgments from multiple experts well versed on groundwater flow and transport. That approach would force the experts (i) to articulate all assumptions, and (ii) to look for interpretations that challenge their conventional wisdom and are consistent with available data. The second point could lead to multiple conceptual models. Finally, the approach could include procedures for allowing the experts to identify

bounding analyses and experimental investigations aimed at distinguishing between alternate conceptualizations and eventually reducing their number.

5.2.9.1 Model Validation

After conceptual models for the disposal system have been assembled, appropriate mathematical models and computer codes must be developed to simulate the behavior of the system over the spatial and temporal scales prescribed by the regulatory requirements. Experts will likely be an integral part of limited-scope activities to build confidence in models and codes. For example, international groups have been formed to select problems of common interest to the radioactive waste-management community. These are simulated by interested parties, and the results are compared. The groups attempt to find discrepancies and their causes among the results from different teams. One important result is that the group may, implicitly or explicitly, agree that, given the current state-of-the-art, existing models and codes are as good as they can be.

Validation, in the strictest sense, means comparing the predictions of the models to experimental or actual results. Because the models' predictive capabilities cannot be fully tested, true validation can never be achieved. The alternative is to build confidence in the models and codes through a synthesis of experiments and calculations and make determinations regarding their fitness for their purposes. Experiments are likely to include laboratory and controlled-field investigations as well as natural analogs. Calculations could consist of bounding analyses and preliminary PA calculations.

In any case, experts will be used to: (i) design experiments and calculations, (ii) establish the validity and limitations of these experiments and calculations, (iii) define appropriate measures to ascertain the predictive capabilities of the models and codes, (iv) ascertain the validity of important couplings in the models that cannot be tested, (v) interpret the results of model runs against existing and new data, and (vi) judge the ability of the models to extrapolate to large temporal and spatial scales.

The experts should make value judgments (tradeoffs) regarding what aspects of models need to be tested. In addition, they should develop criteria for establishing the validity of given models. Because the ultimate validation test at an HLW disposal site cannot be performed and because of the complexity of the models, perhaps one of the major model-validation issues addressed by experts could be the decomposition of models into meaningful pieces. It has been recognized that there are likely to be couplings that cannot be tested; therefore, experts are expected to offer judgments as to the validity of untested couplings.

5.2.9.2 Parameter Uncertainty

To assess the uncertainty in predicting the behavior of HLW repository systems, it is necessary to quantify the uncertainty about the input parameters required to exercise the models and computer codes used. The uncertainty about parameters can be expressed in a variety of ways. One way is to estimate a mean value and the variance about the mean. Another way is to determine the range of possible values and to assess a PDF covering that range. The latter method is conventionally used because it provides a complete description of uncertainty and facilitates the generation of multiple samples of the values of input parameters for carrying out Monte Carlo simulations (e.g., Helton 1993; Stephens et al., 1993; Thompson and Sagar, 1993). Although it is widely accepted that a gamut of expert judgments will be used in calculations, the relevant literature has predominantly focused on expert judgments to encode

probabilities and develop PDFs. For these reasons, the examples below focus on the assessment of PDFs for values of input parameters.

In principle, assessment of the possible range of values and PDFs of input parameters should rely on a very large sample of field data. However, such a large sample is not likely to be collected at a candidate repository site. Expert judgments may be required to determine what samples to take and how to interpret the results and to assess a probability distribution on the basis of the sample. Techniques for the elicitation and use of expert judgment can also be applied to quantify expert knowledge on a given parameter to develop a PDF for that parameter. In this latter case, experts provide information relevant to the value of the parameter (see Section 6.2.1.2). Several recent studies have been performed to develop PDFs of uncertain parameters for use in PA and other technical analyses; two of these studies are Trauth et al. (1992) and Stephens et al. (1993). Tierney (1990) presents a five-step procedure to construct PDFs of uncertain parameters based on expert judgments. The salient feature of Tierney's approach is the use of the maximum entropy formalism (mef) to construct the PDFs when there are no data and the only information is that provided by individuals with substantive knowledge on the parameter.

5.2.9.3 Interpretation of Results and Analysis of Residual Uncertainties

PA and other technical analyses have a bottom line. That is, they produce results (Bernero, 1990). In principle, these results will be used to determine the adequacy of the repository system in its entirety, or of one or more of its components. However, the suite of uncertainties that affect the attendant calculations preclude an unequivocal statement about the repository system and/or its components based solely on the results of the analyses.

In conjunction with the technical results, several fundamental questions need to be addressed with regard to the results themselves, the computational approaches that led to the results, and the manner in which the results are presented and interpreted. Some of these questions are:

- How can the completeness of the analyses be demonstrated and ascertained?
- Can the approach used to present the results affect the determination of the adequacy of the repository or of its components? That is, is there more than one approach to present the results and, if so, what is the possible impact of selecting one approach over the other?
- What other factors or issues need to be considered jointly with the computational results to demonstrate (DOE) or determine (NRC) compliance with the pertinent regulatory requirements?

The containment requirements previously specified by EPA prescribe that PA results need to be assembled into a complementary CCDF that indicates the probability of exceeding various levels of releases of radionuclides to the accessible environment for 10,000 years following closure of the repository. Thus, the CCDF is a fundamental indicator of whether compliance with the release limits established by EPA and implemented in 10 CFR 60.112 has been demonstrated. There are different ways of generating the CCDF that will be compared to the limits established by EPA, such as (i) the generation of an expected value CCDF by normalizing scenario probabilities to combine conditional CCDFs (Bonano and Wahi, 1990), or (ii) development of a family CCDFs from which an average CCDF can be obtained (Helton, 1993). Decision makers are likely to decide which approach will be used to construct the CCDF

that is compared to the regulatory limits. Because no particular approach is necessarily superior to another, the decision will involve expert judgments.

A key issue that will most likely be addressed using expert judgments is the resolution of residual uncertainties (Fehringer and Coplan, 1992). It is recognized that, due to practical limitations, after very reasonable effort has been made to reduce the uncertainties affecting repository performance, some uncertainties will still remain. Some examples of these residual uncertainties are (i) unidentified future states, (ii) uncertainty in computational models that are less than fully validated, and (iii) multiple conceptual models. How specific residual uncertainties will be examined and resolved is still unknown. However, regardless of the specific approach used to resolve these uncertainties, there is no doubt that expert judgments will be pervasive (Fehringer and Coplan, 1992).

5.3 USE OF EXPERT JUDGMENT IN DOE's HIGH-LEVEL NUCLEAR WASTE REPOSITORY PROGRAM

The DOE has used, and is expected to continue to use, expert judgments extensively in its HLW repository program. Earlier in Section 5.1, several examples of past use of expert judgments pertaining to the DOE program, in general, and the proposed YM repository, in particular were discussed. In this section, other examples of DOE's use of expert judgments are discussed.

The DOE has used expert judgments, stemming primarily from informal elicitations, in several aspects of its program; for example, scenario development and selection, development and evaluation of conceptual models, model abstraction, and parameter uncertainty, among others. Ross (1987) developed the first set of scenarios for the proposed repository in the Topopah Spring unit at YM. He started with lists of EPCs developed by Hunter et al. (1982; 1983) and by the International Atomic Energy Agency (1981). From those lists he selected 58 EPCs that he believed might affect the performance of one or more of the repository components. These EPCs were used to construct 84 sequences, later grouped into 17 general categories. All the sequences in a given category were expected to have similar consequences. The 84 sequences were expanded to 99 in the preparation of the Yucca Mountain Site Characterization Plan (U.S. Department of Energy, 1988) mostly due to the consideration of alternative conceptual models of the site not examined by Ross (1987).

More recently, Barr and Dunn (1993) proposed a scenario development approach using generalized event trees. This approach was applied by Barr et al. (1993) to the development of scenarios describing future basaltic igneous activity at YM. The generalized event tree approach requires judgments from technical experts in the different scientific disciplines required to identify relevant features, events, and processes at the site as well as experts in modeling these. A scenario is any combination of (i) a sequence of features, events, and processes and (ii) one possible conceptual model.

In the TSPAs conducted in 1991 (Barnard et al., 1992) and in 1993 (Wilson et al., 1994; Andrews et al., 1994) for the proposed HLW repository at YM, expert judgments were used in a variety of ways. These included the selection of scenarios analyzed, the selection of alternative conceptual models for groundwater flow in the unsaturated zone, the abstraction of detailed models into models suitable for Monte Carlo simulation, and the development of PDFs for uncertain parameters. Unfortunately, Wilson et al. (1994) recognized the development of PDFs for radionuclide sorption and solubility as the only use of expert judgments. In that case, two separate groups of experts, ranging from three to five experts per group, provided the PDFs. The fact that the development of PDFs for sorption and solubility was the

only explicitly recognized use of expert judgments in the SNL version of the 1993 TSPA is cause for concern. In reality, as just mentioned, the use of expert judgments (implicitly or explicitly) was far more extensive that the development of the PDFs for two geochemical parameters. However, if the use of expert judgments is not explicitly indicated it may be difficult to identify and ascertain how the judgments were used.

There is no reason to believe that the use of expert judgments by DOE in PA calculations for the proposed repository at YM will decrease. On the contrary, it is quite possible that it may somewhat increase. As the site characterization program progresses, it could be found that some specific types of data, initially believed to be collectible, may prove to be impossible to obtain. In those cases, the DOE may opt to elicit judgments to close newly found gaps in data collection.

5.4 THE USE OF EXPERT JUDGMENT BY THE NUCLEAR REGULATORY COMMISSION

The NRC is evaluating the applicability of expert judgment for use by its staff in the HLW program. In the past, expert judgment has been used primarily by the staff in the reactor program. Lately, however, formal expert elicitation has been used by the NRC to acquire judgments to support the HLW Iterative Performance Assessment activities. Because the NRC has historically used probabilistic risk assessment (PRA) as an integral part of reactor evaluations, the discussion on the use of expert judgment at NRC will commence there.

5.4.1 Risk Management and Risk Assessment in the Public Sector

The most common early use of the expert judgments in the regulatory process by the NRC was in the area of risk management. Risk management is generally defined as the process of weighing policy alternatives and selecting the most appropriate regulatory action by integrating results of risk assessment with engineering data and social, economic, and political concerns to reach a decision (Nuclear Regulatory Commission, 1984). Risk management is conducted under various legislative mandates by a number of regulatory agencies including, the EPA, the Consumer Product Safety Commission, Food and Drug Administration (within the U.S. Department of Health and Human Services), the Occupational Safety and Health Administration (within the U.S. Department of Labor), and the NRC. The specific definition of risk management differs between agencies depending on the application. The EPA and NRC have jointly defined these concepts in a recent memorandum of understanding on risk (Nuclear Regulatory Commission, 1993a). They define risk assessment as the methods, assumptions, and other considerations involved in quantifying or estimating the health risks associated with a particular activity (Nuclear Regulatory Commission, 1993b). Risk management is the selection of risk objectives and the associated means to achieve these objectives. Most applications of risk management require value judgments to be made by the decision maker regarding issues of acceptability of risk and reasonableness of cost controls. Judgments are also made by technical experts in conducting risk assessments; however, these usually pertain to modeling assumptions about the behavior of complex systems, parameter estimation, desired level of precision, and probabilities of uncertain events.

Important components of risk management include the following (Zimmerman, 1990):

- Formulation of alternative courses of action to avert, avoid, or mitigate risk
- Evaluation of alternatives considering social, health, and economic criteria
- Determination of risk acceptability as a function of perceptions, attitudes, beliefs, and behavior
- Selection and implementation of an approach based on technical evaluation and acceptability
- Monitoring and re-adjustment as necessary

Risk management and risk assessment are synergistic processes. Risk assessment provides necessary inputs to the risk management process. In turn, risk management judgments affect the conduct of risk assessments. For example, identifying the overall objectives of a risk management strategy requires judgments to identify those aspects of a system which are perceived to pose the greatest threat to public health and safety. Such decisions serve to bound the scope of detailed risk assessments, which will provide results to the risk manager to support further judgments on effective risk minimization options. Specific expert judgments used in the risk assessment will affect the assessment conclusions which can then impact risk management decisions. The relationship between risk assessment and risk management is therefore complementary and difficult to clearly delineate. The degree of separation between these two concepts has been a source of debate in the regulatory arena (Silbergeld, 1991). A 1984 report by the National Research Council (National Research Council, 1984) considered this issue and recommended that risk management and risk assessment activities be clearly defined and separated institutionally in regulatory agencies. The extent to which this suggestion was adopted in government is difficult to ascertain. Recently there has been a call for risk-based decision making regarding the management of environmental risks. Over the last decade, for example, EPA has adopted a decision-making process that closely couples risk assessment, risk comparison, and risk management (Bersell, 1994). EPA's use of risk-based decision making in the regulatory context was recently endorsed by the Carnegie Commission on Science, Technology and Government. Most recently, the National Research Council (1994) issued a report on a study it conducted that reached the conclusion that the incorporation of risk assessment in risk management and decision making is an important step in the environmental remediation of DOE results.

A number of decision-making frameworks are available for use in risk management under a variety of circumstances and authority. Lave (1981) has summarized some commonly used techniques:

- Market Based Regulation allows risks to be managed by market forces and consumer choice. Therefore, the consumer is relied upon to weigh the risks and benefits of a particular product. An example is the current approach to regulating risks of smoking tobacco.
- Zero Risk Management Strategy involves rejecting a product or activity if any risk is detected.
- Technology-based Standards rely on engineering judgment regarding state-of-the-art performance of techniques employed to reduce risks with no consideration of costs and

benefits. The HLW Repository licensing standards in 10 CFR Part 60 (Nuclear Regulatory Commission, 1991) are an example of performance standards.

- Direct Risk-Risk Comparison involves quantification and comparison of the risks and benefits of a single health effect related to an activity or exposure (e.g., cancer risk).

- Indirect Risk-Risk Comparison involves quantification and comparison of the risks and benefits of a range of health effects pertaining to an activity or exposure under consideration (i.e., drug licensing).

- Risk Benefit Analysis is a generalized comparison of risks and benefits not restricted to health effects (e.g., National Environmental Protection Act Environmental Impact Statement process), while benefit-cost analysis is similar but with greater quantification and formalization than risk benefit analysis.

- Cost Effectiveness Analysis is aimed at maximizing an objective under the assumption of fixed costs.

- Regulatory Budget only focuses on a subset of costs defined in a budget.

The type of approach selected for a particular situation will vary according a number of factors including overall management structure, statutory authority to regulate, and the level of public concern with the regulated activity.

5.4.2 Probabilistic Risk Assessment

PRA, the most common method for the conduct of risk assessment, involves the application of a variety of analytical techniques to calculate risks based on the products of hazardous event probabilities and resulting consequences. Such weighting of consequence estimates by event probabilities allows a regulator to interpret and communicate risks more effectively than is possible using consequence analysis alone because the likelihood of occurrence of initiating event(s) has been factored into the risk result. PRA therefore requires a thorough understanding of such probabilities and consequences and the relevant parameters influencing each. These relevant parameters are considered in the development of scenarios representing the state of the system to be modeled. Since most systems are dynamic, there will be a range of possible scenarios that must be considered, each with a different probability of occurrence and suite of possible consequences. The intent of a risk assessment is to consider all possible combinations of potential events and consequences to estimate the full range of possible risks. The regulator aims to identify those factors that contribute unacceptable risks that can be controlled by reasonable human intervention. The judgment as to what constitutes an unacceptable risk is a risk management decision.

The overall approach of a given PRA can range from simple to complex depending on the needs of the regulator and the complexity of the systems being analyzed. Although data on certain processes and events may be limited or nonexistent, organizational and legal mandates may require timely decision making from the regulator. Under such circumstances expert elicitation can provide the necessary parameter estimates for PRA models based on the current state of knowledge in the relevant scientific disciplines. This allows regulators to obtain the necessary information to characterize risk to the best of their ability and make informed regulatory decisions in a timely manner. While the nature of a given risk

assessment will vary depending on the intended purpose and scope, there are some components of analysis common to all PRAs. These include hazard identification, exposure assessment, and risk estimation.

Before initiating a risk assessment the regulator must determine whether a potentially hazardous situation exists. When potentially hazardous materials are involved the basis for hazard determinations must be considered. Human health hazards from toxic substances are usually determined by experts analyzing the weight of the evidence of toxicological and epidemiologic research. Such determinations form the basis for risk calculations in risk assessments. Experts conducting a risk assessment must be familiar with the nature of such determinations and the associated levels of uncertainty to reduce bias in the results. A key component of a hazard determination is the dose response relationship (i.e., effect observed per unit of exposure). Data limitations and differing scientific opinions can lead to alternative interpretations of dose response information. An example is provided by the differing dose response hypotheses for biological effects of ionizing radiation (National Research Council, 1990). When differing opinions exist without sufficient experimental data to make a firm conclusion, expert judgments may be used to determine the most credible alternative.

Exposure assessment refers to analyses focussing on determining human exposure following postulated releases of toxic materials to the environment. Exposure can only occur if the toxic material comes into direct contact with sensitive systems of the human body. This requires a detailed understanding of the circumstances involved in releasing the agent into the environment, environmental systems which transport the material to human subjects, and the biological systems of humans which facilitate (and mitigate) entry or exposure to sensitive systems of body. Depending on the focus of the assessment, each of these areas could contain gaps in scientific knowledge which may necessitate the use of informed subjective judgments.

Release mechanisms can involve very simple situations such as a continuous stack emission, or vast complexity as in a postulated nuclear reactor malfunction or industrial plant failure. If the release mechanism is not a well-known regular occurrence, then causal events leading to release and their associated probabilities will need to be considered and factored into the risk assessment.

In this case, each triggering event and its associated probability will need to be determined. Furthermore, relationships between triggering events are also important considerations. For complex systems, this process quickly leads to an unmanageable array of possible scenarios and modeling information needs. As a result, it becomes necessary to group similar scenarios into classes so they can be kept to a manageable number. Expert judgments can be used at this stage to assist in the identification of possible scenarios and scenario classes. Similarly, areas where data are scarce or nonexistent and impractical to obtain may also be suitable for expert judgments.

The environmental fate of a hazardous material can be as simple as direct dilution into a nearby drinking water source or a complex web of environmental system interactions including sources, sinks, bioconcentrators, food chains, and transformations. Such characteristics of the transport path must be considered in order to estimate the amount of material which reaches the target population. Simplifying assumptions are often used to avoid excessive complexity of analyses (in light of scope of study) and resolve data/knowledge limitations. Models applicable to the specific type of flow and transport situation are used to determine the amount of material which reaches the target population and the duration of time necessary for transport. Subjective judgments are used in selection of models, application of models, and selection of appropriate data.

While some analyses use environmental concentration as an endpoint, further modeling is necessary if the desired result is dose or health effects. Biological fate modeling includes consideration of intake, transport, allocation, transformation, storage, and excretion of toxic material. Depending on the form of material, exposures can occur by dermal contact or penetration, inhalation, and ingestion. The end result is an amount of exposure to a target organ or system which is known to exhibit a toxic response to the material. Expert opinion is used in cases where the specific biological mechanism for toxicity has not been clearly established. In addition, assumptions regarding the characteristics of the target population which affect exposure (e.g., demographics, exposure/risk related behaviors) require some level of subjective judgment.

Risk estimation considers the dose-response information from the hazard identification and estimated doses (and their probabilities) from the exposure assessment to calculate risks for each scenario weighted by scenario probabilities. Risks represent the product of event probabilities and their consequences. Analysis and interpretation of calculated risks involves consideration of uncertainties. Uncertainty analysis provides a framework for combining and describing the uncertainties associated with elements of the analysis in order to determine the overall uncertainty in the results. Uncertainties come from a variety of sources but consideration of both model and parameter uncertainties has been emphasized (Chhibber et al., 1992; Nuclear Regulatory Commission, 1990). Data uncertainties are those due to lack of complete data and model uncertainties result from difference between the model results and actual system behavior. A number of methods for uncertainty analysis exist. One common method is to use parameter ranges instead of single numbers for important variable inputs. A statistical sampling technique [e.g., Latin Hypercube Sampling (LHS)] then randomly samples input data values from all such distributions in order to generate distributions of results as output. These distributions can be combined to show the overall variability in results due to uncertainties. This allows for a better understanding of the precision of results. Sensitivity analysis results can be used to determine which variables are contributing the most uncertainty to the results. To use these methods, it is necessary to have distributions for the important modeling parameters. When such distributions are not in the available literature, and it is not practical to obtain the information through research, expert judgment is used.

5.4.3 Risk Management in the Regulatory Arena

As stated earlier, risk management deals with the policy and societal issues regarding risk, and the cost, practical feasibility, and timing of risk mitigating actions. Thus, a key attribute of a good risk management program is consistency in the manner decision makers respond to risk. This entails ensuring that all potentially hazardous events and their associated risks are considered on the same basis so that there is a framework for conducting appropriate and meaningful risk comparisons.

In the regulatory arena, risk management is typically used to determine adequate margins of safety. Margins of safety are closely tied to societal perceptions and tolerance of risk, which commonly range from totally unacceptable risks (i.e., risks that cannot be accepted under any set of circumstances) to risk that can be safely neglected. In the jargon of regulatory agencies, risks that fall in the latter category are called de minimis risks and these need not be regulated. Risk assessment plays a key role in determining de minimis risks. In between the two end points in the risk scale lies a gray area in which the tolerance and mitigation of risks depend on two factors: (i) the benefits reaped from taking the risks and (ii) the cost of risk reduction. Managing risks that fall in this gray area is the essence of risk management. For regulatory agencies, risk management translates into decisions regarding how these risks should be regulated.

While risk assessments provide key input to risk management and many quantitative judgments used in the former will have a direct impact on the latter, the most critical judgments in risk management are value judgments. No one could have said this better than Professor Omenn, Dean of the University of Washington's School of Public Health, who stated (Bersell, 1994): "Even in a world with perfect scientific knowledge, it would be still be true that many of the judgments necessary to compare risks transcend science...science cannot tell us which risk is worse...."

5.4.4 Past Use of Expert Judgments by the Nuclear Regulatory Commission

While discussion of expert judgment case histories are limited in NRC documentation, the cautious application of expert judgment methods has been discussed in early PRA guidance (American Nuclear Society/Institute of Electrical and Electronic Engineers, 1983) and routine use of expert judgment in risk analyses and other technical studies has been acknowledged (Nuclear Regulatory Commission, 1990). Examples of historical use of expert judgment in risk analysis at NRC are discussed in NUREG/CR-4962 (Mosleh et al., 1987). These include NRCs initial PRA effort in the Reactor Safety Study (Nuclear Regulatory Commission, 1975), efforts of the Steam Explosion Review Group considering containment failure from in-vessel steam explosions in NUREG-1116 (Nuclear Regulatory Commission, 1985), analysis of severe accident risks in NUREG/CR-4551 (Benjamin et al., 1986), the NRC human reliability handbook comprising NUREG/CR-1278 (Swain and Guttmann, 1983), the assessment of human error probabilities (Embrey et al., 1984) and NRC sponsored precursor studies regarding severe core damage accidents (Minarick and Kukielka., 1982; Minarick et al., 1986).

NRC use of PRA to assess nuclear power plant accidents began in 1975 with the development of the Reactor Safety Study (Nuclear Regulatory Commission, 1975). Following the Three Mile Island (TMI) incident in 1979, the NRC intensified its research efforts on accident processes to provide a better foundation for risk assessments (American Nuclear Society/Institute of Electrical and Electronic Engineers, 1983). A report to the Commissioners on the TMI incident recommended greater use of PRA by NRC staff to compliment the traditional nonprobabilistic methods of assessing accident risks (Rogovin, 1980). The NRC gradually introduced PRA into its regulatory process, conducting PRAs for additional plant designs (Carlson et al., 1981). In 1988, the NRC requested information on severe accident vulnerabilities from each licensed power plant in the United States, offering a choice of methods including PRA. Virtually all plants responded with their intentions to use PRA.

The formal use of expert elicitation in NRC PRAs was introduced with the development of NUREG-1150 (Nuclear Regulatory Commission, 1990). NUREG-1150 is a risk assessment of five nuclear power plants in the United States. This example represents a greater degree of sophistication and formalization in NRC application of expert judgment in PRA. During the preparation of NUREG-1150, the NRC also undertook an effort to examine the risk management implications of the results of NUREG-1150 by commissioning the NUREG/CR-5263 study. NUREG/CR-5263 identifies two primary risk management goals for nuclear power plants. The first is to minimize the public health risk from nuclear power plants and the second is to provide the capability for operators and decision makers to effectively respond to and thereby reduce the probability and consequences of severe accidents (Nuclear Regulatory Commission, 1989). Five phases of the risk management approach were identified as:

- Prevention of accident initiators (reliability management)
- Prevention of core damage (accident management)

- Implementation of an effective emergency response (emergency response management)

- Prevention of vessel breach and mitigation of radionuclide releases from the reactor coolant system (accident management)

- Retention of fission products in the containment and other surrounding buildings (accident management)

The procedures for the elicitation were based upon those identified in NUREG/CR-4551 Volume 1 (Gorham et al., 1993) and Volume 2 (Harper et al., 1990). The issues selected for elicitation were those important to risk, with large uncertainties, and little or no widely acceptable data.

The primary results of NUREG-1150 focus on core damage frequency; accident progression and containment performance; severe accident source terms; offsite consequences; and public risk. NUREG-1150 discusses NRC intended uses of the results (Nuclear Regulatory Commission, 1990):

Other examples of recent use of formal expert elicitation to produce expert judgments for use by the NRC staff are as follows:

(i) The Nuclear Reactor Safety Study (NUREG 1150). The goal of this study was to estimate the uncertainties and consequences of sever core damage at selected nuclear power plants. Seven panels of experts were assembled to study accident frequency, reactor coolant pump seal performance, in-vessel accident progression, containment loading, molten core-containment interaction, containment structural response, and source term uncertainties. Probabilistic estimates were obtained individually from a total of 40 experts involved in the 7 panels.

(ii) Clarification of the Substantially Complete Containment Performance Requirement. This project involved a prioritization exercise with members of the NRC staff to provide guidance and interpret the substantially complete containment performance requirement for an engineered barrier system in the proposed HLW repository. The performance requirement for the engineered barrier system consists of two parts: (a) a containment requirement for HLW packages, and (b) a radionuclide release limit from the engineered barrier system. Although the requirement in the rule for limited release from the engineered barrier system in the post-containment period is clearly stated in numerical terms, the associated requirement for substantially complete containment during the containment period is qualitative and subject to interpretation. This study used a formal expert elicitation to define a quantitative criteria for substantially complete containment (Tschoepe and Abramson, 1992).

(iii) Expert Elicitation of Future Climate in the Yucca Mountain Vicinity. The study examined potential future climate scenarios in the vicinity of the proposed HLW repository at YM and their associated probabilities. A panel of five climatologists was assembled to develop probability distributions of climatic parameters, such as temperature and precipitation, that will serve as input to PA models. Each expert was interviewed individually to elicit his judgment of the technical issues and to provide the technical basis for his assessment (DeWispelare et al., 1993).

The probabilistic models of possible accident sequences, containment events, and offsite consequences of severe accidents will be used in:

- Development of guidance for individual plant examinations of internally and externally initiated accidents
- Accident management strategies
- Analysis of the need and means for improving containment performance under severe accident conditions
- Characterization of the importance of plant operational features and areas requiring improvement
- Analysis of alternative safety goal implementation strategies
- The effect of emergency preparedness on consequences

The data on major contributors to risk and uncertainty information will be used in:

- Prioritization of research
- Prioritization of generic issues
- Integration and prioritization of issues for inspections

This example illustrates how the application of expert elicitation in PRA enabled the NRC to conduct a detailed safety analysis of five nuclear facilities and generate a useful information base for fulfilling its regulatory responsibilities.

5.4.5 Policy Analysis and the Regulatory Process

In the current regulatory framework, agencies with mandates to protect health and safety [e.g., NRC, EPA, Occupational Safety and Health Administration (OSHA), Federal Aviation Administration (FAA)] must regulate well-known risks in addition to potential risks. When the NRC considers a LA for a nuclear power plant the regulations require licensing actions to be guided by the principle in 10 CFR Part 50.40(a) that...the health and safety of the public must not be endangered (Nuclear Regulatory Commission, 1993b). Continually evolving, improving, and changing scientific knowledge and public perceptions of safety hazards compels the regulator to review existing policy and consider the need for revision or development of new standards or guidance pertaining to licensing procedures, regulatory standards, and enforcement actions. In addition, unforeseen events (e.g., TMI Incident) can create a sudden need for reassessment of existing policy. Often decisions must be made prior to development of the necessary scientific knowledge needed to fully address issues of risk. When a licensing proceeding begins or an important safety concern must be addressed there will likely not be the time nor resources to study all the issues to resolution. Inevitably, the regulatory decision maker must assess current information, however incomplete, and take action to comply with legal mandates, revise outdated standards, or address current public safety concerns. The risk management process, supported by a

diverse body of information including PRA and the judicious application of expert elicitation, provides a means by which the regulator can utilize existing information and expertise to make informed decisions in a timely manner.

The NRC acknowledges the use of expert judgment in PRA to supplement and interpret available data for parameters where little or no data exists (Nuclear Regulatory Commission, 1990) and in developing probability distributions for PRAs when information is inadequate (Nuclear Regulatory Commission, 1984). The regulatory applications of PRA at NRC are outlined in NUREG-1050. These include prioritization of resources, generic regulatory decision making, and plant-specific applications (Nuclear Regulatory Commission, 1984). Outside of reactor regulation there are applications for HLW repository licensing and the safety of devices that use radioactive materials (Nuclear Regulatory Commission, 1993b). While PRA is an important contributor of information, it must still be viewed in the context of risk management as only one of many factors considered in regulatory decisions. The judgment of technical experts, the degree of conservatism required, and the estimated accuracy of results are acknowledged as factors which enhance the credibility of regulatory decisions (Nuclear Regulatory Commission, 1984).

The general process for decision making is described in NUREG-1050. In the first step, the necessary information required for the decision and the analytical methods required to obtain the information are determined. Methods can be qualitative or quantitative, deterministic or probabilistic, and can use operating experience and involve value-impact analyses. Once the appropriate methods are selected, analyses are conducted and then assessed for technical credibility, using peer review when appropriate. All the information is then evaluated to gain insights into safety significance and determine and assess alternative resolutions. The process concludes with recommendations for regulatory action considering inherent uncertainties in the analyses. By allowing a documentable, broad analysis of issues, alternatives, and uncertainties, PRA is well suited to the needs of the regulatory decision process.

For prioritization of resources, the integrated nature of PRA and its reliance on realistic information provide valuable information on ways which critical safety functions of nuclear power plants can fail (Nuclear Regulatory Commission, 1984). In this light, limited resources can be more efficiently allocated to upgrade systems that are shown to be significant contributors to risk. Results are tempered with an engineering evaluation for the reasonableness of priority assignments. The documentation of a disciplined analysis rather than sole reliance on subjective judgments for prioritization is a primary benefit. Also, assumptions and uncertainties are clearly identified and documented. This process has been applied in a number of situations, including the prioritization of generic safety issues and action items following the TMI incident (Nuclear Regulatory Commission, 1984). Other applications include allocation of resources for NRC inspection and enforcement programs (Nuclear Regulatory Commission, 1984). Generic and plant specific identification of important contributors to system failures and risk help both the regulator and the licensee ensure safety.

NUREG-1050 emphasizes the application of PRA to provide technical support for generic decision making as the greatest potential utility of PRA (Nuclear Regulatory Commission, 1984). PRAs can support decisions to strengthen, relax, or support existing regulations. This is beneficial due to the need to take an integrated approach to analyzing complex systems. Generic lessons-learned from plant-specific PRAs have provided the impetus for several regulatory actions. PRAs have also identified areas where regulatory effort has been overemphasized on issues found to have negligible safety significance. In such cases, there is a need for very careful consideration of uncertainties and underlying assumptions in analyses.

5.4.6 Expert Judgment in the Nuclear Regulatory Commission High-Level Nuclear Waste Program

The discussion in the previous section indicates that the NRC has recognized the need to use, and has effectively used, expert judgments in PRA applied to reactor safety. There is little reason to believe that the use of expert judgment in the HLW program will be dramatically different than in the reactor licensing program. As a matter of fact, the NRC has already used expert judgment in the HLW program. Both value and quantitative judgments were likely used in the development of the regulatory requirements in 10 CFR Part 60 (Nuclear Regulatory Commission, 1991). Expert judgments are currently being used and will continue to be used in the examination and development of possible future revisions to this regulation. For example, a formal elicitation process was used to obtain judgments about the definition of the quantitative criteria constituting the substantially complete containment performance requirements in 10 CFR 60.112 (Tschoepe and Abramson, 1992).

The development and exercise of the NRC's technical assessment capability includes the use of expert judgments. For instance, NRC staff do not necessarily adhere, as they should not, to DOE's interpretations of the available data to determine: (i) current and future EPCs that need to be considered in PA and other technical analyses, (ii) the extent to which these EPCs need to be modeled, and (iii) PDFs for uncertain parameters; etc. Rather, fundamental to the development of this technical assessment capability is the need to rely on the NRC's staff's own expertise and insights to consider what ifs, and as necessary, to question and challenge the DOE's approach and interpretation.

The conduct of the Iterative Performance Assessment (IPA) exercises will require the NRC to elicit judgments because all of the data needed to carry out the analyses are not readily available. IPA Phases 1 and 2 relied on considerable amounts of expert judgments related to: (i) scenario selection and assessment of the probability of occurrence, (ii) selection of conceptual models, (iii) uncertain parameter PDFs, and (iv) interpretation of results from uncertainty analysis and sensitivity analysis. To date, most of these judgments have been provided using less than formal procedures. The exception is the expert elicitation exercise for the identification of climate change modes and their probability of occurrence for the YM region in which a strictly forma process was employed (DeWispelare et al., 1993) As IPA progresses and the level of sophistication in the model increases, there will most likely be a need to combine the available measured data with subjective judgments to construct the data base necessary to exercise the models.

The NRC relied on technical expert judgments in the development of the Compliance Determination Strategies (CDSs) for the License Application Review Plan (LARP) within the context of the Systematic Regulatory Analysis. These judgments mainly involved the identification of key technical uncertainties (KTUs) and the level of review that each would need. Currently, KTUs are being examined to (i) decide which KTUs the NRC should address and which should be DOE's responsibility, (ii) identify KTUs that warrant combination into "global" KTUs, and (iii) rank and prioritize the KTUs that NRC will address. This review and the accompanying decisions will be based largely on expert judgments.

The next step in the development of the LARP is the identification, selection, and testing of Compliance Determination Methods (CDMs). The nature of the CDMs will depend on the nature of the KTU each CDM addresses. For each CDM, the NRC will identify data, models, techniques, etc. needed to test and apply the CDM. The assessment method (models, techniques, etc.) comprising each CDM will be tested against an appropriate data set the staff deems representative of the conditions at YM. Expert

judgments will be involved in: (i) identifying and determining the appropriate data set, and (ii) establishing the adequacy of the CDM.

The review of the DOE's LA will require the NRC to use expert judgments to address a number of issues or questions. The NRC will need to determine the completeness of the PA and other technical analyses with regards to whether the estimates of the attendant performance measures are credible and defensible. In particular, decisions will need to be made regarding the handling of residual uncertainties and their impact on the defensibility of the quantitative performance measures. The NRC will also need to determine whether or not the regulatory requirements have been met. This determination will not depend exclusively on the scientific body of evidence. Qualitative issues that also affect the decision will need to be considered, and these issues by necessity will require expert judgments.

The use of expert judgments and their elicitation via a formal process, when warranted, will not only allow the NRC to carry out the aforementioned activities, but it will also put the NRC staff in a position to effectively review the DOE's use of expert judgments. The underlying message throughout this report is that the elicitation of expert judgments, even when employing a formal process, is not always amenable to a cookbook recipe approach. In reality, the elicitation of expert judgments, particularly for complex and potentially controversial problems such as a HLW repository, is an art. Therefore, the NRC's own expert judgment elicitation studies will provide the staff with the necessary insights to evaluate DOE's use of expert judgments both in pre-LA activities and in activities supporting the demonstration of compliance case in the LA. It is known that, to date, the DOE's TSPAs have included both implicit and explicit, formal and informal expert judgments. Insights gained from the NRC's elicitation studies will be critical to identifying implicit and informal expert judgments that could have a significant impact, and therefore, should have been obtained using a formal elicitation process.

5.5 SUMMARY

Expert elicitation is a tool that provides a type of data, expert judgments, that can be used to support PA, other technical analyses, and licensing activities. These data are valid to the extent that they are collected using formal procedures that can be subjected to scientific scrutiny. Two basic types can be obtained: quantitative and value judgments.

Expert judgments have routinely been used in risk analyses supporting regulatory action. Regulatory agencies are engaged in the management of risk by integrating the results of risk assessments with engineering data and social, economic, and political concerns to arrive at decisions. Expert judgment will continue to be a source of data for this process.

Expert judgments collected by formal elicitation procedures have a variety of potential uses in the HLW program. In conditions in which collecting data will result in site destruction or in which models are unreliable on the time scales examined, expert judgments may be the only source of the data needed to evaluate the alternatives and make decisions. The DOE has used and will continue to use expert judgement for scenario development and selection, development and evaluation of conceptual models, model abstraction, and parameter estimation.

6 REFERENCES

American Nuclear Society/Institute of Electrical and Electronic Engineers. 1983. *PRA Procedures Guide: A Guide to the Performance of Probabilistic Risk Assessments for Nuclear Power Plants.* NUREG/CR-2300. Washington DC: Nuclear Regulatory Commission.

Andersson, J., T. Carlsson, T. Eng, F. Kautsky, E. Soderman, and S. Wingefords. 1989. *The Joint SKI/SKB Scenario Development Project.* SKI Technical Report 89:14. Stockholm, Sweden: Swedish Nuclear Power Inspectorate.

Andrews, R.W., T.F. Dale, and J.A. McNeish. 1994. *Total System Performance Assessment-1993: An Evaluation of the Potential Yucca Mountain Repository.* B00000000-01717-2200-00099, Revision 01. Las Vegas, NV: INTERA, Inc.

Apostolakis, G.E. 1990. The concept of probability in safety assessment of technological systems. *Science* 250: 1,359–1,364.

Apostolakis, G.E., and J.S. Wu. 1990. The interpretation of probability, de Finetti's representation theorem, and their implications to the use of expert opinions in safety assessment. *Conference on Reliability and Decision Making.* Siena, Italy.

Apostolakis, G.E., R. Bras, L. Price, J. Valdes, K. Wahi, and E. Webb. 1991. *Techniques for Determining Probabilities of Events and Processes Affecting the Performance of Geologic Repositories.* NUREG/CR-3964, Vol. 2. Washington, DC: Nuclear Regulatory Commission.

Armstrong, J.S. 1985. *Long-Range Forecasting from Crystal Ball to Computer, Second Edition.* New York, NY: John Wiley & Sons.

Armstrong, J.S., W.B. Denniston, and M.M. Gordon. 1975. The use of the decomposition principle in making judgements. *Organizational Behavior and Human Performance* 14: 257–263.

Arrow, K.J. 1951. *Social Choice and Individual Values.* New York, NY: John Wiley & Sons.

Aumann, R.J. 1976. Agreeing to disagree. *Annals of Statistics* 4: 1,236–1,239.

Bacharach, M. 1979. Normal Bayesian dialogues. *Journal of the American Statistical Association* 74: 837–846.

Barnard, R.W., M.L. Wilson, H.A. Dockery, J.H. Gauthier, P.G. Kaplan, R.R. Easton, S.W. Binghmam, and T.H. Robey. 1992. *TSPA 1991: An Initial Total-System Performance Assessment for Yucca Mountain.* SAND91-2795. Albuquerque, NM: Sandia National Laboratories.

Barr, G.E., and E. Dunn. 1993. A working definition of scenario and a method of scenario construction. *Proceedings of the Fourth Annual International Conference on High-Level Radioactive Waste Management.* La Grange Park, IL: American Nuclear Society: 1,093–1,098.

Barr, G.E., E. Dunn, H. Dockery, R. Barnard, G. Valentine, and B. Crowe. 1993. *Scenarios Constructed for Basaltic Igneous Activity at Yucca Mountain and Vicinity*. SAND91-1653. Albuquerque, NM: Sandia National Laboratories.

Bates, J.M., and C.W.J. Granger. 1969. The combination of forecasts. *Operational Research Quarterly* 20: 451–468.

Benjamin, A.S., D. Kunsman, S. Lewis, W. Murfin, and D. Williams. 1986. *Evaluation of Severe Accident Risks and the Potential for Risk Reduction: Surry Power Station, Unit 1.* NUREG/CR-4551. Washington, DC: Nuclear Regulatory Commission.

Bernero, R. 1990. Are you sure? Performance assessment beyond proof. *Safety Assessment of Radioactive Waste Repositories.* Paris, France: Organization for Economic Cooperation and Development: 53–58.

Berry, D.C., and D.E. Broadbent. 1984. On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology* 36A: 209–231.

Bersell, S.D. 1994. *Risk and Environmental Decision-Making.* Document No. 8300. New York, NY: American Institute of Chemical Engineers.

Blattenberger, G., and R. Fowles. 1994. Road closure: Operational variable weights for combining data and expert opinion. *International Journal of Forecasting.* In preparation.

Bonano, E.J., and G.E. Apostolakis. 1991. Theoretical foundations and practical issues for using expert judgements in uncertainty analysis of high-level radioactive waste disposal. *Radioactive Waste Management and the Nuclear Fuel Cycle* 16: 137–159.

Bonano, E.J., and R.G. Baca. 1994. *Review of Scenario Selection Approaches for Performance Assessment of High-Level Waste Repositories and Related Issues.* CNWRA 94-002. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

Bonano, E.J., and R.M. Cranwell. 1988. Treatment of uncertainties in the performance assessment of geologic high-level radioactive waste repositories. *Mathematical Geology* 20: 543-565.

Bonano, E.J., and J.A. Thies. 1994. Bridging the gap between performance assessment and data collection for waste disposal systems: A proposed approach using genetic algorithms. *Technology and Programs for Radioactive Waste Management and Environmental Restoration.* R.G. Post and M.E. Wacks, eds. Tucson, AZ: WM Symposia, Inc.: 1,925–1,931.

Bonano, E.J., and K.K. Wahi. 1990. *Use of Performance Assessment in Assessing Compliance with the Containment Requirements in 10 CFR Part 191.* NUREG/CR-5521. Washington, DC: Nuclear Regulatory Commission.

Bonano, E.J., S.C. Hora, R.L. Keeney, and D. von Winterfeldt. 1990. *Elicitation and Use of Expert Judgement in Performance Assessment for High-Level Radioactive Waste Repositories.* NUREG/CR-5411. Washington, DC: Nuclear Regulatory Commission.

Boyd, D., and S.G. Regulinski. 1979. *Characterizing Uncertainty in Technology Cost and Performance.* Project No. 1114. Menlo Park, CA: Decision Focus Incorporated.

Brockhoff, K. 1975. The performance of forecasting groups in computer dialogue and face-to-face discussion. *The Delphi Method: Techniques and Applications.* H. Linstone and M. Turoff, eds. Reading, MA: Addison-Wesley.

Bunn, D., and G. Wright. 1991. Interaction of judgmental and statistical forecasting methods: Issues and analysis. *Management Science* 37: 501–518.

Burns, M., and J. Pearl. 1981. Causal and diagnostic inferences: A comparison of validity. *Organizational Behavior and Human Performance* 28: 379–394.

Cambridge Decision Analysts Limited. 1992. *Procedures for the Elicitation of Expert Judgements in the Probabilistic Risk Analysis of Radioactive Waste Disposal: An Overview.* Cambridge, England: Cambridge Decision Analysts Limited.

Cambridge, R.M., and R.C. Shreckengost. 1978. *Are You Sure? The Subjective Probability Assessment Test.* Unpublished manuscript. Langley, VA: Central Intelligence Agency.

Carlson, d.d., W.R. Cramond, J.W. Hickman, S.V. Asselin, and P. Cybulskis. 1981. *Reactor Safety Study Methodology Applications Program.* NUREG CR-1659, Volume 1. Washington, DC: Nuclear Regulatory Commission.

Chhibber, S., and G. Apostolakis. 1993. Some approximations useful to the use of dependent information sources. *Reliability Engineering and System Safety* 42: 67–86.

Chhibber, S., G. Apostolakis, and D. Okrent. 1991. A probabilistic framework for the analysis of model uncertainty. *Transactions Institution of Chemical Engineers* 69B: 67–75.

Chhibber, S., G. Apostolakis, and D. Okrent. 1992. A taxonomy of issues related to the use of expert judgements in probabilistic safety studies. *Reliability Engineering and System Safety* 38: 27–45.

Clemen, R.T. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5: 559–583.

Clemen, R.T., and A.H. Murphy. 1986. Objective and subjective precipitation probability forecasts: Statistical analysis of some interrelationships. *Weather and Forecasting* 1: 56–65.

Clemen, R.T., and R.L. Winkler. 1985. Limits for the precision and value of information from dependent sources. *Operations Research* 33: 427–442.

Clemen, R.T., and R.L. Winkler. 1986. Combining economic forecasts. *Journal of Business and Economic Statistics* 4: 39–46.

Clemen, R.T., and R.L. Winkler. 1987. Calibrating and combining precipitation probability forecasts. R. Viertl, ed. *Probability and Bayesian Statistics.* New York, NY: Plenum: 97–110.

Clemen, R.T., and R.L. Winkler. 1990. Unanimity and compromise among probability forecasters. *Management Science* 36: 767–779.

Clemen, R.T., and R.L. Winkler. 1993. Aggregating point estimates: A flexible modeling approach. *Management Science* 39: 501–515.

Clemen, R.T., S.K. Jones, and R.L. Winkler. 1994. Aggregating forecasts: An empirical evaluation of some Bayesian methods. *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*. D. Berry, K. Chaloner, and J. Geweke, eds. In Press.

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

Cooke, R., S. French, and J. van Steen. 1990. *The Use of Expert Judgment in Risk Assessment*. The Netherlands: Delft University of Technology.

Coppersmith, K.J., R.C. Perman, and R.R. Youngs. 1993. *Earthquakes and Tectonics Expert Judgment Elicitation Project*. EPRI TR-102000. Palo Alto, CA: Electric Power Research Institute.

Cranwell, R.M., R.W. Guzowski, J.E. Campbell, and N.R. Ortiz. 1990. *Risk Methodology for Geologic Disposal of Radioactive Waste: Scenario-Selection Procedure*. NUREG/CR-1667. Washington, DC: Nuclear Regulatory Commission.

Dalkey, N.C. 1967. *Delphi*. Report P 3704. Santa Monica, CA: The Rand Corporation.

Dalkey, N.C. 1969. *The Delphi Method: An Experimental Study of Group Opinions*. Report No. RM-5888-PR. Santa Monica, CA: The Rand Corporation.

Dalkey, N.C., and B. Brown. 1971. *Comparison of Group Judgement Techniques with Short-Range Predictions and Almanac Questions*. Report No. R-678-ARPA. Santa Monica, CA: The Rand Corporation.

Dalrymple, G.J. 1990. The use of expert opinion in specifying input distributions for use in probabilistic risk analysis of radioactive waste disposal. *Safety Assessment of Radioactive Waste Repositories*. Paris, France: Organization for Economic Cooperation and Development: 683–696.

Dalrymple, G.J., and M. Willows. 1992. *Dry Run 3—A Trial Assessment of Underground Disposal of Radioactive Wastes Based on Probabilistic Risk Analysis Volume 4: Elicitation of Subjective Data*. DoE/HMIP/RR/92-043. London, England: Her Majesty's Inspectorate of Pollution, Department of the Environment.

Davis, P.A. 1994. *Presentation on SNL Approach to Optimize Experiments and Engineered Alternatives for the WIPP*. Presentation at U.S. Department of Energy, Carlsbad, NM: Carlsbad Area Office, March 7, 1994.

Dawes, R. 1979. The robust beauty of improper linear models in decision making. *American Psychologist* 34: 571–582.

Dawes, R.M., D. Faust, and P.A. Meehl. 1989. Clinical versus actuarial judgement. *Science* 243: 1,668–1,673.

de Finetti, B. 1937. La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales De L'Institut Henri Poincaré* 7.

de Finetti, B. 1974. *Theory of Probability.* Volumes 1 and 2. New York, NY: John Wiley & Sons.

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

DeSmet, A.A., D.C. Fryback, and J.R. Thornbury. 1979. A second look at the utility of radiographic skull examination for trauma. *American Journal of Radiology* 132: 95–99.

DeWispelare, A.R., L.T. Herren, R.T. Clemen, and M.P. Miklas. 1993. *Expert Elicitation of Future Climate in the Yucca Mountain Region.* CNWRA 93-016. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

Einhorn, H.J. 1972. Expert measurement and mechanical combination. *Organization Behavior and Human Performance* 7: 86–106.

Einhorn, H.J., and R. Hogarth. 1975. Unit weighting schemes for decision making. *Organizational Behavior and Human Performance* 13: 171–192.

Embrey, D.E., P.C. Humphreys, E.A. Rosa, B. Kirwan, and K. Rea. 1984. *SLIM-MAUD: An Aproach to Assessing Human Error Probabilities Using Structured Expert Judgment.* NUREG/CR-3518. Washington, DC: Nuclear Regulatory Commission.

EPA (1986), *Air Quality Criteria for Ozone and Other Photochemical Oxidants,"* EPA/600/8-84/020a to 020e, U.S. Environmental Protection Agency, Research Triangle Park, NC.

Ericsson, K.A., and H.A. Simon. 1984. *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press.

Eslinger, P.W., and B. Sagar. 1989. Use of Bayesian analysis for incorporating subjective information. *Geostatistical, Sensitivity, and Uncertainty Methods for Ground-Water Flow and Radionuclide Transport Modeling.* Columbus, OH: Battelle Press: 613–627.

Evans, J.B.T. 1989. *Bias in Human Reasoning: Causes and Consequences.* London, United Kingdom: Lawrence Erlbaum Associates.

Evans, J.B.T., and A.E. Dusoir. 1975. *Sample Size and Subjective Probability Judgements: A Test of Kahneman and Tversky's Hypothesis.* Paper read to the Experimental Psychology Society, Oxford University.

Evans, J.St. B.T., and A.E. Dusoir. 1977. Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psychologica* 41: 129–137.

Fehringer, D., and S. Coplan. 1992. Uncertainty in regulatory decision-making. *Proceedings of the Third Annual International Conference on High-Level Radioactive Waste Management.* LaGrange Park, IL: American Nuclear Society: 106-109.

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

Fischhoff, B. 1982. Debiasing. *Judgment Under Uncertainty: Heuristics and Biases.* D. Kahneman, P. Slovic, and A. Tversky, eds. New York, NY: Cambridge University Press: 422–444.

Flores, B. E., and E.M. White. 1989. Subjective versus objective combining of forecasts: An experiment. *Journal of Forecasting* 8: 331–341.

Fong, G.T., D.H. Krantz, and R.E. Nisbett. 1986. The effects of statistical training on thinking about everyday problems. *Cognitive Psychology* 18: 253–292.

French, S. 1985. Group consensus probability distributions: A critical survey. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds. *Bayesian Statistics* 2: 183–197.

Genest, C. 1984. Pooling operators with the marginalization property. *Canadian Journal of Statistics* 12: 153–163.

Genest, C., and K.J. McConway. 1990. Allocating the weights in the linear opinion pool. *Journal of Forecasting* 9: 53–73.

Genest, C., and M.J. Schervish. 1985. Modeling expert judgements for Bayesian updating. *Annals of Statistics* 13: 1198–1212.

Genest, C., and J.V. Zidek. 1986. Combining probability distributions: A critique and annotated bibliography. *Statistical Science* 1: 114–148.

Goldberg, L.R. 1968. Simple or simple processes? Some research on clinical judgments. *American Psychologist* 23: 483–496.

Goodman, B. 1972. Action selection and likelihood estimation by individuals and groups. *Organizational Behavior and Human Performance* 7: 121–141.

Goodwin, P., and G. Wright. 1991. *Decision Analysis for Management Judgement.* Chichester, England: John Wiley & Sons.

Gorham, E.D., R.J. Breeding, J.C. Helton, T.D. Brown, W.B. Murfin, F.T. Harper, and S.C. Hora. 1993. *Evaluation of Severe Accident Risks: Methodology for the Accident Progression, Source Term, Consequence, Risk Integration, and Uncertainty Analyses.* NUREG/CR-4551, Volume 1. Washington, DC: Nuclear Regulatory Commission.

Harper, F.T., T.D. Brown, R.J. Breeding, J.J. Gregory, A.C. Payne, E.D. Gorham, and C.A. Amos. 1990. *Evaluation of Severe Accident Risks: Quantification of Major Input Parameters.* NUREG/CR-4551, Volume 2. Washington, DC: Nuclear Regulatory Commission.

Hastie, R. 1986. Review essay: Experimental evidence on group accuracy. B. Grofman and G. Owen, eds. *Information Pooling and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy.* Greenwich, CT: JAI Press.

Hazard, T.H., and C.R. Peterson. 1973. *Odds Versus Probabilities for Categorical Events.* McLean, VA: Decisions and Designs, Inc.

Heger, A.S., and J.R. Hill. 1993. Probability networks for handling uncertainty in the performance assessment of high-level nuclear waste repositories. *Reliability Engineering and System Safety* 41: 13–20.

Helton, J.C. 1993. Risk, uncertainty in risk, and the EPA release limits for radioactive waste disposal. *Nuclear Technology* 101: 18–39.

Henrion, M., G.F. Fischer, and T. Mullin. 1989. *Divide and Conquer? The Effect of Decomposition on Accuracy and Calibration.* Pittsburgh, PA: Carnegie Mellon University: Department of Social and Decision Science.

Hill, G.W. 1982. Group versus individual performance: Are N+1 heads better than one? *Psychological Bulletin* 91: 517–539.

Hoch, S.J. 1985. Canterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11: 719–731.

Hogarth, R.M. 1977. Methods for aggregating opinions. *Decision Making and Change in Human Affairs.* H. Jungermann and G. DeZeeuw, eds. Dordrecht, Netherlands: Reidel: 231–255.

Hogarth, R.M. 1987. *Judgement and Choice: 2nd Ed.* Chichester, England: Wiley.

Hong, Y., and G. Apostolakis. 1993. Conditional influence diagrams in risk management. *Risk Analysis* 13: 625–636.

Hora, S.C. 1992. *Probabilities of Human Intrusion into the WIPP Methodology for 1992 Preliminary Comparison.* Internal Report to Department 6342. Albuquerque, NM: Sandia National Laboratories.

Hora, S.C., and R.L. Iman. 1989. Expert opinion in risk analysis: The NUREG-1150 methodology. *Nuclear Science and Engineering* 102: 323–331.

Hora, S.C., D. von Winterfeldt, and K.M. Trauth. 1991. *Expert Judgment on Inadvertent Human Intrusion into the Waste Isolation Pilot Plant.* SAND90-3063. Albuquerque, NM: Sandia National Laboratories.

Hora, S.C., N.G. Dodd, and J.A. Hora. 1993. The use of decomposition in probability assessments on continuous variables. *Journal of Behavioral Decision Making* 6: 133–147.

Humphreys, P. 1988. *Human Reliability Assessors Guide.* United Kingdom Atomic Energy Authority: Safety and Reliability Directorate.

Hunter, R.L., and C.J. Mann, eds. 1989. *Techniques for Determining Probabilities of Events and Processes Affecting the Performance of Geologic Repositories.* NUREG/CR-3964, Vol. 1. Washington, DC: Nuclear Regulatory Commission.

Hunter, R.L., G.E. Barr, and F.W. Bingham. 1982. *Preliminary Scenarios for Consequence Assessments of Radioactive-Waste Repositories at the Nevada Test Site.* SAND82-0426. Albuquerque, NM: Sandia National Laboratories.

Hunter, R.L., G.E. Barr, and F.W. Bingham. 1983. *Scenarios for Consequence Assessments of Radioactive-Waste Repositories at Yucca Mountain, Nevada Test Site.* SAND82-1277. Albuquerque, NM: Sandia National Laboratories.

International Atomic Energy Agency. 1981. *Safety Assessment for the Underground Disposal of Radioactive Wastes.* Safety Series No. 56. Vienna, Austria: International Atomic Energy Agency.

Janis, I.L. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes, 2nd Ed.* Boston, MA: Houghton Mifflin.

Janis, I.L., and L. Mann. 1977. *Decision Making.* New York, NY: Free Press.

Johnson, E.J. 1980. *Expertise in Admissions Judgment.* Unpublished Doctoral Dissertation. Carnegie-Mellon University: Department of Psychology.

Johnson, E.J., and A. Sathi. 1988. *Expertise in Security Analysts.* In Preparation.

Jouini, M.N., and R.T. Clemen. 1994. *Copula models for aggregating expert opinions.* Working paper. University of Oregon.

Kahn, H., and A.J. Wiener. 1967. *The Year 2000, A Framework for Speculation.* New York, NY: Macmillan.

Kahneman, D., and A. Tversky. 1972. Subjective probability: A judgement of representativeness. *Cognitive Psychology* 3: 430–454.

Kahneman, D., and A. Tversky. 1973. On the psychology of prediction. *Psychological Review* 80: 237–251.

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

Kaplan, S. 1992. Expert information versus expert opinions: Another approach to the problem of eliciting/combing/using expert knowledge in PRA. *Journal of Reliability Engineering and System Safety* 35: 61–72.

Keeney, R.L. 1992. *Value-Focused Thinking*. Cambridge, MA: Harvard University Press.

Keeney, R.L., and H. Raiffa. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York, NY: John Wiley & Sons.

Keeney, R.L., and D. von Winterfelt. 1989. *On the Uses of Expert Judgment on Complex Technical Problems*. IEEE Transactions on Engineering Management, 36, 83-86.

Keeney, R.L., and D. von Winterfeldt. 1991. Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management* 38: 191–201.

Kidd, J.B. 1970. The utilization of subjective probabilities in production planning. *Acta Psychologica* 34: 338–347.

Kyburg, H.E., and H.E. Smokler, eds. 1964. *Studies in Subjective Probability*. New York, NY: John Wiley & Sons.

Lave, L.B. 1981. *The Strategy of Social Regulation: Decision Frameworks for Policy*. Washington, DC: The Brookings Institution.

Lawrence, M.J., R.H. Edmundson, and M.J. O'Connor. 1986. The accuracy of combining judgmental and statistical forecasts. *Management Science* 32: 1521–1532.

Lichtenstein, S., and B. Fischhoff. 1980. Training for calibration. *Organizational Behavior and Human Performance* 26: 149–171.

Lichtenstein, S., B. Fischhoff, and L.D. Phillips. 1982. Calibration of probabilities: The state of the art to 1980. *Judgment under Uncertainty: Heuristics and Biases*. D. Kahneman, P. Slovic, and A. Tversky, eds. New York, NY: Cambridge University Press: 305–334.

Lindley, D.V. 1983. Reconciliation of probability distributions. *Operations Research* 31: 866–880.

Lindley, D.V. 1985a. *Making Decisions*. Second Edition. London, England: John Wiley & Sons Ltd.

Lindley, D.V. 1985b. Reconciliation of Discrete Probability Distributions. M.H. Bernardo, D.V. DeGroot, D.V. Lindley, and A.F.M. Smith, eds. *Bayesian Statistics* 2: 375–390.

Lindley, D., and A. Smith. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society* B(34): 1–41.

Linstone, H.A., and M. Turoff. 1975. *The Delphi Method: Techniques and Applications*. Reading, MA: Addison-Wesley.

Lock, A. 1987. Integrating group judgements in subjective forecasts. G. Wright and P. Ayton, eds. Chichester, England: Wiley. *Judgmental Forecasting*: 109–127.

Ludke, R.L., F.F. Strauss, and D.H. Gustafson. 1977. Comparison of five methods for estimating subjective probability distributions. *Organizational Behavior and Human Performance* 19: 162–179.

Lusted, L.B. 1977. *A Study of the Efficacy of Diagnostic Radiologic Procedures.* Chicago, IL: American College of Radiology.

MacGregor, D., S. Lichtenstein, and P. Slovic. 1988. Structuring knowledge retrieval: An analysis of decomposing quantitative judgements. *Organizational Behavior and Human Decision Processes* 42: 303–323.

Maines, L. 1994. *An Experimental Examination of Factors Influencing the Subjective Combining of Forecasts.* Duke University: Fuqua School of Business. In preparation.

McGarity, T.O. 1984. Judicial review and scientific rulemaking. *Science, Technology & Human Values* 9: 97–106.

McGuire, R.K., ed. 1990. *Demonstration of Risk-Based Approach to High-Level Waste Repository Evaluation.* EPRI NP-7057. Palo Alto, CA: Electric Power Research Institute.

McGuire, R.K., ed. 1992. *Demonstration of Risk-Based Approach to High-Level Waste Repository Evaluation: Phase 2.* EPRI TR-100384. Palo Alto, CA: Electric Power Research Institute.

Mendel, M.B., and T.B. Sheridan. 1989. Filtering information from human experts. *IEEE Transactions on Systems, Man, and Cybernetics* 36: 6–16.

Merkhofer, M.W., and A.K. Runchal. 1989. Probability encoding: Quantifying uncertainty over hydrologic parameters for basalt. *Geostatistical, Sensitivity, and Uncertainty Methods for Ground-Water Flow and Radionuclide Transport Modeling.* Columbus, OH: Battelle Press: 629–648.

Meyer, M.A., and J.M. Booker. 1987. *Sources of Correlation Between Experts: Empirical Results from Two Extremes.* NUREG/CR-4814. Washington, DC: Nuclear Regulatory Commission.

Meyer, M.A., and J.M. Booker. 1991. *Eliciting and Analyzing Expert Judgement, A Practical Guide, Knowledge-Based Systems.* Volume 5. San Diego, CA: Academic Press.

Minarick, J.W., and C.A. Kukielka. 1982. *Precursors to Potential Severe Core Damage Accidents: 1969-1979, A Status Report.* NUREG/CR-2497. Washington, DC: Nuclear Regulatory Commission.

Minarick, J.W., J.D. Harris, P.N. Austin, J.W. Cletcher, and E.W. Hagen. 1986. *Precursors to Potential Severe Core Damage Accidents: 1985, A Status Report.* NUREG/CR-4674. Washington, DC: Nuclear Regulatory Commission.

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

Morris, P.A. 1974. Decision analysis expert use. *Management Science* 20: 1,233–1,241.

Morris, P.A. 1977. Combining expert judgements: A Bayesian approach. *Management Science* 23: 679–693.

Mosleh, A., V.M. Bier, and G. Apostolakis. 1987. *Methods for the Elicitation and Use of Expert Opinion in Risk Assessment*. NUREG/CR-4962, PLG-0533. Washington, DC: Nuclear Regulatory Commission.

Mosleh, A., V.M. Bier, and G. Apostolakis. 1988. A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering and System Safety* 20: 63–85.

Murphy, A.H., and R.L. Winkler. 1987. A general framework for forecast verification. *Monthly Weather Review* 115: 1,330–1,338.

Murphy, A.H., and R.L. Winkler. 1992. Diagnostic verification of probability forecasts. *International Journal of Forecasting* 7: 435–455.

National Research Council. 1984. *Risk Assessment in the Federal Government: Managing the Process*. Washington, DC: National Academy Press.

National Research Council. 1990. *Health Effects of Exposure to Low Levels of Ionizing Radiation: BEIR V*. Washington, DC: National Academy Press: 20–21. do I need pages? is this an entire book?

National Research Council. 1992. *Combining Information*. Washington, DC: National Academy Press.

National Research Council. 1994. *Building Consensus Through Risk Assessment and Management of the Department of Energy's Environmental Remediation Program*. Washington, DC: National Academy Press.

Neapolitan, R.E. 1990. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. New York, NY: John Wiley & Sons.

Newell, A., and H.A. Simon. 1972. *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nuclear Energy Agency. 1987. *Uncertainty Analysis for Performance Assessment of Radioactive Waste Disposal Systems*. Paris, France: Nuclear Energy Agency: Organization for Economic Cooperation and Development.

Nuclear Energy Agency. 1989. *Risk Associated with Human Intrusion at Radioactive Waste Disposal Sites*. Paris, France: Nuclear Energy Agency: Organization for Economic Cooperation and Development.

Nuclear Energy Agency. 1990. *Safety Assessment of Radioactive Waste Repositories: Paris Symposium.* Paris, France: Nuclear Energy Agency: Organization for Economic Cooperation and Development.

Nuclear Energy Agency. 1992. *Systematic Approaches to Scenario Development.* Paris, France: Nuclear Energy Agency: Organization for Economic Cooperation and Development.

Nuclear Energy Agency. 1993. *Future Human Actions at Radioactive Waste Disposal Sites, Draft Report.* Paris, France: Nuclear Energy Agency: Organization for Economic Cooperation and Development.

Nuclear Regulatory Commission. 1975. *Reactor Safety Study—An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants (WASH-1400).* NUREG-75/014. Washington, DC: Nuclear Regulatory Commission.

Nuclear Regulatory Commission. 1984. *Probabilistic Risk Assessment (PRA) Reference Document: Final Report.* NUREG-1050. Washington, DC: Nuclear Regulatory Commission.

Nuclear Regulatory Commission. 1985. *A Review of the Current Understanding of the Potential for Containment Failure from In-Vessel Steam Explosions.* NUREG-1116. Washington, DC: Nuclear Regulatory Commission.

Nuclear Regulatory Commission. 1989. *Reactor Risk Reference Document.* NUREG-1150. Washington, DC: Nuclear Regulatory Commission.

Nuclear Regulatory Commission. 1990. *Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants: Final Summary Report.* NUREG-1150, Vol. 1. Washington, DC: Nuclear Regulatory Commission.

Nuclear Regulatory Commission. 1991. *Disposal of High-Level Radioactive Wastes in Geologic Repositories.* Title 10, Energy, Part 60 (10 CFR Part 60). Washington, DC: U.S. Government Printing Office.

Nuclear Regulatory Commission. 1993a. *Status of Risk Harmonization with the Environmental Protection Agency (EPA) under Section D of the March 1992 Memorandum of Understanding (MOU).* SECY-93-134. Washington, DC: Nuclear Regulatory Commission.

Nuclear Regulatory Commission. 1993b. *Domestic Licensing of Production and Utilization Facilities.* Title 10, Energy, Part 50 (10 CFR Part 50). Washington, DC: U.S. Government Printing Office.

Otway, H., and D. von Winterfeldt. 1992. *Expert Judgement in Risk Analysis and Management: Process, Context, and Pitfalls.* Risk Analysis. 12: 83-93.

Parenté, F.J., and J.K. Anderson-Parenté. 1987. Delphi inquiry systems. *Judgmental Forecasting* 129-156.

Park, W. 1990. A review of research on groupthink. *Journal of Behavioral Decision Making* 3: 229–245.

Park, J.R., N.A. Eisenberg, J.T. Buckley, R.G. Baca, and E.J. Bonano. 1994. *The Nuclear Regulatory Commission Strategic Plan for Postclosure Performance Assessment Activities for the High-Level Waste Geologic Repository, Draft Report.* Washington, DC: Nuclear Regulatory Commission. In preparation.

Phillips, L.D. 1984. A theory of requisite decision models. *Acta Psychologica* 56: 29–48.

Phillips, L.D. 1987. On the adequacy of judgmental forecasts. *Judgmental Forecasting* 11–30.

Phillips, L.D., and M.C. Phillips. 1990. *Facilitated Work Groups: Theory and Practice.* Unpublished manuscript. London, England: London School of Economics and Political Science.

Plous, S. 1993. *The Psychology of Judgement and Decision Making.* New York, NY: McGraw-Hill.

Raiffa, H. 1968. *Decision Analysis.* Reading, MA: Addison-Wesley.

Ramsey, F.P. 1931. *The Foundation of Mathematics and Other Logical Essays.* London, England: Kegan Paul.

Ravinder, H.V., D.N. Kleinmuntz, and J.S. Dyer. 1988. The reliability of subjective probabilities obtained through decomposition. *Management Science* 34: 186–199.

Reber, A.S. 1976. Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory* 2: 88–94.

Rechard, R.P., K.M. Trauth, J.S. Rath, R.V. Guzowski, S.C. Hora, and M.S. Tierney. 1993. *The Use of Formal and Informal Expert Judgments for Performance Assessments.* SAND92-1148. Albuquerque, NM: Sandia National Laboratories.

Reckhow, K. 1988. A comparison of robust Bayes and classical estimators for regional lake models of fish response to acidification. *Water Resources Research* 24: 1,061–1,068.

Redus, K.S. 1994. A Bayesian approach for data fusion in waste characterization. *Proceedings from Spectrum '94, Nuclear and Hazardous Waste Management International topical Meeting, August 14-18, 1994, Atlanta, Georgia.* LaGrange Park, IL: American Nuclear Society.

Reid, D.J. 1968. Combining three estimates of gross domestic product. *Economica* 35: 431–444.

Roberds, W.J. 1990. Methods for developing defensible subjective probability assessments. *Soils, Geology and Foundations in Geotechnical Engineering.* Transportation Research Board Document No. 1288. Washington, DC: National Research Council: 183-190.

Rogovin, M. 1980. *Three Mile Island—A Report to the Commissioners and to the Public.* NUREG/CR-1250, Volume 1. Washington, DC: Nuclear Regulatory Commission.

Rohrbaugh, J. 1979. Improving the quality of group judgement: Social judgment analysis and the Delphi technique. *Organizational Behavior and Human Performance* 24: 73–92.

Ross, B.. 1987. *A First Survey of Disruption Scenarios for a High-Level Waste Repository at Yucca Mountain, Nevada*. SAND85-7117. Albuquerque, NM: Sandia National Laboratories.

Russo, J.E., and P.J.H. Schoemaker. 1992. Managing overconfidence. *Sloan Management Review* 33: 7–17.

Savage, L.J. 1954. *The Foundations of Statistics*. New York, NY: John Wiley & Sons.

Schlaefer, R. 1959. *Probability and Statistics for Business Decisions*. New York: McGraw-Hill.

Seaver, D.A. 1978. *Assessing Probability with Multiple Individuals: Group Interaction Versus Mathematical Aggregation*. Report No. 78-3. Los Angeles, CA: University of Southern California: Social Science Research Institute.

Silbergeld, E.K. 1991. *Risk Assessment and Risk Management: An Uneasy Divorce*. Acceptable Evidence: Science and Values in Risk Management. D.G. Mayo and R.D. Hollander, eds. Oxford University Press: 99-114.

Slovic, P. 1991. Beyond numbers: A broader perspective on risk perception and risk communication. *Acceptable Evidence: Science and Values in Risk Management*. D.G. Mayo and R.D. Hollander, eds. New York, NY: Oxford University Press: 48–65.

Slovic, P., B. Fischholl, and S. Lichtenstein. 1982. Facts versus fears: Understanding perceived risk. *Judgment Under Uncertainty*. D. Kahneman, P. Slovic, and A. Tversky, eds. Cambridge, MA: Cambridge University Press.

Smith, A.F.M. 1990. The Bayesian approach to probabilistic risk assessment. *Methods for the Treatment of Different Types of Uncertainty*. PSAC/DOC (90)11. Paris, France: Nuclear Energy Agency: Organization for Economic Cooperation and Development: 89–104.

Sniezek, J.A. 1989. An examination of group process in judgmental forecasting. *International Journal of Forecasting* 5: 171–178.

Sniezek, J.A., and R.A. Henry. 1989. Accuracy and confidence in group judgement. *Organizational Behavior and Human Decision Processes* 43: 1–28.

Sniezek, J.A., and R.A. Henry. 1990. Revision, weighting, and commitment in consensus group judgement. *Organizational Behavior and Human Decision Processes* 45: 66–84.

Spetzler, C.S., and C.-A.S. Stael von Holstein. 1975. *Probability Encoding in Decision Analysis*. Management Science. Vol. 22. pp. 340-352.

Stael von Holstein, C.A. 1972. Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance* 8: 139–158.

Stephens, M.E., and B.W. Goodwin. 1990. Scenario analysis for the postclosure assessment of the Canadian concept for nuclear fuel waste disposal. *Safety Assessment of Radioactive Waste Repositories.* Paris, France: Nuclear Energy Agency: Organization for Economic Cooperation and Development: 405–416.

Stephens, M.E., B.G. Goodwin, and T.H. Andres. 1993. Deriving parameter probability density functions. *Reliability Engineering and System Safety* 42: 271-292.

Stone, M. 1961. The opinion pool. *Annals of Mathematical Statistics* 32: 1,339–1,342.

Swain, A.D, and H.E. Guttmann. 1983. *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Applications.* NUREG/CR-1278. Washington, DC: Nuclear Regulatory Commission.

Swedish Nuclear Power Inspectorate. 1991. *SKI Project 90.* Technical Report 91:23. Stockholm, Sweden: Swedish Nuclear Power Inspectorate.

Thompson, B.G.J. 1988. *A Method of Overcoming the Limitation of Conventional Scenario-Based Assessments by Using Monte Carlo Simulation of Possible Future Environmental Changes.* PAAG/DOC/88/11. Paris, France: Nuclear Energy Agency: Organization for Economic Cooperation and Development.

Thompson, B.G.J., and B. Sagar. 1993. The development and application of integrated procedures for post-closure assessment, based upon Monte Carlo simulation: The probabilistic systems approach (PSA). *Reliability Engineering and System Safety* 42: 125–160.

Thorne, M.C. 1992. *Development of a Conceptual Model of a Radioactive Waste Disposal System for the Purpose of Conducting an Assessment of the Biases Inherent in a Post-Closure Radiological Assessment of that System.* Horsham, England: Electrowatt Engineering Services (UK) Limited.

Thorne, M.C. 1993. The use of expert opinion in formulating conceptual models of underground disposal systems and the treatment of associated biases. *Reliability Engineering and System Safety* 42: 161–180.

Tierney, M.S. 1990. *Constructing Probability Distributions of Uncertain Variables in Models of the Performance of the Waste Isolation Pilot Plant: The 1990 Performance Simulations.* SAND90-2510. Albuquerque, NM: Sandia National Laboratories.

Trauth, K.M., S.C. Hora, R.P. Rechard, and D.R. Anderson. 1992. *The Use of Expert Judgment to Quantify Uncertainty in Solubility and Sorption Parameters for Waste Isolation Plant Performance Assessment.* SAND92-0479. Albuquerque, NM: Sandia National Laboratories.

Trauth, K.M., S.C. Hora, and R.V. Guzowski. 1993. *Expert Judgment on Markers to Deter Inadvertent Human Intrusion into the Waste Isolation Pilot Plant*. SAND92-1382. Albuquerque, NM: Sandia National Laboratories.

Tschoepe, E., and L. Abramson. 1992. *Substantially Complete Containment Elicitation Report*. CNWRA 92-016. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

Tversky, A., and D. Kahneman. 1982a. Evidential impact of base rates. Judgment Under Uncertainty. New York, NY: Cambridge University Press.

Tversky, A., and D. Kahneman. 1982b. Heuristics and biases. D. Kahneman, P. Slovic, and A. Tversky, eds. *Judgment Under Uncertainty*. Cambridge, MA: University Press.

Tweney, R.D., M.E. Doherty, W.J. Warner, and D.B. Pliske. 1980. Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology* 32: 109–124.

U.S. Department of Energy. 1986. *A Multiattribute Utility Analysis of Sites Nominated for Characterization for the First Radioactive-Waste Repository—A Decision Aiding Methodology*. DOE/RQ-0074. Washington, DC: Department of Energy: Office of Civilian Radioactive Waste Management.

U.S. Department of Energy. 1988. *Site Characterization Plan, Yucca Mountain Site, Nevada Research and Development Area, Nevada*. Washington, DC: Department of Energy: Office of Civilian Radiation Waste Management. Volume II, Part A, Chapter 5: 5–1 to 5–106.

von Mises, R. 1957. *Probability, Statistics, and Truth*. London, England: George Allen & Unwin.

von Winterfeldt, D., and W. Edwards. 1986. *Decision Analysis and Behavioral Research*. New York, NY: Cambridge University Press.

Wallsten, T.S., and D.V. Budescu. 1983. Encoding subjective probabilities: A psychological and psychometric review. *Management Science* 29: 151–173.

Wason, P.C. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology* 12: 129–140.

West, M. 1994. Multi-agent opinion models. *Annals of Statistics*. In Press.

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: 197-207.

Wilson, M. L., J.H. Gauthier, R.W. Barnard, G.E. Barr, H.A. Dockery, E. Dunn, R.R. Eaton, D.C. Guerin, N. Lu, M.J. Martinez, R. Nilson, C.A. Rautman, T.H. Robey, B. Ross, E.E. Ryder, A.R. Schenker, S.A. Shannon, L.H. Skinner, W.G. Halsey, J.D. Gansemer, L.C. Lewis, A.D. Lamont, I.R. Triay, A. Meijer, and D.E. Morris. 1994. *Total-System Performance Assessment for Yucca Mountain—SNL Second Iteration (TSPA 1993)*. SAND93-2675, 3 Volumes. Albuquerque, NM: Sandia National Laboratories.

Winkler, R.L., T.S. Wallsten, R.G. Whitfield, H.M. Richmond, S.R. Hayes, and A.S. Hayes. 1994. *An Assessment of The Risk of Chronic Lung Injury Attributable to Long-Term Ozone Exposure.* Operations Research.

Winkler, R.L. 1968. The consensus of subjective probability distributions. *Management Science* 15: 361–375.

Winkler, R.L. 1981. Combining probability distributions from dependent information sources. *Management Science* 27: 479–488.

Winkler, R.L., and R. Clemen. 1989. *Combining Dependent Information: Empirical Data and Expert Judgments*. University of Oregon. Unpublished Manuscript.

Winkler, R.L., and R.M. Poses. 1993. Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science* 39: 1526–1543.

Winkler, R.L, S.C. Hora, and R.G. Baca. 1992. *The Quality of Experts' Probabilities Obtained Through Formal Elicitation Techniques.* CNWRA Technical Letter Report. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

Winkler, R., W. Smith, and R. Kulkarni. 1978. Adaptive forecasting models based on predictive distributions. *Management Science* 24: 977–986.

Wright, W.F. 1979. Properties of judgment models in a financial setting. *Organizational Behavior and Human Performance* 23: 73–85.

Wright, G., C. Saunders, and P. Ayton. 1988. The consistency, coherence and calibration of holistic, decomposed, and recomposed judgmental probability forecasts. *Journal of Forecasting* 7: 185–199.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York, NY: John Wiley & Sons.

Zimmerman, R. 1990. *Governmental Management of Chemical Risk, Regulatory Processes for Environmental Health.* Lewis Publishers.

Zimmerman, D.A., E.J. Bonano, P.A. Davis, C.P. Harlan, and M.S.Y. Chu. 1992. *Peer Review of the U.K. DoE Dry Run 3 Exercise (A Trial Probabilistic Risk Assessment of a Hypothetical Nuclear Waste Repository at Harwell, Oxfordshire).* SAND92-1945. Albuquerque, NM: Sandia National Laboratories.