A Practical Guide on Conducting Expert-Opinion Elicitation of Probabilities and Consequences for Corps Facilities

by Bilal M. Ayyub, PhD, PE

Prepared
for U.S. Army Corps of Engineers
Institute for Water Resources
Alexandria, VA 22315-3868

January 2001 **IWR Report 01-R-01**

Preface

This report is a product of the Corps of Engineers' Risk Analysis for Water Resources Investments Research Program managed by the Institute for Water Resources. The report was prepared to fulfill work units in the research program concerning risk management strategy. These work units focus on developing and applying the concepts of risk communication to water resources issues. The report conforms to the basic planning model and to the risk and uncertainty analysis recommendations presented in "Economic and Environmental Principles and Guidelines for Water related Land Resources Implementation Studies" (P&G).

The purpose of this research project was to define techniques for eliciting expert opinion on possible events and their consequences for Corps facilities for use by planners, engineers, and others. The report consists of three chapters, a bibliography, and three appendices. The chapters provide a practical discussion of terms and concepts; detailed discussion of the expert elicitation process followed by an example of processes with results. Appendices provided detailed discussion of pertinent statistical nomenclature, failure consequences, and heuristics, elicitation, scoring, and aggregation of data.

This report was prepared by Bilal M. Ayyub, PhD, PE, under terms of a contract with the U.S. Army Corps of Engineers Institute for Water Resources. Dr. David A. Moser was the contract manager for the report and is the manager of the Risk Analysis for Water Resources Investments Research Program. Dr. Moser, assisted by Ms. Darlene R. Guinto, served as final editors. This research was initially prepared under the supervision of Mr. Michael Krouse, retired Chief of the Decision Methodologies Division and Mr. Kyle Schilling, retired Director of IWR.

Acknowledgments

The authors would like to acknowledge the financial support of the U.S. Army Corps of Engineers of this study, and the facilitation and support provided by the Planning & Management Consultants, Limited, and Dr. John F. Langowski, Jr. Also, the opportunity, support and valuable comments provided by Dr. David Moser are greatly appreciated.

Table of Contents

PREFACE iii
ACKNOWLEDGEMENTS v
LIST OF FIGURES ix
LIST OF TABLES ix
ABSTRACT xi

1. INTRODUCTION 1
   1.1. Ignorance, Knowledge and Uncertainty 1
   1.2. Historical Background 3
      1.2.1. Delphi Method 5
      1.2.2. Scenario Analysis 9
   1.3. Objectives and Scope 10

2. THE EXPERT-OPINION ELICITATION PROCESS 11
   2.1. Introduction and Terminology 11
      2.1.1. Theoretical Bases 11
      2.1.2. Terminology 11
      2.1.3. Classification of Issues, Study Levels, Experts, and Process Outcomes 13
   2.2. Process Definition 17
      2.2.1. Need Identification for Expert-Opinion Elicitation 18
      2.2.2. Selection of Study Level and Study Leader 18
      2.2.3. Selection of Peer Reviewers and Experts 19
         2.2.3.1. Selection of Peer Reviewers 19
         2.2.3.2. Identification and Selection of Experts 19
         2.2.3.3. Items to be Sent to Experts and Reviewers Before the Expert-Opinion Elicitation Meeting 21
      2.2.4. Identification, Selection and Development of Technical Issues 21
      2.2.5. Elicitation of Opinions 22
         2.2.5.1. Issue Familiarization of Experts 22
         2.2.5.2. Training of Experts 23
         2.2.5.3. Elicitation and Collection of Opinions 23
         2.2.5.4. Aggregation and Presentation of Results 24
         2.2.5.5. Group Interaction, Discussion and Revision by Experts 24
      2.2.6. Documentation and Communication 24
   2.3. Example Expert-Opinion Elicitation Processes with Results 24
      2.3.1. Cargo Elevators Onboard Ships 24
      2.3.2. Navigation Locks 27

3. CONCLUSIONS 33

4. BIBLIOGRAPHY 35

APPENDIX A. OCCURRENCE PROBABILITIES, MOMENTS AND PERCENTILES A-1

A.1. Background A-1
A.2. Definition of Probability A-2
A.2.1. Linguistic Probabilities A-3
A.2.2. Unsatisfactory-Performance Rate A-5
A.3. Central Tendency Measures A-5
A.3.1. Mean (or Average) Value A-5
A.3.2. Average Time Between Unsatisfactory Performances A-5
A.3.3. Median Value A-6
A.4. Dispersion (or Variability) A-6
A.5. Percentiles A-7
A.6. Statistical Uncertainty A-8

APPENDIX B. UNSATISFACTORY-PERFORMANCE CONSEQUENCES B-1
B.1. Consequence Types B-1
B.1.1. Production Loss B-1
B.1.2. Property Damage B-2
B.1.3. Flood Inundation B-2
B.1.4. Other Consequence Types B-3
B.2. Assessment of Consequences B-5
B.2.1. Unsatisfactory-Performance and Loss Records B-5
B.2.2. Unsatisfactory-Performance Databases B-5
B.2.3. Cause-Consequence Diagrams B-6
B.2.4. Formal Expert-Opinion Elicitation and Questionnaires B-7

APPENDIX C. HEURISTICS, ELICITATION, SCORING AND AGGREGATION C-1
C.1. Introduction C-1
C.2. Scientific Heuristics C-1
C.3. Elicitation and Scoring Methods C-5
C.3.1. Elicitation Methods C-6
C.3.1.1. Indirect Elicitation C-6
C.3.1.2. Direct Method C-7
C.3.1.3. Parametric Estimation C-8
C.3.2. Scoring Methods C-8
C.3.2.1. Self Scoring C-8
C.3.2.2. Collective Scoring C-8
C.3.2.3. Entropy and Discrepancy Measures C-8
C.4. Combining Expert Opinions C-9
C.4.1. Rational Consensus C-9
C.4.2. Consensus Combination of Opinions C-9
C.4.3. Percentiles for Combining Opinions C-10
C.4.4. Weighted Combinations of Opinions C-10
C.4.5. Opinion Aggregation Using Interval Analysis, Fuzzy Numbers and Uncertainty Measures C-10
C.5. Methods of Educational and Psychological Testing, and Social Research C-11
C.5.1. Standards for Educational and Psychological Testing C-11
C.5.2. Methods of Social Research C-14

LIST OF FIGURES

Figure 1-1: Estimated Duration for Thermonuclear Postwar Economic Recuperation.....5
Figure 2-1: Outcomes of the Expert-opinion Elicitation Process.....16
Figure 2-1: Expert-Opinion Elicitation Process.....16
Figure 2-3a: Emsworth Navigation Lock on the Ohio River (Ayyub et. al. 1996) 29
Figure 2-3b: Details for Emsworth Navigation Lock on the Ohio River (Ayyub et. al. 1996) 29
Figure B-1: Example Calculation of Potential Loss of Life for a warning time of One Hour .B-4
Figure C-1: Heuristics C-2
Figure C-2: Stages of Social Research C-15

LIST OF TABLES

Table 1-1: Delphi Questionnaire (Helmer 1968).....7
Table 1-2: Likelihood of Occurrence (Wiggins 1985).....8
Table 1-3: Consequence (Wiggins 1985).....8
Table 1-4: Risk Assessment Matrix (Wiggins 1985).....9
Table 2-1: Terminology and Definitions.....12
Table 2-2: Issue Degrees and Study Levels (Constructed based on NRC, 1997).....15
Table 2-3: Guidance on the Use of Peer Reviews (NRC 1997).....15
Table 2-4: Expert-opinion Elicitation for Example Issue I (Ayyub 1992 and Ayyub et. al. 1996).....25
Table 2-5: Expert-opinion Elicitation for Example Issue 2 (Ayyub 1992 and Ayyub et. al. 1996).....26
Table A-1: Linguistic Probabilities and Translations A-4
Table C-1: Reliability and Accuracy Ratings in Intelligence Information C-5
Table C-2: A Kent Chart C-5
Table C-3: Money Required to Double Happiness C-7
Table C-4: Selected Validity Standards from the Standards for Educational and Psychological Testing C-14

Abstract

Risk analysis and risk-based decision making for maintaining the integrity of the U. S. Army Corps of Engineers (USACE) facilities require the knowledge of two main quantities for components, and systems, their unsatisfactory-performance probability and consequences. Sometimes this information is not available from historical records, prediction methods or literature review. Also, sometimes there is a need to perform a preliminary risk evaluation of components within a system for the purpose of planning future reliability and risk analyses or for the purpose of performing initial screening of components. Also, in many situations classical frequency analysis cannot be used or is prohibitively expensive to quantify the existing risk or the change in risk related to USACE activities. These cases typically lack historical data, or USACE activities create new conditions without useful data for risk analysis. In addition, models to assist in quantifying risk may not be available or may be very data intensive. The USACE is increasingly turning to using expert opinions in a variety of analyses. Some of the issues already investigated by the Corps include (a) quantifying the probability of failure of navigation lock components, (b) quantifying the probability and consequence of navigation lock closure, (c) estimating the probability of events requiring emergency gate usage at hydropower plants, and (d) predicting the vessel safety improvements from deep channel widening. Expert-opinion elicitation can provide the USACE with a means of gaining information on these essential risk-related quantities.

The expert-opinion elicitation process is defined as a formal, heuristic process of obtaining information or answers to specific questions about certain quantities, called issues, such as unsatisfactory-performance rates, unsatisfactory-performance consequences and expected service life. Expert-opinion elicitation should not be used in lieu of rigorous reliability and risk analytical methods, but should be used to supplement them and to prepare for them. not a big deal, but an example of how this technique could supplement risk analysis might be helpful. Also, it should be used in cases where reliability and risk analytical methods are inappropriate or inconsistent. what is a case where this is true? It should be preferably performed during a face-to-face meeting of members of an expert panel that is developed specifically for the issues under consideration. The meeting of the expert panel should be conducted after communicating to the expert in advance to the meeting background information, objectives, list of issues, and anticipated outcome from the meeting. In this document, the different components of the expert-opinion elicitation process are described, and the process itself is outlined and discussed.

This guide defines a process for conducting expert-opinion elicitation of probabilities and consequences for Corps facilities for the use of planners, engineers, and others should they choose to use expert judgment. The guide documents techniques for eliciting expert opinion on possible events and their probabilities for application to Corps facilities. Historical background on the development of expert-opinion elicitation, its limitations, current uses, and example applications relevant to different engineering, planning, and operations decisions problems are provided in the guide. Because using expert judgment

can be easily abused, the guide provides a process for the use of this technique and limitations of the method. suggestion: Due to the vulnerability of expert elicitation methods to bias and error, the guide details procedures and limitations of the method. The guide provides users with acceptable practices for expert-opinion elicitation in situations with a scarcity of historical data and areas where they are best suited.

1. Introduction

Risk studies for maintaining the integrity of Corps facilities require the assessment of unsatisfactory-performance probabilities and consequences of engineering systems. This chapter provides introductions to uncertainty in the context of knowledge and ignorance as background information for presenting expert-opinion elicitation methods. Also, historical background on expert opinion elicitation is provided.

This practical guide was developed based on the reported study by Ayyub (1999) on this subject. This report covers in detail with references various topics presented in that guide, which is recommended for further reading.

1.1. Ignorance, Knowledge and Uncertainty

The development of engineering models requires knowledge and information. Knowledge can be defined as the body of truth, information, and principles acquired by humankind about a system of interest. Information is a subset of knowledge that is acquired by investigation, study, or instruction about a system. However, knowledge and information about the system might not constitute the absolute state of the system's existence, i.e., its absolute truth. Knowledge is defined in the context of the humankind, and cannot be removed from it. As a result, knowledge would always reflect the imperfect nature of humans that can be attributed to their reliance on their senses for knowledge acquisition, and mind for extrapolation, creativity and imagination, bias, and their preconceived notions due to time asymmetry. An important domain in defining the absolute truth of a system is non-knowledge or ignorance.

Engineering is a practice that often tries to make statements about the future. However, knowledge is primarily the product of the past as we know more about the past than the future. For example, we can precisely describe past daily temperatures, but cannot accurately forecast future temperatures. Time asymmetry of knowledge can be attributed to several factors of which the significant ones are

1. our limited capacity to free ourselves from the past in order to forecast in the future;
2. our inability to go back in time and verify historical claims, therefore it gives us overconfidence in the superiority of our present knowledge; and
3. the unidirectional nature of causation to the past but not the future. We tend to explain phenomena based on antecedents rather than consequences. Therefore, we assume that causes precede effects. Although, the order can be switched for some systems, as someone might be creating the effects needed for some causes. The unidirectional temporal nature of explanation might not be true all the times, and sometimes can be non-verifiable.

Engineers tend to be preoccupied more with what will happen than what has happed. This preoccupation might result in bias and time asymmetry. Engineering systems can be characterized by their goals as well as by their causes, thereby removing some of this asymmetry.

Knowledge and ignorance can be rightfully argued that they are not absolute, and are socially constructed and negotiated. A non-absolute working definition of ignorance can be taken as "Expert A is ignorant from B's viewpoint if A fails to agree with or show awareness of ideas which B defines as actually or potentially valid (Smithson 1988)." This definition allows for self-attributed ignorance, and either Expert A or B can be attributer or perpetrator of ignorance. Ignorance can be classified based on its nature into two types, error and irrelevance. Error is a state of being ignorant of something as defined by its components. Irrelevance can be due to untopicality, taboo, and undecidability. Untopicality can be attributed to intuitions of experts that are negotiated with others in terms of cognitive relevance. Taboo is due to socially reinforced irrelevance. Issues that people must not know, deal with, inquire about, or investigate define the domain of taboo. Undecidability deals with issues that cannot be designated true or false because they are considered insoluble, or solutions that are not verifiable.

Error has two primary components of distortion and incompleteness. Distortion can result from assignments and substitutions that are wrong, conflicting or biased producing confusion, conflict or inaccuracy, respectively. Incompleteness consists of absence due to incompleteness in kind, and uncertainty. Uncertainty can be due to ambiguity, probability and/or vagueness. Ambiguity includes unspecificity and nonspecificity as a result of outcomes or assignments that are incompletely and improperly defined, respectively. Probability can be due to physical randomness, statistical or modeling uncertainty. Statistical uncertainty arises from using samples to characterize populations. Modeling uncertainty arises from using analytical models to predict system behavior. Vagueness is due to uncertainties of memberships to sets (i.e., fuzziness) and boundaries of sets (i.e., roughness).

System engineering provides a general framework for engineering analysis and design. The system definition can be based on observations at different system levels in the form of a hierarchy. An epistemological hierarchy of systems suited to the representation of engineering problems with a generalized treatment of uncertainty can provide realistic assessments of systems (Klir 1985, Klir and Folger 1988).

Uncertainty modeling and analysis in engineering started with the employment of safety factors using deterministic analysis, then was followed by probabilistic analysis with reliability-based safety factors. Uncertainty in engineering was also classified into objective and subjective types. The objective types included the physical, statistical and modeling sources of uncertainty. The subjective types were based on lack of knowledge and expert-based assessment of engineering variables and parameters. Similar classifications are utilized in quantitative risk analysis for policy related areas (Morgan and Henrion 1992).

Uncertainties in engineering systems can mainly be attributed to ambiguity and vagueness in defining the architecture, parameters and governing prediction models for the systems (Ayyub 1992 and 1994).

Stochastic modeling and analysis is needed in cases of probabilistic, ambiguous or epistemic uncertainty. Cognitive, vague or aleatory uncertainty can be handled using fuzzy sets and logic in other modeling scenarios (Paté-Cornell 1996, and Blair and Ayyub 1999). The ambiguity component is generally due to non-cognitive sources. These sources include (1) physical randomness; (2) statistical uncertainty due to the use of sampled information to estimate the characteristics of these parameters; (3) lack of knowledge; and (4) modeling uncertainty which is due to simplifying assumptions in analytical and prediction models, simplified methods, and idealized representations of real performances. The vagueness-related uncertainty is due to cognitive sources that include (1) the definition of certain parameters, e.g., structural performance (failure or survival), quality, deterioration, skill and experience of construction workers and engineers, environmental impact of projects, conditions of existing structures; (2) other human factors; and (3) defining the inter-relationships among the parameters of the problems, especially for complex systems. Other sources of uncertainty can include conflict in information, and human and organizational errors.

Analysis of engineering systems commonly starts with a definition of a system that can be viewed as an abstraction of the real system. The abstraction is performed at different epistemological levels (Ayyub 1992 and 1994). The resulting model can depend largely on an analyst or engineer; hence the subjective nature of this process. During the process of abstraction, the engineer needs to make decisions regarding what aspects should or should not be included in the model. These aspects include the previously identified uncertainty types. In addition to the abstracted and non-abstracted aspects, unknown aspects of the system can exist, and they are more difficult to deal with because of their unknown nature, sources, extents, and impact on the system.

Uncertainty modeling and analysis for the abstracted aspects of the system need to be performed with a proper consideration of the non-abstracted aspects of a system. The division between abstracted and non-abstracted aspects can be a division of convenience that is driven by the objectives of the system modeling, or simplification of the model. However, the unknown aspects of the systems are due to ignorance and lack of knowledge. These aspects depend on the knowledge of the analyst, and the state of knowledge about the system in general. The effects of the unknown aspects on the ability of the system model to predict the behavior of the real system can range from none to significant.

1.2. Historical Background

The development of structured methods for expert-opinion elicitation was done by the RAND (**R***esearch* **AND D***evelopment*) corporation of Sante Monica, California. The RAND corporation resulted from a joint U. S. Air Force and Douglas Aircraft effort in 1946 called *Project RAND*. In its first year of operation, RAND predicted the first space satellite would be launched in the middle of 1957. The prediction was accurately validated by the Russian Sputnik launch on October 4, 1957. In 1948, RAND split off from Douglas Aircraft as the first think-tank type of a corporation. The research of RAND can be classified into four broad categories: (1) methodology, (2) strategic and tactical planning, (3) international relations, and (4) new technology. Almost all of these categories can rely heavily on expert opinions. In its early days between World War II and Vietnam War, RAND

developed two methods for structured elicitation of expert opinions: (1) Delphi method, and (2) scenario analysis.

Example 1-1. Fallacy of Civil Defense Strategic Planning of the 1960s

Herman Kahn led several RAND studies that were funded by the U. S. Air Force on the effects of thermonuclear war and civil defense (Cooke 1991). He later founded the Hudson Institute in New York. He articulated the strategic posture of *finite deterrence* and its upgrade to *credible first strike capability* for *Thermonuclear War* (Kahn 1960). The finite deterrence requires maintaining an ability to inflict unacceptable damage on an enemy after absorbing a surprise nuclear attack. This strategy can be augmented by *counterforce measures* to limit enemy-attack effects by building, for example, fallout shelters. By having enough *counterforce measures* with the ability to deliver and knock out enemy missiles before they are launched, a *credible first strike capability* is achieved. Kahn argument includes the initiation of a nuclear war in the case of a *desperate crisis* or *provocation* that would be morally acceptable. A *desperate crisis* is defined as "a circumstance in which, destabilizing as it would be, we would feel we would need an ability to rescue ourselves from a more dire eventuality by increasing our bargaining power or by actual use of the credible first strike capability" (Kahn 1960).

The argument of RAND for *credible first strike capability* is based on expert opinion of the acceptable nature of retaliatory blow by an enemy as demonstrated in Figure 1-1 in the form of an estimated duration in years for thermonuclear postwar economic recuperation. Kahn goes further to state "… Our calculations indicate that even without special stockpiling, dispersal, or protection, the restoration of our prewar gross national product should take place in a relatively short time – if we can hold the damage to the equivalent of something like 53 metropolitan areas destroyed." The results were based on the assumptions of "(1) favorable political environment (i.e., not losing the war), (2) immediate survival and patch-up, (3) maintenance of economic momentum, (4) specific bottlenecks alleviated, (5) "bourgeois" virtues survive, (6) workable postwar standards adopted, and (7) neglected effects uncertain assumptions that were arguably justified by Kahn (1960) and were set at levels that were described as more likely to be pessimistic than optimistic.

The analysis by RAND did not adequately deal with uncertainty and ignorance. It weighed heavily cognitive knowledge and expert opinion creating overconfidence in the results. Newman (1961) provided a review in *Scientific America* of Kahn's book in which he conjectured that the entire book was a staff joke in poor taste (Newman 1961, and Freeman 1969). The RAND study failed in properly assessing ignorance that places limits on human knowledge. Since the publication of *Thermonuclear War* (Kahn 1960), the phenomenon of electromagnetic pulse and potential climatological changes as a result of thermonuclear war were identified. These problems were not considered by RAND. The latter problem can result from the injection of millions of tons of dust and smoke in the upper atmosphere resulting in subfreezing land temperatures for months, and perhaps destroying human food resources such as crops. The effect of 100 to 10,000 total megatons of nuclear exchange could conceivably reduce the "population size of homosapians to prehistoric levels or below, and the extinction of human species itself cannot be excluded" (Science 1983). Another failure of the RAND study is in logic used to conduct reasoning under uncertainty. For example, Kahn arguably concludes that after a small nuclear destruction scenario of 53 metropolitan areas, we will *probably*

restore our gross national product (GNP) quickly. He argues that it is *likely* that we can handle radiation, it is *likely* that we can handle death, it is *likely* that we can handle destruction, therefore it is *likely* that we can handle jointly radiation, death and destruction. As a result he concludes that we will *probably* restore our GNP quickly. A fallacy of this logic in probabilistic reasoning is that high probabilistic likeliness of three propositions does not necessarily lead to a high probabilistic likeliness of their joint proposition. Uncertainty does not propagate in this simple manner as was used by Kahn. A proper treatment of uncertainty through assessment, modeling, propagation and integration is essential in conjecture.

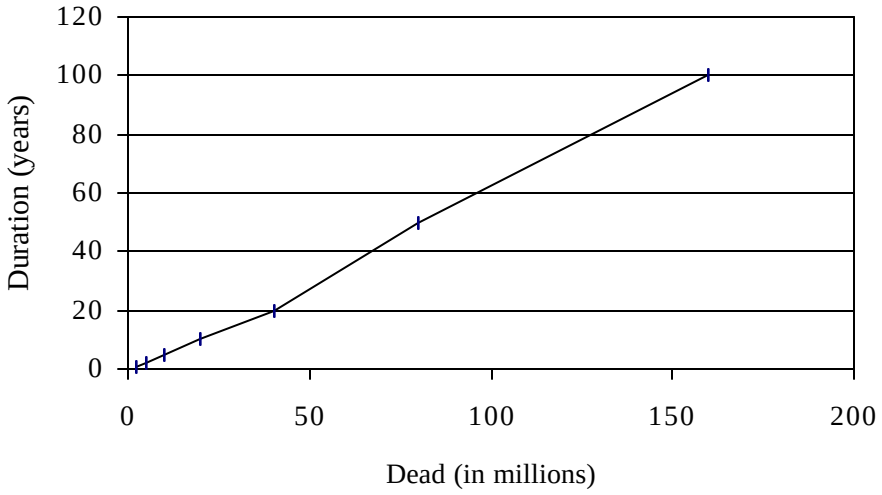

Figure 1-1. Estimated Duration for Thermonuclear Postwar Economic Recuperation

1.2.1. Delphi Method

The Delphi method is by far the most known method for eliciting and synthesizing expert opinions. The RAND corporation developed the Delphi method for the U. S. Air Force in the 1950s. In 1963, Helmer and Gordon used the Delphi method for a highly publicized long-range forecasting study (Helmer 1968). The method was extensively used in a wide variety of applications in the 1960s and 1970s exceeding 10,000 studies in 1974 on primarily technology forecasting and policy analysis (Linstone and Turoff 1975).

The purpose and steps of the Delphi method depend on the nature of use. Primarily the uses can be categorized into (1) technological forecasting, and (2) policy analysis. The technological forecasting relies on a group of experts on a subject matter of interest. The experts should be the most knowledgeable about issues or questions of concern. The issues and/or questions need to be stated by the study facilitators or analysts or a monitoring team, and high degree of consensus is sought from the experts. On the other hand, the policy analysis Delphi method seeks to incorporate the opinions and views of the entire spectrum of stakeholders, and seeks to communicate the spread of opinions to decision-makers. In engineering, we are generally interested in the former type of consensus opinion.

The basic Delphi method consists of the following steps (Helmer 1968):

1. Selection of issues or questions and development of questionnaires.
2. Selection of experts who are most knowledgeable about issues or questions of concern.
3. Issue familiarization of experts by providing sufficient details on the issues on the questionnaires.
4. Elicitation of experts about the issues. The experts might not know who the other respondents are.
5. Aggregation and presentation of results in the form of median values and an inter-quartile range (i.e., 25% and 75% percentile values).
6. Review of results by the experts and revision of initial answers by experts. This iterative reexamination of issues would sometimes increase the accuracy of results. Respondents who provide answers outside the inter-quartile range need to provide written justifications or arguments on the second cycle of completing the questionnaires.
7. Revision of results and re-review for another cycle. The process should be repeated until a complete consensus is achieved. Typically, the Delphi method requires two to four cycles or iterations.
8. A summary of the results is prepared with argument summary for out of inter-quartile range values.

The responses on the final iteration usually show less spread in comparison to spreads in earlier iterations. The median values are commonly taken as the best estimates for the issues or questions.

The Delphi method offers an adequate basis for expert-opinion elicitation, however, there is need to develop guidelines on its use to ensure consistency and result reliability. Chapter 2 provides a guide on conducting expert opinion elicitation.

Example 1-2. Helmer (1968) Delphi Questionnaire

This example provides a Delphi questionnaire as was originally developed and used by Helmer (1968). Table 1-1 shows the first part out of four parts of the questionnaire on technological innovations and use in the United States. These questions were also used in 1963 long range forecasting study by RAND, and in 1966 using 23 RAND employees as participants. The differences among the results of three studies range from 0 to 21 years with an average of six years.

Table 1-1. Delphi Questionnaire, Helmer (1968)

Questionnaire #1

- This is the first in a series of four questionnaires intended to demonstrate the use of the Delphi technique in obtaining reasoned opinions from a group of respondents.
- Each of the following six questions is concerned with developments in the United States within the next few decades.
- In addition to giving your answer to each question, you are also being asked to rank the questions from 1 to 7. Here "1" means that in comparing your own ability to answer this question with what you expect the ability of the other participants to be, you feel that you have the relatively best chance of coming closer to the truth than most of the others, while a "7" means that you regard that chance as relatively least.

Rank	Question	Answer*
☐	1. In your opinion, in what year will the median family income (in 1967 dollars) reach twice its present amount?	
☐	2. In what year will the percentage of electric automobiles among all automobiles in use reach 50%?	
☐	3. In what year will the percentage of households that are equipped with computer consoles tied to a central computer and data bank reach 50%?	
☐	4. By what year will the per-capita amount of personal cash transactions (in1967 dollars) be reduced to one-tenth of what it is now?	
☐	5. In what year will the power generation by thermonuclear fusion become commercially competitive with hydroelectric power?	
☐	6. By what year will it be possible by commercial carriers to get from New York to San Francisco in half the time that is now required to make that trip?	
☐	7. In what year will a man for the first time travel to the moon, stay at least one month, and return to earth?	

* "Never" is also an acceptable answer.

- Please also answer the following question, and give your name, (this is for identification purposes during the exercise only; no opinions will be attributed to a particular person).

Check one:
☐ I would like to participate in the three remaining questionnaires
☐ I am willing but not anxious to participate in the three remaining questionnaires
☐ I would prefer not to participate in the three remaining questionnaires

Name (block letters please): _____

Example 1-3. NASA's Challenger Space Shuttle Risk Study

The National Aeronautics and Space Administration (NASA) sponsored a study to assess the risks associated with the space shuttle (Colglazier and Weatherwax 1986, and Cooke 1991). In this study, an estimate of the solid rocker booster failure probability per launch, based on subjective probabilities and operating experience, was estimated to be about 1 in 35. The probability was based on Bayesian analysis utilizing prior experience of 32 confirmed failures from 1902 launches of various solid rocket

motors. This estimate was disregarded by NASA, and a number of 1 in 100,000 was dictated based on subjective judgments by managers and administrators (Colglazier and Weatherwax 1986, and Cooke 1991). The dictated number was not in agreement with published data (Bell and Esch 1989). The catastrophic Challenger explosion occurred on the twenty-fifth launch of a space shuttle on January 28, 1986.

Historically, NASA was distrustful of absolute reliability numbers for various reasons. It was publicized that the reliability numbers tend to be optimistic (pessimistic?), or taken as facts which they are not (Wiggins 1985). In reality, failure probabilities can be threatening to the survival to NASA's mission programs. For example, a General Electric qualitative probabilistic study on the probability of successfully landing a man on the moon was 95%. NASA felt that such numbers could do an irreparable harm, and efforts of this type should be disbanded (Bell and Esh 1989).

At the present, NASA is aggressively pursuing safety studies using probabilistic risk analysis of its various space missions. This change in NASA's practices can be attributed to the extensive investigations following the 1986 shuttle disaster.

The NASA has used risk assessment matrices to avoid the problem of managers treating the values of probability and risk as absolute judgements (Wiggins 1985). The Department of Defense offers the use of risk assessment matrices as a tool to prioritize risk (Defense Acquisition University 1998). Qualitatively, the likelihood of occurrence and consequences of an adverse scenarios may be described as shown in Tables 1-2 and 1-3, respectively. Levels of occurrence may be based on expert-opinion elicitation or actual probability data. The consequences described in Table 1-3 may be best determined using expert-opinion elicitation. Tables 1-2 and 1-3 can be combined to from the risk matrix. Risk assessment is based on the pairing of the likelihood of occurrence and consequences. Table 1-4 shows this pairing and is called a risk assessment matrix.

Table 1-2. Likelihood of Occurrence (Wiggins 1985)

Level	Description	Detailed Description
A	Frequent	Likely to occur frequently
B	Probable	Will occur several times in life of a system
C	Occasional	Likely to occur at sometime in life of a system
D	Remote	Unlikely but possible to occur in life of a system
E	Improbable	So unlikely that it can be assumed its occurrence may not be experienced

Table 1-3. Consequence (Wiggins 1985)

Level	Description	Mishap Definition
I	Catastrophic	Death or system loss
II	Critical	Severe injury, severe occupational illness, or major system damage
III	Marginal	Minor injury, minor occupational illness, or minor system damage
V	Negligible	Less than minor injury, occupational illness, or system damage

Table 1-4. Risk Assessment Matrix (Wiggins 1985)

Likelihood level	Consequence level			
	I Catastrophic	II Critical	III Marginal	IV Negligible
A: Frequent	1	3	7	13
B: Probable	2	5	9	16
C: Occasional	4	6	11	18
D: Remote	8	10	14	19
E: Improbable	12	15	17	20

Risk Index	Suggested Criteria
1-5	Unacceptable
6-9	Undesirable (project management decision required)
10-17	Acceptable with review by project management
18-20	Acceptable without review

1.2.2. Scenario Analysis

The development of scenario analysis can be attributed to Kahn and Wiener (1967). A scenario is defined as a hypothetical sequence of events that are constructed to focus attention on causal processes and decision points or nodes. Scenario analysis attempts to answer two questions: (1) how might some hypothetical situation come about, step by step, and (2) what alternatives or choices exist for each actor or party to the situation, at each step, for preventing, diverting, or facilitating the process. The first question is addressed in a similar manner to what is called event tree analysis as described by Ayyub and McCuen (1997). The second question is commonly handled nowadays using decision tree as described by Ayyub and McCuen (1997). Kahn and Wiener (1967) used scenario analysis to predict technological innovations for the year 2000. An examination of their top *likely* 25 technological innovations would reveal a success rate of about 40%. The predictions are based on 50% occurrence likelihood.

The scenario analysis by Kahn and Wiener (1967) did not use scenario probabilities, and relied on identifying what is termed the *surprise-free scenario* that is used as a basis for defining *alternative futures* or *canonical variations*. The alternative futures or canonical variations are generated by varying key parameters of the surprise-free scenario. Probabilities, that are absent from such an analysis, are arguably justified by Kahn and Wiener (1967) due to long-term projections making all

scenarios of small likelihood. The surprise-free scenario is considered important due to its ability in defining the long-term trend rather than its likelihood. Therefore, it is important to bear in mind this limitation of scenario analysis, its inability to deliver likelihood predictions to us but only long-term trend. At the present this limitation can be elevated by using event and decision tree analyses.

1.3. Objectives and Scope

Risk analysis and risk-based decision making for maintaining the integrity of the U. S. Army Corps of Engineers (USACE) facilities require the knowledge of two main quantities for components, and systems, their unsatisfactory-performance probability and consequences. Sometimes this information is not available from historical records, prediction methods or literature review. Also, sometimes there is a need to perform a preliminary risk evaluation of components within a system for the purpose of planning future reliability and risk analyses or for the purpose of performing initial screening of components. Also, in many situations classical frequency analysis cannot be used or is prohibitively expensive to quantify the existing risk or the change in risk related to USACE activities. The USACE is increasingly turning to using expert opinions in cases such as in (a) quantifying the probability of failure of navigation lock components, (b) quantifying the probability and consequence of navigation lock closure, (c) estimating the probability of events requiring emergency gate usage at hydropower plants, (d) quantifying the probability of failure and unsatisfactory performance of embankment dams, (e) quantifying the probability of failure and unsatisfactory performance of multi-purpose navigation, hydropower and recreational dams, and (f) predicting the vessel safety improvements from deep channel widening. Expert-opinion elicitation can provide the USACE with a means of gaining information on these essential risk-related quantities.

This guide describes an expert-opinion elicitation process of probabilities and consequences for Corps facilities. The process is designed for the use of Corps planners, economists and engineers. Historical and technical background materials, limitations, recommendations, and references are provided in this guide. Also, guidance on when and how to use expert-opinion elicitation is provided.

Chapter 1 provides background on ignorance, knowledge and uncertainty in modeling engineering systems. Chapter 2 describes a process for conducting expert-opinion elicitation. Chapter 3 has some concluding remarks. Cited references and a bibliography are provided at the end of the guide.

Several appendices are provided in the guide as background and additional materials. Appendix A contains background materials on failure probabilities, rates, and assessment methods. Appendix B provides background materials on failure consequences, and assessment methods of consequences. Appendix C provides additional materials on heuristics, elicitation, scoring and aggregation for expert opinions.

2. The Expert-Opinion Elicitation Process

2.1. Introduction and Terminology

2.1.1. Theoretical Bases

Expert-opinion elicitation can be defined as a heuristic process of gathering informing and data or answering questions on issues or problems of concern. In this study, the focus is on occurrence probabilities and consequences of events related to civil works for the use of USACE planners, engineers, and others should they choose to use expert judgment. For this purpose, the expert-opinion elicitation process can be defined as a formal process of obtaining information or answers to specific questions about certain quantities, called issues, such as unsatisfactory-performance rates, unsatisfactory-performance consequences and expected service life. Expert-opinion elicitation should not be used in lieu of rigorous reliability and risk analytical methods, but should be used to supplement them and to prepare for them. The suggested expert-opinion elicitation process in this chapter is a variation of the Delphi technique (Helmer 1968) scenario analysis (Kahn and Wiener 1967) based on uncertainty models (Ayyub, 1999, Ayyub 1991, 1992 and 1993, Haldar et al 1997, Ayyub et al 1997, Ayyub and Gupta 1997, Ayyub 1998, Cooke 1991), social research (Bailey 1994), USACE studies (Ayyub et al 1996, and Baecher 1998), ignorance, knowledge, information and uncertainty of Chapter 1, nuclear industry recommendations (NRC 1997), Stanford Research Institute protocol (Spetzler and Stael von Holstein 1975).sentence is not clear

2.1.2. Terminology

The terminology of Table 2-1 is needed for defining the expert-opinion elicitation process, in addition to other related definitions of Appendices A and B for probabilities and consequences, respectively.

Table 2-1. Terminology and Definitions

Term	Definition
Evaluators	Evaluators consider available data, become familiar with the views of proponents and other evaluators, question the technical bases of data, and challenge the views of proponents.
Expert	A person with related or unique experience to an issue or question of interest for the process.
Expert-opinion elicitation (EE) process	A formal, heuristic process of gathering informing and data or answering questions on issues or problems of concern.
Leader of EE process	An entity having managerial and technical responsibility for organizing and executing the project, overseeing all participants, and intellectually *owning* the results.
Observers	Observers can contribute to the discussion, but cannot provide expert opinion that enters in the aggregated opinion of the experts.
Peer reviewers	Experts that can provide an unbiased assessment and critical review of an expert-opinion elicitation process, its technical issues, and results.
Proponents	Proponents are experts who advocate a particular hypothesis or technical position. In science, a proponent evaluates experimental data and professionally offers a hypothesis that would be challenges by the proponent's peers until proven correct or wrong.
Resource experts	Resource experts are technical experts with detailed and deep knowledge of particular data, issue aspects, particular methodologies, or use of evaluators.
Sponsor of EE process	An entity that provides financial support and *owns* the rights to the results of the EE process. Ownership is in the sense of property ownership.
Subject	A person who might be affected or might affect an issue or question of interest for the process.
Technical facilitator (TF)	An entity responsible for structuring and facilitating the discussions and interactions of experts in the EE process; staging effective interactions among experts; ensuring equity in presented views; eliciting formal evaluations from each expert; and creating conditions for direct, non-controversial integration of expert opinions.
Technical integrator (TI)	An entity responsible for developing the composite representation of issues based on informed members and/or sources of related technical communities and experts; explaining and defending composite results to experts and outside experts, peer reviewers, regulators, and policy makers; and obtaining feedback and revising composite results.
Technical integrator and facilitator (TIF)	An entity responsible for both functions of TI and TF.

2.1.3. Classification of Issues, Study Levels, Experts, and Process Outcomes

The NRC (1997) classified issues for expert-opinion elicitation purposes into three complexity degrees (A, B, or C), with four level of study in the expert-opinion elicitation process (I, II, III, and IV) as shown in Table 2-1. A given issue is assigned a complexity degree and a level of study that depend on (1) the significance of the issue to the final goal of the study, (2) the issue's technical complexity and uncertainty level, (3) the amount of non-technical contention about the issue in the technical community, and (4) important non-technical consideration such as budgetary, regulatory, scheduling, public perception, or other concerns.

Experts can be classified into five types (NRC 1997): (1) proponents, (2) evaluators, (3) resource experts, (4) observers, and (5) peer reviewers. A proponent is an expert who advocates a particular hypothesis or technical position. In science, a proponent evaluates experimental data and professionally offers a hypothesis that would be challenges by the proponent's peers until proven correct or wrong. An evaluator is an expert who has the role of evaluating the relative credibility and plausibility of multiple hypotheses to explain observations. Evaluators consider available data, become familiar with the views of proponents and other evaluators, questions the technical bases of data, and challenges the views of proponents. A resource expert is a technical expert with detailed and deep knowledge of particular data, issue aspects, particular methodologies, or use of evaluators. An observer can contribute to the discussion, but cannot provide expert opinion that enters in the aggregated opinion of the experts. A peer reviewer is an expert that can provide an unbiased assessment and critical review of an expert-opinion elicitation process, its technical issues, and results.

The study level as shown in Table 2-1 involves a technical integrator (TI) or a technical integrator and facilitator (TIF). A TI can be one person or a team (i.e., an entity) that is responsible for developing the composite representation of issues based on informed members and/or sources of related technical communities and experts; explaining and defending composite results to experts and outside experts, peer reviewers, regulators, and policy makers; and obtaining feedback and revising composite results. A TIF can be one person or a team (i.e., an entity) that is responsible for the functions of a TI, and structuring and facilitating the discussions and interactions of experts in the EE process; staging effective interactions among experts; ensuring equity in presented views; eliciting formal evaluations from each expert; and creating conditions for direct, non-controversial integration of expert opinions. The primary difference between the TI and the TIF is in the intellectual responsibility for the study where it lies with only the TI, and the TIF and the experts, respectively. The TIF has also the added responsibility of maintaining the professional integrity of the process and its implementation.

The TI and TIF processes are required to utilize peer reviewers for quality assurance purposes. Peer review can be classified according to peer-review method, and according to peer-review subject. Two methods of peer review can be performed: (1) participatory peer review that would be conducted as an ongoing review throughout all study stages, and (2) late-stage peer review that would be performed as the final stage of the study. The former method allows for affecting the course of the study, whereas the latter one might not be able to affect the study without a substantial rework of the study. The second classification of peer reviews is by peer-review subject and has two types: (1) technical peer review that

focuses on the technical scope, coverage, contents and results, and (2) process peer review that focuses on the structure, format and execution of the expert-opinion elicitation process. A guidance on the use of peer reviewers is provided in Table 2-2 (NRC 1997).

The expert-opinion elicitation process should preferably be conducted to include a face-to-face meeting of experts that is developed specifically for the issues under consideration. The meeting of the experts should be conducted after communicating to the experts in advance to the meeting background information, objectives, list of issues, and anticipated outcome from the meeting. The expert-opinion elicitation based on the technical integrator and facilitator (IIF) concept can result in consensus or disagreement as shown in Figure 2-1. Consensus can be of four types as shown in Figure 2-1 (NRC 1997). Commonly, the expert-opinion elicitation process has the objective of achieving consensus type 4, i.e., experts agree that a particular probability distribution represents the overall scientific community. The TIF plays a major role in building consensus by acting as a facilitator. Disagreement among experts, whether it is intentional or unintentional, requires the TIF to act as an integrator by using equal or non-equal weight factors. Sometimes, expert opinions need to be weighed for appropriateness and relevance rather than strictly weighted by factors in a mathematical aggregation procedure.

Table 2-2. Issue Degrees and Study Levels (Constructed based on NRC 1997)

Issue Complexity Degree		Study Level	
Degree	**Description**	**Level**	**Requirements**
A	Non-controversial Insignificant effect on risk	I	A technical integrator (TI) evaluates and weighs models based on literature review and experience, and estimates needed quantities.
B	Significant uncertainty Significant diversity Controversial Complex	II	technical integrator (TI) interacts with proponents & resource experts, asses interpretations, and estimates needed quantities.
C	Highly contentious Significant effect on risk Highly complex	III	technical integrator (TI) brings together proponents & resource experts for debate and interaction. TI focuses the debate, evaluates interpretations, and estimates needed quantities.
		IV	technical integrator (TI) and technical facilitator (TF) (that can be one entity, i.e., ITF) organize a panel of experts to interpret and evaluate, focus discussions, keep the experts debate orderly, summarize and integrate opinions, and estimates needed quantities.

Table 2-3. Guidance on Use of Peer Reviewers (NRC 1997)

Expert-opinion elicitation Process	Peer Review Subject	Peer Review Method	Recommendation
Technical integrator and facilitator	Technical	Participatory	Recommended
		Late stage	Can be acceptable
	Process	Participatory	Strongly recommended
		Late stage	Risky: unlikely to be successful
Technical integrator	Technical	Participatory	Strongly recommended
		Late stage	Risky but can be acceptable
	Process	Participatory	Strongly recommended
		Late stage	Risky but can be acceptable

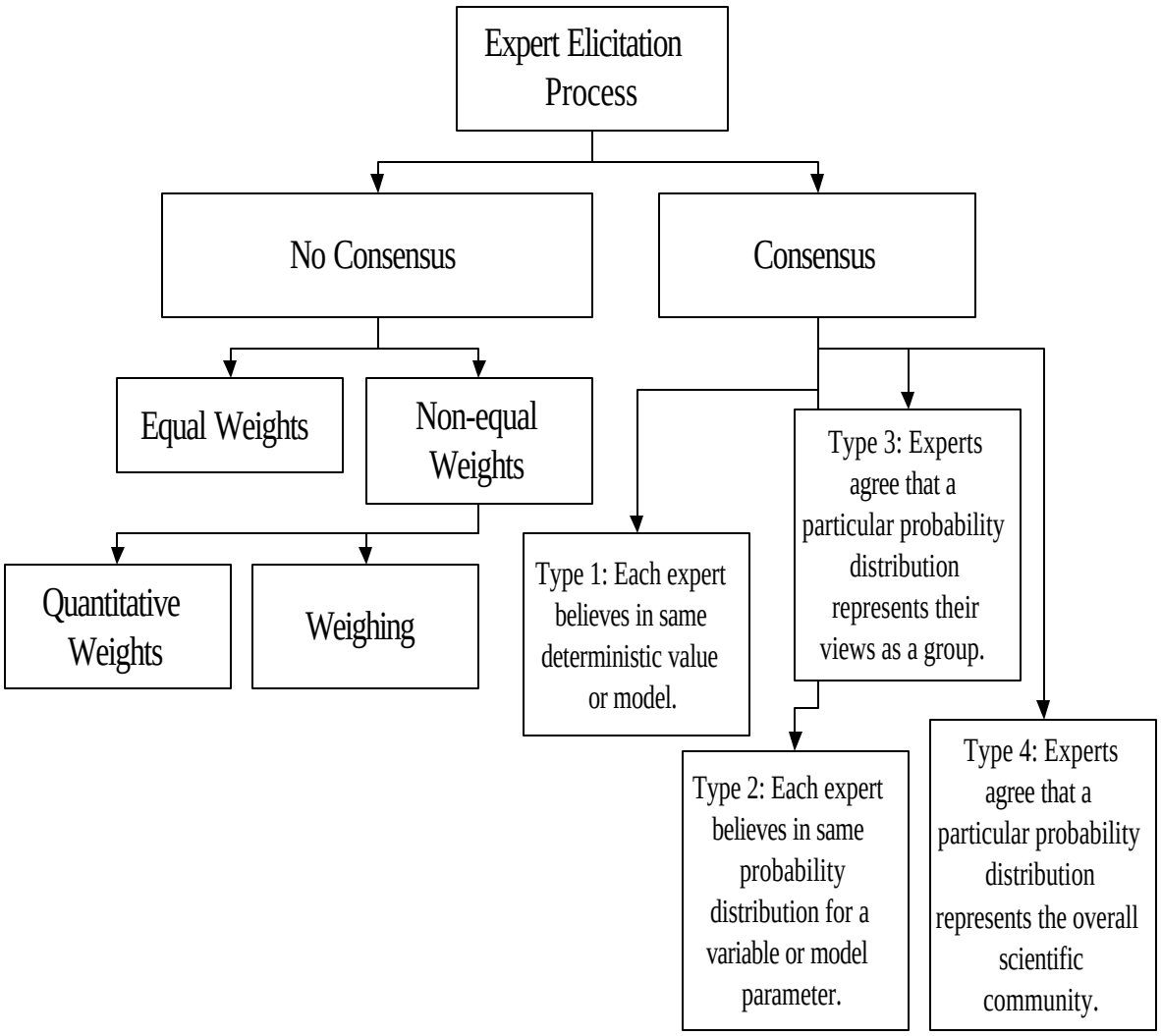

Figure 2-1. Outcomes of the Expert-Opinion Elicitation Process

2.2. Process Definition

Expert-opinion elicitation was defined as a formal, heuristic process of obtaining information or answers to specific questions about certain quantities, called issues, such as unsatisfactory-performance rates, unsatisfactory-performance consequences and expected service lives. The suggested steps for an expert-opinion elicitation process depend on the use of a technical integrator (TI) or a technical integrator and facilitator (TIF) as shown in Figure 2-2. Figure 2-2 was constructed based on NRC (1997), supplemented with details, and added steps. The details of the steps involved in these two processes are defined in subsequent subsections.

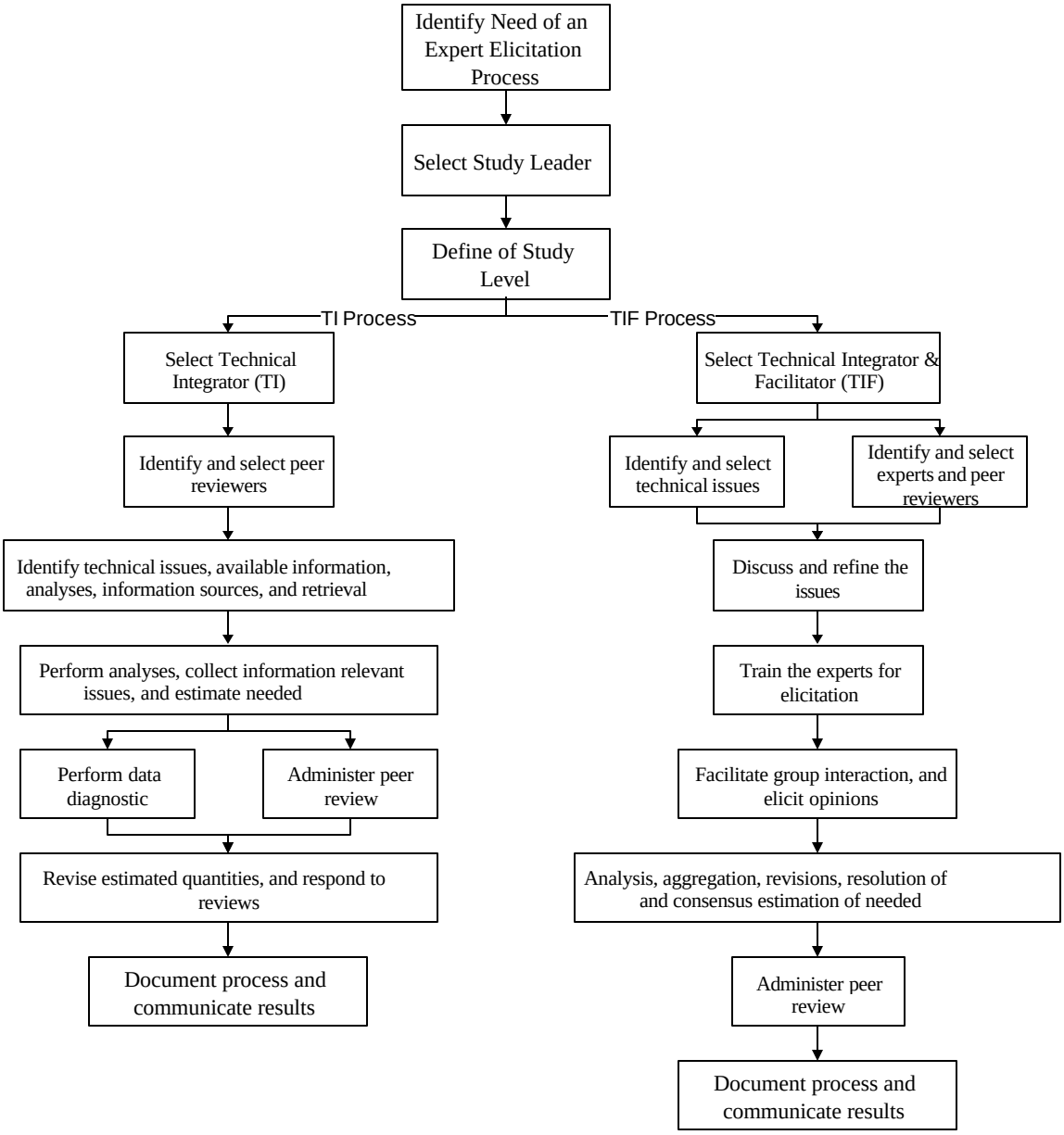

Figure 2-2. Expert-Opinion Elicitation Process

2.2.1. Need Identification for Expert-Opinion Elicitation

The primary reason for using expert-opinion elicitation is to deal with uncertainty in selected technical issues related to a system of interest. Issues with significant uncertainty, issues that are controversial and/or contentious, issues that are complex, and/or issues that can have a significant effect on risk are most suited for expert-opinion elicitation. The value of the expert-opinion elicitation comes from its initial intended uses as a heuristic tool, not a scientific tool, for exploring vague and unknown issues that are otherwise inaccessible. It is not a substitute to scientific, rigorous research.

The identification of need and its communication to experts are essential for the success of the expert-opinion elicitation process. The need identification and communication should include the definition of the goal of the study and relevance of issues to this goal. Establishing this relevance would make the experts stake holders and thereby increase their attention and sincerity levels. Relevance of each issues and/or question to the study needs to be established. This question-to-study relevance is essential to enhancing the reliability of collected data from the experts. Each question or issue needs to be relevant to each expert especially when dealing with subjects with diverse views and backgrounds.

2.2.2. Selection of Study Level and Study Leader

The goal of a study and nature of issues determine the study level as shown in Table 2-1. The study leader can be either a technical integrator (TI), technical facilitator (TF), or a combined technical integrator and facilitator (TIF). The leader of the study is an entity having managerial and technical responsibility for organizing and executing the project, overseeing all participants, and intellectually *owning* the results. The primary difference between the TI and the TIF is in the intellectual responsibility for the study where it lies with only the TI, and the TIF and the experts, respectively. The TIF has also the added responsibility of maintaining the professional integrity of the process and its implementation. The TI is required to utilize peer reviewers for quality assurance purposes. A study leader should be selected based on the following attributes:

1. an outstanding professional reputation, and wide recognition and competence based on academic training and relevant experience;
2. strong communication skills, interpersonal skills, flexibility, impartiality, and ability to generalize and simplify;
3. a large contact base of industry leaders, researcher, engineers, scientists, and decision makers; and
4. ability to build consensus, and leadership qualities.

The study leader does not need to be a subject expert, but should be knowledgeable of the subject matter.

2.2.3. Selection of Peer Reviewers and Experts

2.2.3.1. Selection of Peer Reviewers

Peer review can be classified according to peer-review method, and according to peer-review subject. Two methods of peer review can be performed: (1) participatory peer review that would be conducted as an ongoing review throughout all study stages, and (2) late-stage peer review that would be performed as the final stage of the study. The second classification of peer reviews is by peer-review subject and has two types: (1) technical peer review that focuses on the technical scope, coverage, contents and results, and (2) process peer review that focuses on the structure, format and execution of the expert-opinion elicitation process. These classifications were previously discussed.

Peer reviewers are needed for both the TI and TIF processes. The peer reviewers should be selected by the study leader in close consultation with perhaps the study sponsor. The following individuals should be sought after in peer reviewers:

1. Researchers, scientists, and/or engineers that have outstanding professional reputation, and widely recognized competence based on academic training and relevant experience.
2. Researchers, scientists, and/or engineers with general understanding of the issues in other related areas, and/or with relevant expertise and experiences from other areas.
3. Researchers, scientists, and/or engineers who are available and willing to devote the needed time and effort.
4. Researchers, scientists, and/or engineers with strong communication skills, interpersonal skills, flexibility, impartiality, and ability to generalize and simplify.

2.2.3.2. Identification and Selection of Experts

The size of an expert panel should be determined on case by case basis. The size should be large enough to achieve a needed diversity of opinion, credibility, and result reliability. In recent expert-opinion elicitation studies, a nomination process was used to establish a list of candidate experts by consulting archival literature, technical societies, governmental organization, and other knowledgeable experts (Trauth et al 1993). Formal nomination and selection processes should establish appropriate criteria for nomination, selection and removal of experts. For example, the following criteria were used in an ongoing Yucca Mountain seismic hazard analysis (NRC 1997) to select experts:

1. Strong relevant expertise through academic training, professional accomplishment and experiences, and peer-reviewed publications;
2. Familiarity and knowledge of various aspects related to the issues of interest;
3. Willingness to acts as proponents or impartial evaluators;
4. Availability and willingness to commit needed time and effort;
5. Specific related knowledge and expertise of the issues of interest;
6. Willingness to effectively participate in needed debates, to prepare for discussions, and provide needed evaluations and interpretations; and
7. Strong communication skills, interpersonal skills, flexibility, impartiality, and ability to generalize and simplify.

In this NRC study, criteria were set for expert removal that include failure to perform according to commitments and demands as set in the selection criteria, and unwillingness to interact with members of the study.

The panel of experts for an expert-opinion elicitation process should have a balance and broad spectrum of viewpoints, expertise, technical points of view, and organizational representation. The diversity and completeness of the panel of experts is essential for the success of the elicitation process. For example, it can include the following:

1. Proponents who advocate a particular hypothesis or technical position;
2. Evaluators who consider available data, become familiar with the views of proponents and other evaluators, questions the technical bases of data, and challenges the views of proponents; and
3. Resource experts who are technical experts with detailed and deep knowledge of particular data, issue aspects, particular methodologies, or use of evaluators.

The experts should be familiar with the design, construction, operational, inspection, maintenance, reliability and engineering aspects of the equipment and components of a facility of interest. It is essential to select people with basic engineering or technological knowledge, however they do not necessarily need to be engineers. It might be necessary to include one or two experts from management with engineering knowledge of the equipment and components, consequences, safety aspects, administrative and logistic aspects of operation, expert-opinion elicitation process, and objectives of this study. One or two experts with a broader knowledge of the equipment and components might be needed. Also, one or two experts with a background in risk analysis and risk-based decision making and their uses in areas related to the facility of interest might be needed.

Observers can be invited to participate in the elicitation process. Observers can contribute to the discussion, but cannot provide expert opinion that enters in the aggregated opinion of the experts. The observers provide expertise in the elicitation process, probabilistic and statistical analyses, risk analysis and other support areas. The composition and contribution of the observers are essential for the success of this process. The observers may include the following:

1. Individuals with research or administrative-related background from research laboratories or headquarters of the US Army Corps of Engineers with engineering knowledge of equipment and components of Corps facilities.
2. Individuals with expertise in probabilistic analysis, probabilistic computations, consequence computations and assessment, and expert-opinion elicitation.

A list of names with biographical statements of the study leader, technical integrator, technical facilitator, experts, observers, and peer reviewers should be developed and documented. All attendees can participate in the discussions during the meeting. However, only the experts can provide the needed answers to questions on the selected issues. The integrators and facilitators are responsible for conducting the expert-opinion elicitation process. They can be considered to be a part of the observers or experts depending on the circumstances and the needs of the process.

2.2.3.3. Items to be Sent to Experts and Reviewers Before the Expert-Opinion Elicitation Meeting

The experts and observers need to receive the following items before the expert-opinion elicitation meeting:

1. An objective statement of the study;
2. A list of experts, observers, integrators, facilitators, study leader, sponsors, and their biographical statements;
3. A description of the facility, systems, equipment and components;
4. Basic terminology, definitions that should include probability, unsatisfactory-performance rate, average time between unsatisfactory performances, mean (or average) value, median value, and uncertainty;
5. Unsatisfactory-performance consequence estimation;
6. A description of the expert-opinion elicitation process;
7. A related example on the expert-opinion elicitation process and its results, if available;
8. Aggregation methods of expert opinions such as computations of percentiles;
9. A description of the issues in the form of a list of questions with background descriptions. Each issue should be presented on a separate page with spaces for recording an expert's judgment, any revisions and comments. Clear statements of expectations from the experts in terms of time, effort, responses, communication, and discussion style and format.

It might be necessary to personally contact individual experts for the purpose of establishing clear understanding of expectations.

2.2.4. Identification, Selection and Development of Technical Issues

The technical issues of interest should be carefully selected to achieve certain objectives. In these guidelines, the technical issues are related to the quantitative assessment of unsatisfactory-performance probabilities and consequences for selected components, subsystems and systems within a facility. The issues should be selected such that they would have a significant impact on the study results. These issues should be structured in a logical sequence starting by background statement, followed by questions, and then answer selections or answer format and scales. Personnel with risk-analysis background that are familiar with the construction, design, operation, and maintenance of the facility need to define these issues in the form of specific questions. Also, background materials about these issues need to be assembled. The materials will be used to familiarize and train the experts about the issues of interest as described subsequent steps.

An introductory statement for the expert-opinion elicitation process should be developed that includes the goal of the study and establishes relevance. Instructions should be provided with guidance on expectations, answering the questions, and reporting. The following are guidelines on constructing questions and issues based social research practices (Bailey 1994):

1. Each issue can include several questions, however, each question should consist of only one sought after answer. It is a poor practice to include two questions in one.
2. Question and issue statements should not be ambiguous. Also, the use of ambiguous words should be avoided. In expert-opinion elicitation of failure probabilities, the word "failure" might be vague or ambiguous to some subjects. Special attention should be given to its definition within the context of each issue or question. The level of wording should be kept to a minimum. Also, the choice of the words might affect the connotation of an issue especially by different subjects.
3. The use of factual questions is preferred over abstract questions. Questions that refer to concrete and specific matters result in desirable concrete and specific answers.
4. Questions should be carefully structured in order to reduce biases of subjects. Questions should be asked in a neutral format, sometimes more appropriately without lead statements.
5. Sensitive topics might require stating questions with lead statements that would establish supposedly accepted social norms in order to encourage subjects to answers the questions truthfully.

Questions can be classified into *open-ended questions* and *closed-ended questions* as was previously discussed. The format of the question should be selected carefully. The format, scale and units for the response categories should be selected to best achieve the goal of the study. The minimum number of questions and question order should be selected using practices and methods of educational and psychological testing and social research as provided in Appendix C.

Once the issues are developed, they should be pretested by administering them a few subjects for the purpose of identifying and correcting flaws. The results of this pretesting should be used to revise the issues.

2.2.5. Elicitation of Opinions

The elicitation process of opinions should be systematic for all the issues according to the steps presented in this section.

2.2.5.1. Issue Familiarization of Experts

The background materials that were assembled in the previous step should be sent to the experts about one to two weeks in advance of the meeting with the objective of providing sufficient time for them to become familiar with the issues. The objective of this step is, also, to ensure that there is a common understanding among the experts of the issues. The background material should include the objectives of the study, description of the issues and lists of questions for the issues, description of systems and processes, their equipment and components, the elicitation process, selection methods of experts, and biographical information on the selected experts. Also, example results and their meaning, methods of analysis of the results, and lessons learned from previous elicitation processes should be made available to them. It is important to breakdown the questions or issues in components that can be easily addressed. Preliminary discussion meetings or telephone conversations between the facilitator and experts might be necessary in some cases in preparation for the elicitation process.

2.2.5.2. Training of Experts

This step is performed during the meeting of the experts, observers and facilitators. During the training the facilitator needs to maintain flexibility to refine wording or even change approach based on feedback from experts. For instance, experts may not be comfortable with "probability" but they may answer on year" or "recurrence interval." Additional information on the indirect elicitation is provided in Appendix C. The meeting should be started with presentations of background material to establish relevance of the study to the experts, and study goals in order to establish rapport with the experts. Then, information on uncertainty sources and types, occurrence probabilities and consequences, expert-opinion elicitation process, technical issues and questions, aggregation of expert opinions should be presented. Also, experts need to be trained on providing answers in an acceptable format that can be used in the analytical evaluation of the unsatisfactory-performance probabilities or consequences. The experts need to be trained in certain areas such as the meaning of probability, central tendency, and dispersion measures especially to experts who are not familiar with the language of probability. Additional training might be needed on consequences, subjective assessment, logic trees, problem structuring tools such as influence diagrams, and methods of combining expert evaluations. Sources of bias that include overconfidence, and base-rate fallacy and their contribution to bias and error should be discussed. This step should include a search for any motivational bias of experts due to, for example, previous positions experts have taken in public, wanting to influence decisions and funding allocations, preconceived notions that they will be evaluated by their superiors as a result of their answers, and/or to be perceived as an authoritative expert. These motivational biases, once identified, can be sometimes overcome by redefining the incentive structure for the experts.

2.2.5.3. Elicitation and Collection of Opinions

The opinion elicitation step starts with a technical presentation of an issue, and by decomposing the issue to its components, discussing potential influences, and describing event sequences that might lead to top events of interest. These top events are the basis for questions related to the issue in the next stage of the opinion elicitation step. Factors, limitations, test results, analytical models, and uncertainty types and sources need to be presented. The presentation should allow for questions to eliminate any ambiguity and clarify scope and conditions for the issue. The discussion of the issue should be encouraged. The discussion and questions might result in refining the definition of the issue. Then, a form with a statement of the issue should be given to the expert to record their evaluation or input. The experts' judgment along with their supportive reasoning should be documented about the issue. It is common that experts would be asked to provide several conditional probabilities in order to reduce the complexity of the questions and thereby obtain reliable answers. These conditional probabilities can be based on fault tree and event tree diagrams. Conditioning has the benefit of simplifying the questions by decomposing the problems. Also, it results in a conditional event that has a larger occurrence probability than its underlying events; therefore making the elicitation less prone to biases since experts tend to have a better handle on larger probabilities in comparison to very small ones. It is desirable to have the elicited probabilities in the range of 0.1 to 0.9 if possible. Sometimes it might be desirable to elicit conditional probabilities using linguistic terms as shown in Table 2-1. If correlation among variables exits, it should be presented to the experts in great detail and conditional probabilities need to be elicited.

Issues should be dealt with one issue at a time, although sometimes similar or related issues might be considered simultaneously.

2.2.5.4. Aggregation and Presentation of Results

The collected assessments from the experts for an issue should be assessed for internal consistency, analyzed and aggregated to obtain composite judgments for the issue. The means, medians, percentile values and standard deviations need to be computed for the issues. Also, a summary of the reasoning provided during the meeting about the issues needs to be developed. Uncertainty levels in the assessments should also be quantified. A summary of methods for combining expert opinions was provided in Appendix C. The methods can be classified into consensus methods and mathematical methods. The mathematical methods can be based on assigning equal weights to the experts or different weights.

2.2.5.5. Group Interaction, Discussion and Revision by Experts

The aggregated results need to be presented to the experts for a second round of discussion and revision. The experts should be given the opportunity to revise their assessments of the individual issues at the end of discussion. Also, the experts should be asked to state the rationale for their statements and revisions. The revised assessments of the experts need to be collected for aggregation and analysis. This step can produce either consensus or no consensus as shown in Figure 2-1. The selected aggregation procedure might require eliciting weight factors from the experts. In this step the technical facilitator plays a major role in developing a consensus, and maintaining the integrity and credibility of the elicitation process. Also, the technical integrator is needed to aggregate the results without biases with reliability measures. The integrator might need to deal with varying expertise levels for the experts, outliers (i.e., extreme views), non-independent experts, and expert biases.

2.2.6. Documentation and Communication

A comprehensive documentation of the process is essential in order to ensure acceptance and credibility of the results. The document should include complete descriptions of the steps, the initial results, revised results, consensus results, and aggregated results spreads and reliability measures.

2.3. Example Expert-Opinion Elicitation Processes with Results

2.3.1. Cargo Elevators Onboard Ships

This example illustrates the use of expert-opinion elicitation to obtain unsatisfactory-performance probabilities needed to study the safety of cargo elevators onboard naval ships (Ayyub 1992). In order to study the safety of the elevators and the effect of add-on safety features, a fault tree analysis was performed. The fault tree analysis requires the knowledge of unsatisfactory-performance probabilities of basic events, such as the unsatisfactory performance of mechanical or electrical components and human errors.

Generally, the unsatisfactory-performance probabilities can be obtained from several sources, such as unsatisfactory-performance records, unsatisfactory-performance databases, literature review, or industry-based reports and documents. However, in some cases these sources do not contain the needed probabilities for some basic events. In such cases, expert-opinion elicitation can be used to obtain the needed information. For example, the unsatisfactory-performance rate of the hoisting machinery brake was obtained from unsatisfactory-performance records, and the probability that a passerby falls into an open elevator trunk (human error) required expert-opinion elicitation.

In the elevator safety study, about 250 issues were identified for the expert-opinion elicitation process. The issues were presented to the experts with the needed background information over a three-day period. All the issues were discussed and addressed in this time period.

This section provides examples issues and results of expert-opinion elicitation. Since the background information on the types of elevators, their use and limitation are not provided in this section, the reported results herein can be considered to be hypothetical and should not be used for other purposes.

Two example issues are described in this section. The issues are:

1. How often does the load on a platform shift as a result of being poorly stacked?
2. During one loading revolution at one deck level, what is the probability that a fork truck driver will place the load such that it overhangs the edge of the platform?

Eight experts were used in the expert-opinion elicitation process. The results of the process were summarized in the form of percentiles. The percentiles were computed using the equations in Table 2-2. Tables 2-3 and 2-4 were used to summarize the results of the expert-opinion elicitation for issues 1 and 2, respectively. It can be noted from the tables that the results are expressed as the number of unsatisfactory performances per year and a percent for issues 1 and 2, respectively. These results were used to compute the needed probabilities in the fault tree analysis. It is desirable in expert-opinion elicitation to state the issues in the most suitable form and units in order to obtain the best results from the experts.

Table 2-4. Expert-opinion elicitation for Example Issue 1 (Ayyub 1992, and Ayyub et al 1996)

Event Name	Full Description	Expert-opinion elicitation (8 experts)				Summary
		First Response	Median	Second Response	Median	
Load is poorly stacked.	The load on the platform is stacked in such a manner that it is shifted by normal starting and stopping of the platform. Assume that the ship is in calm sea state. <u>Issue:</u> On one elevator, how often does the load on the platform shift as a result of being poorly stacked?	<u>Issue:</u> 1 in 1 yr 1 in 1 yr 1 in 0.5 yr 1 in 2 yrs 1 in 0.1 yr 1 in 1 yr 1 in 0.1 yr 1 in 15 yr	1 in 1 yr	<u>Issue:</u> 1 in 1 yr 1 in 1 yr 1 in 0.5 yr 1 in 1 yr 1 in 0.5 yr 1 in 1 yr 1 in 0.5 yr 1 in 1 yr	1 in 1 yr	<u>Low</u> 1 in 1 year <u>25 percentile</u> 1 in 1 year <u>Median</u> 1 in 1 year <u>75 percentile</u> 1 in 0.5 year <u>High</u> 1 in 0.5 year

Table 2-5. Expert-opinion elicitation for Example Issue 2 (Ayyub 1992, and Ayyub et al 1996)

Event Name	Full Description	Expert-opinion elicitation (8 experts)				Summary
		First Response	Median	Second Response	Median	
Fork truck driver places load over-hanging platform.	Fork truck driver places load such that it overhangs platform despite the existence of adequate lighting. Assume that there are no yellow margins painted on the platform. <u>Issue:</u> During one loading evolution at one deck level, what is the probability that a fork truck driver will place the load such that it overhangs the edge of the platform?	<u>Issue:</u> 1% 1% 10% 0.1% 0.5% 1% 0.5% 0.5%	0.75%	<u>Issue:</u> 1% 1% 10% 1% 0.5% 1% 0.5% 0.5%	1%	<u>Low</u> 0.5% <u>25 percentile</u> 0.5% <u>Median</u> 1% <u>75 percentile</u> 1% <u>High</u> 10%

2.3.2. Navigation Locks

Detailed descriptions of technical issues are essential for the success of an expert-opinion elicitation process, and need to be provided to the experts. The descriptions should provide the experts of background materials, clear statements of issues, objectives, format, opinion aggregation that would be used in elicitation sessions. In this example, a description of a navigation lock and fault scenarios are presented for demonstration purposes. The equipment and components are based on the Emsworth navigation lock on the Ohio River. The background materials were used to develop technical issues (Ayyub et al 1996).

A navigation lock can be considered to constitute a system that consists of equipment, each equipment consists of components that consist of elements. The equipment, components and elements are called levels of analysis. In estimating unsatisfactory-performance likelihood and consequences, decisions are needed on the level of computation for the equipment in the process, i.e., equipment, component or element level. The decision can be based on the availability of information, the logistic of inspection that might define the entity or unit, the objectives of risk analyses that will be performed on the lock or other considerations. Accordingly, the level of computation does not need to be the same for all equipment within the process.

General Description

The operation of the lock is shown in the form of a logic diagram in Figures 2-3a and 2-3b (Ayyub et al 1996).

Two adjacent, parallel lock chambers are located along the right bank of the main channel. The large lock chamber occupies the landward position and has clear dimensions of 110 feet x 600 feet. The smaller river chamber measures 56 feet x 360 feet. Normal lift is 18 feet. The lock walls and sills are the gravity type and founded on rock. Both the upper and lower guide and guard walls are concrete gravity sections but the upper and lower guard walls have been extended using steel sheet pile cells. The filling and emptying of the lock chambers is accomplished through ports in the middle and river walls. The large chamber is filled by 16 cylindrical valves located in the upper end of the middle wall and emptied by 16 similar valves which pass the water through a culvert under the smaller chamber and into the river below the dam. A supplemental filling system was instituted during a recent major rehabilitation and involved the reopening of a 10-foot diameter turbine tunnel, providing of a slide gate, plugging of the tailrace exit, and the cutting of filling ports through the land wall at lock floor level. The small chamber uses only six filling and six emptying valves in the river wall. The lock gates are of the mitering type, hinged to embedded anchorages at the top and supported at the bottom on steel pintles. Each leaf is a rectangular frame with vertical girders at each end, and vertical beams and horizontal intercoastals on the gate leaves for the 110-foot chamber, or horizontal beams and vertical intercoastals on the leaves for the 56-foot chamber. Upstream closure of the large chamber is accomplished using trestles stored underwater that are raised from notches in a concrete sill upstream of the miter gates and then fitted with bulkheads. The small chamber uses a coffer beam and needle type closure. Downstream closure for both chambers is accomplished with poiree dams. The average number of annual lockages has remained fairly constant over the last 30 years at about 9950, with commercial lockages decreasing and recreational lockages increasing in recent years.

Description of Components

The Emsworth navigation lock on the Ohio River as a system consists of gates, dam, walls, channel, equipment, and users. The following are descriptions of its components:

1. Filling and Emptying Valves: The filling and emptying of the lock chambers are accomplished through culverts placed in the middle and river walls. The main lock is filled by 16 cylindrical valves located in the upper end of the middle wall and emptied by 16 similar valves which pass the water through the lower end of the wall and under the riverward chamber into the river below the dam.

2. Filling and Emptying Equipment: The hydraulic system consists of three constant delivery oil pumps and one pressure holding oil pump, located on the first floor in the operation building on the land wall. The pumps supply oil under pressure to the hydraulic piping system for operation of the lock gate and culvert valve operating machinery on the lock walls. This system was installed in 1968 and replaced the original compressed air system for operation of the miter gates and the original hydraulic system installed for operation of the emptying and filling valves.

3. Lock Wall: The lock walls are the gravity type founded on rock. Width of wall at the top is 5 feet minimum and 24 feet maximum. The sills are concrete gravity sections and anchor rods installed where computations indicated their need.

4. Guide Wall: The upper guide wall is 1,023.19 feet long measured from the upstream nose of the middle wall, and the lower guide wall is 650.0 feet long measured from the downstream nose of the middle wall. They are gravity structures founded on rock, except for the upper guide wall extension which is constructed of individual steel sheet pile cells.

5. Miter Gates: The lock gates are constructed of structural steel shapes and plates. The gate leaves for the 110-foot chamber are vertically framed. Each gate consists of two leaves which are hinged to embedded anchorages at the top by gudgeon pins and are supported at the bottom on steel pintles with the pintle bases embedded in concrete. Each leaf is a rectangular frame with vertical quoin and miter girders at the fixed and free ends respectively, and vertical beams and horizontal intercostals on the gate leaves for the 110-foot chamber.

6. Miter Gate Operating Equipment: The hydraulic system consists of three constant delivery oil pumps and one pressure holding oil pump, located on the first floor in the operation building on the land wall. The pumps supply oil under pressure to the hydraulic piping system for operation of the lock gate and culvert valve operating machinery on the lock walls. This system was installed in 1968 and replaced the original compressed air system for operation of the miter gates and the original hydraulic system installed for operation of the emptying and filling valves.

7. Dam Gates: The 13 submergible lift gates are steel structures arranged to travel on vertical tracks on the piers. Each gate can be raised to a point where its bottom is 39.4 feet above the sill and lowered to a point where its top is 3 feet below normal pool level. There is one Sidney gate located on the back channel dam. This gate combines features of both the tainter and vertical lift gates. The

gate works like a tainter gate until the gate reaches the limits of its rotation, after which the entire gate is raised by the lifting chains up to the maximum travel limit, which is 38 feet above the sill.

8. Dam Gate Operating Equipment: Two hoist motors and two synchronous tie motors of the slip-ring induction type are provided for each gate. A full magnetic reverse control panel operates the two hoist motors and the two synchronous tie motors for each gate from a remotely mounted master switch. In case of emergency, either hoisting motor may be cut out by means of switches and the gate can be operated by the remaining motor through the synchronous tie motors.

9. Tow Haulage Unit: All the tow haulage equipment is located on the middle wall and is used to assist tows in leaving the 110-ft land chamber. This equipment consists of the following: an electric motor driven pump; hydraulic motor driven grooved winch drum; towing bitt; controls; and miscellaneous items including rails, wire rope and sheaves. The system is designed for towing a maximum load of 18,000 pounds at a speed of 70 feet-per-minute.

10. Mooring Equipment: There are 20 check posts present for the 110= land chamber, 10 on the land wall and 10 on the land side of the middle wall. These are embedded on the top of the walls for routine tow stopping. One floating mooring bitt was installed on the land wall of the 110= chamber during the major rehabilitation in 1982. This installation facilitates locking through up-bound tows.

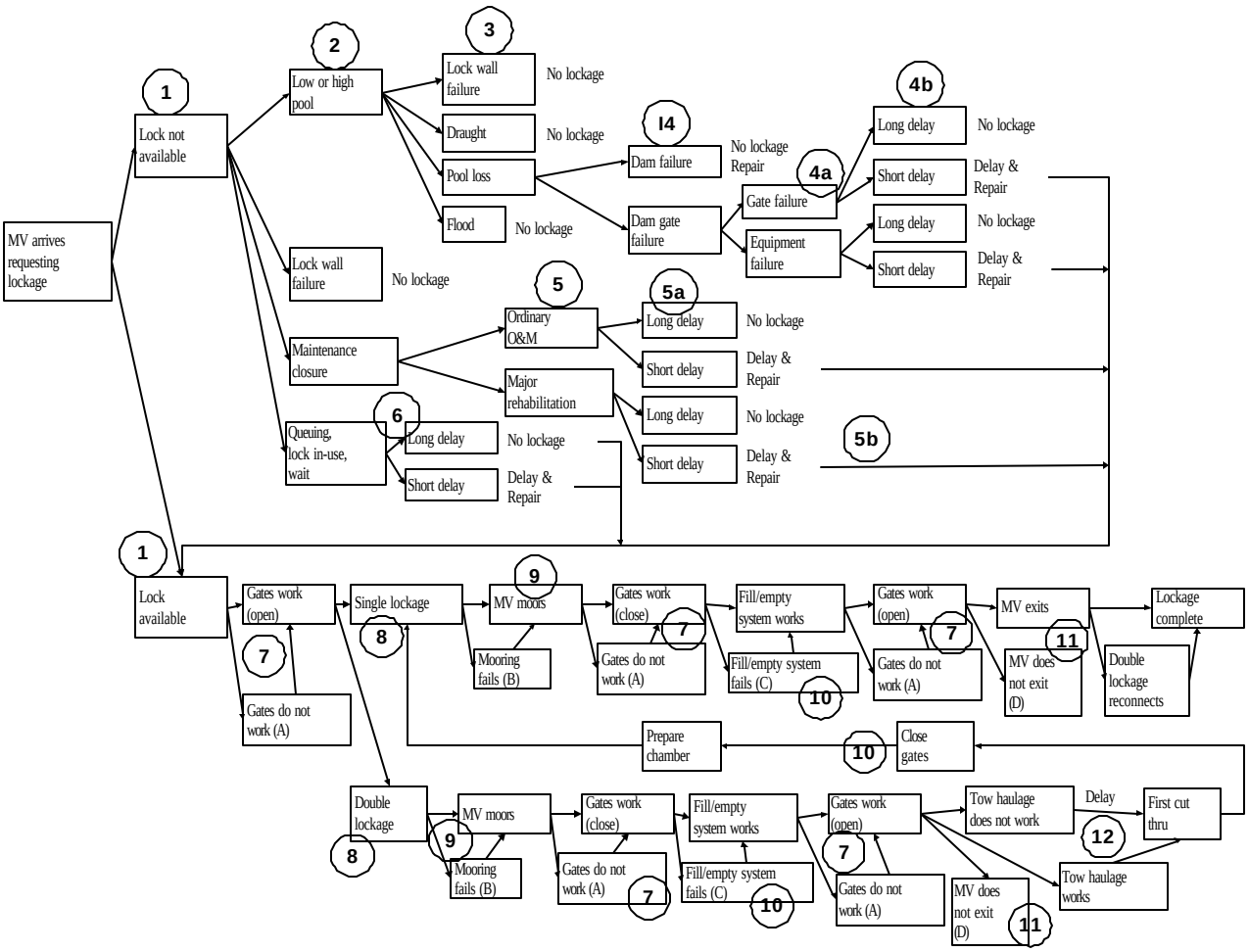

Figure 2-3a. Emsworth Navigation Lock on the Ohio River (Ayyub et al 1996)

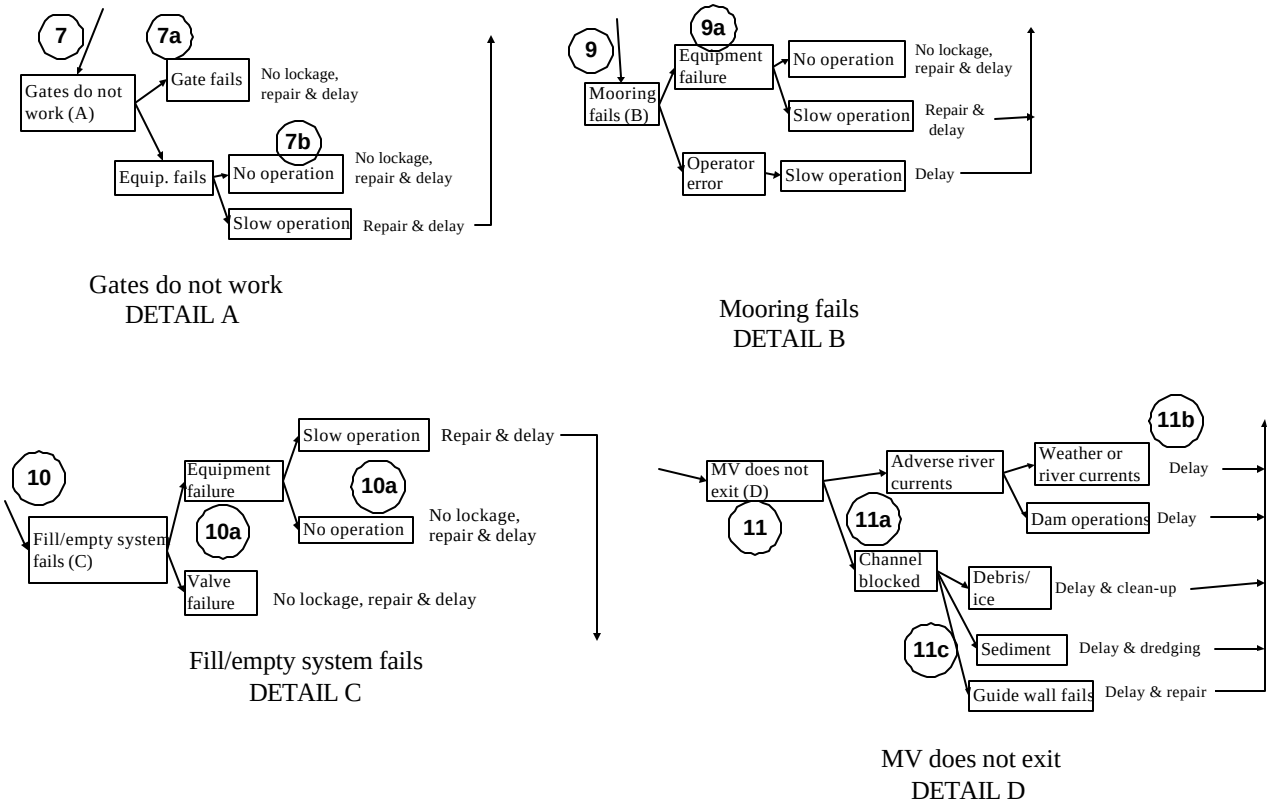

Figure 2-3b. Details for Emsworth Navigation Lock on the Ohio River (Ayyub et al 1996)

3. Conclusions

The expert-opinion elicitation process was defined as a formal, heuristic process of obtaining information or answers to specific questions about certain quantities, called issues, such as unsatisfactory-performance rates, unsatisfactory-performance consequences and expected service life. Historical, philosophical and analytical background on expert-opinion elicitation, its limitations, current uses, and example applications relevant to different engineering, planning, and operations decisions problems are provided in the guide. The guide provides a process for expert-opinion elicitation of probabilities and consequences for Corps facilities for the use of planners, engineers, and others should they choose to use expert judgment. The development of this guide resulted in the following conclusions:

1. Judgement and expert opinion in the presence of uncertainty frequently rely on simple cognitive heuristics, the outcomes of which depend on the issues and experts that are selected for this purpose. Although these cognitive heuristics commonly achieve the intended goal in most circumstances, they can be a source of bias and sometimes error.
2. Expert-opinion elicitation should not be used in lieu of rigorous reliability and risk analytical methods, but should be used to supplement them and to prepare for them. Also, it should be used in cases where reliability and risk analytical methods are inappropriate or inconsistent.
3. It should be preferably performed during a face-to-face meeting of members of an expert panel that is developed specifically for the issues under consideration. The meeting of the expert panel should be conducted after communicating to the expert in advance to the meeting background information, objectives, list of issues, and anticipated outcome from the meeting. In this document, the different components of the expert-opinion elicitation process are described, and then the process itself is outlined and discussed.
4. Because using expert judgment can be easily abused, the guide provides a process for the use of this technique and limitations of the method. The guide provides users with acceptable practice and usage of expert opinion in situations with a scarcity of historical data.
5. The selection of a scoring technique and an aggregation method of opinions should be made on a case by case basis.
6. The guide provides suggestions for avoiding pitfalls based on previous critique of Delphi techniques, methods of social research, and the Standards for Educational and Psychological Testing of the American Psychological Association.

4. Bibliography

1. Alpert, M., and Raiffa, H., 1982, "A Progress Report on the Training of Probability Assessors," in Kahneman, et al., (Eds.), *Judgement Under Uncertainty, Heuristics and Biases*, Cambridge University Press, Cambridge, 294-306.
2. Amendola, A., 1986. System Reliability Benchmark Exercise Parts I and II, EUR-10696, EN/I, Joint Research Center of Ispra, Italy.
3. American Institute of Aeronautics and Astronautics, 1998. Guide for Verification and Validation of Computational Fluid Dynamics Simulation, AIAA G-077-1998.
4. American Psychological Association, 1985. Standards for Educational and Psychological Testing, Washington, DC.
5. Ang, A., and Tang, W., Probability Concepts in Engineering Planning and Design, John Wiley, NY, 1975.
6. Ayyub, B. M., 1999. "Guidelines on Expert Elicitation of Probabilities and Consequences for Corps Facilities," Report, Institute for Water Resources, USACE, Alexandria, VA.
7. Ayyub, B.M., and Chao, R.-J., 1998. "Chapter 1. Uncertainty Modeling in Civil Engineering with Structural and Reliability Applications," in Uncertainty Modeling and Analysis in Civil Engineering, edited by B. Ayyub, CRC Press, 1-32.
8. Ayyub, B. M. 1994. "The Nature of Uncertainty in Structural Engineering," in Uncertainty Modelling and Analysis: Theory and Applications, edited by Ayyub, and Gupta, North-Holland-Elsevier Scientific Publishers, 195-210.
9. Ayyub, B. M., 1991. "Systems Framework for Fuzzy Sets in Civil Engineering," international journal of Fuzzy Sets and Systems, North-Holland, Amsterdam, 40(3), 491-508.
10. Ayyub, B. M., 1992. "Generalized Treatment of Uncertainties in Structural Engineering." *Analysis and Management of Uncertainty: Theory and Applications*, Edited by Ayyub and Gupta, Elsevier Science Publisher, NY, 235-246.
11. Ayyub, B. M., and McCuen, R., Probability, Statistics and Reliability for Engineers, CRC Press, FL, 1997.
12. Ayyub, B. M., Fault Tree Analysis of Cargo Elevators Onboard Ships, BMA Engineering Report, prepared for Naval Sea System Command, U.S. Navy, Crystal City, VA, 1992.
13. Ayyub, B. M., Handbook for Risk-Based Plant Integrity, BMA Engineering Report, prepared for Chevron Research and Technology Corporation, Richmond, CA, 1993.
14. Ayyub, B. M., Riley, B. C., and Hoge, M. T., 1996. Expert Elicitation of Unsatisfactory-Performance Probabilities and Consequences for Civil Works Facilities, Technical Report, USACE, Pittsburgh District, PA.
15. Ayyub, B.M., (Editor), 1998, Uncertainty Modeling and Analysis in Civil Engineering, CRC Press.

16. Ayyub, B.M., and Gupta, M.M., (Editors), 1997, Uncertainty Analysis in Engineering and the Sciences: Fuzzy Logic, Statistics, and Neural Network Approach, Kluwer Academic Publisher.
17. Ayyub, B.M., Guran, A., and Haldar, A., (Editors), 1997, Uncertainty Modeling in Vibration, Control, and Fuzzy Analysis of Structural Systems, World Scientific, 1997.
18. Bailey, K. D., 1994. *Methods of Social Research*. The Free Press, Maxwell Macmillan, NY.
19. Beacher, G., Expert Elicitation in Geotechnical Risk Assessment, USACE Draft Report, University of Maryland, College Park, MD.
20. Bell, T. E., and Esch, K., 1989. "The Space Shuttle: A Case of Subjective Engineering," IEEE Spectrum, June 1989, 42-46.
21. Blair, A. N., and Ayyub, B. M., "Fuzzy Stochastic Cost And Schedule Risk Analysis: MOB Case Study," Proceedings of the Symposium on Very Large Floating Structures, Elsveir, North Holland.
22. Blockley, D. I., 1975, "Predicting the Likelihood of Structural Accidents," *Proceedings*, Institution of Civil Engineers, London, England, 59, Part 2, 659-668.
23. Blockley, D. I., 1979a, "The Calculations of Uncertainty in Civil Engineering," *Proceedings*, Institution of Civil Engineers, London, England, 67, Part 2, 313-326.
24. Blockley, D. I., 1979b, "The Role of Fuzzy Sets in Civil Engineering," *Fuzzy Sets and Systems*, 2, 267-278.
25. Blockley, D. I., 1980, *The Nature of Structural Design and Safety*, Ellis Horwood, Chichester, UK.
26. Blockley, D. I., Pilsworth, B. W. and Baldwin, J.F., 1983, "Measures of Uncertainty," *Civil Engineering Systems*, 1, 3-9.
27. Bowles, D. 1990. "Risk Assessment in Dam Safety Decisionmaking." Risk-Based Decision Making in Water Resources, Proceedings of the Fourth Conference, Ed. by Y. Y. Haimes and E. Z. Stakhiv, 254-83.
28. Brown, C. B. and Yao, J. T. P., 1983, "Fuzzy Sets and Structural Engineering," *Journal of Structural Engineering*, ASCE, 109(5), 1211-1225.
29. Brown, C. B., 1979, "A Fuzzy Safety Measure," *Journal of Engineering Mechanics Division*, ASCE, 105(EM5), 855-872.
30. Brown, C. B., 1980, "The Merging of Fuzzy and Crisp Information," Journal of Engineering Mechanics Division, ASCE, 106(EM1), 123-133.
31. Brune, R., Weinstein, M., and Fitzwater, M., 1983. Peer Review Study of the Draft Handbook for Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications," NUREG/CR-1278.
32. Clemen, R. T., 1989. "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting, 5*, 559-583.
33. Colglazier, E. W., and Weatherwax, R. K., 1986. "Failure Estimates for the Space Shuttle," Abstracts from the Society of Risk Analysis, annual meeting, Boston, MA, Nov. 9-12, 1986, 80.
34. Committee on Safety Criteria for Dams, 1985. Safety of Dams: Flood and Earthquake Criteria. National Academy Press, Washington, D.C.
35. Committee on the Safety of Existing Dams, 1983. Safety of Existing Dams, Evaluation and Improvement. National Research Council, National Academy Press, Washington, D.C.
36. Cooke, R. M., 1986. "Problems with Empirical Bayes," Risk Analysis, 6(3), 269-272.
37. Cooke, R. M., 1991. *Experts in Uncertainty*, Oxford University Press.

38. De Finetti, B., 1937, English Translation in 1964 by H. Kyburg and H. Smokler (eds.). *Studies in Subjective Probabilities*, Wiley, NY.
39. Defense Acquisition University, 1998. "Risk Management Guide," Defense Systems Management College Press, Fort Belvoir, VA
40. DeKay, M. L., and McClelland, G. H., 1993, "Predicting Loss of Life in Cases of Dam Failure and Flash Flood," Risk Analysis, 13(2), 193-205.
41. Ferrell, W. R., 1985. "Combining Individuals Judgments," in Wright, G. (Ed.), *Behavioral Decision Making*, Plenum, NY.
42. Ferrell, W. R., 1994. "Discrete Subjective Probabilities and Decision Analysis: Elicitation, Calibration and Combination," in, Wright, G., and Ayton, P. (Eds.), *Subjective Probability*, John Wiley and Sons, NY.
43. Freeman, W. M., 1969. Readings from Scientific America: Science, Conflict and Society, San Francisco, CA.
44. French, S., 1985. Group Consensus Probability Distributions: A Critical Survey," J. M. Bernardo et al (eds.) Bayesian Statistics, Elsevier, North Holland, 183-201.
45. Furuta, H., Fu, K. S., and Yao, J. T. P., 1985, "Structural Engineering Applications of Expert Systems," Computer Aided Design, 17(9), 410-419.
46. Furuta, H., Shiraishi, N., and Yao, J. T. P., 1986, "An Expert System for Evaluation of Structural Durability," *Proceedings of fifth OMAE Symposium*, 1, 11-15.
47. Galanter, E., 1962. "The Direct Measurement of Utility and Subjective Probability," American J. of Psychology, 75, 208-220.
48. Genest, C., and Zidek, J., 1986. "Combining Probability Distributions: Critique and an Annotated -148.
49. Gustafson, D. H., Shukla, R. K., Delbecq, A., and Walster, G. W., (1973). "A Comparative Study of Differences in Subjective Likelihood Estimates Made by Individuals, Interacting Groups, Delphi Groups, and Nominal Groups," Organizational Behavior and Human Performance, 9, 200-291.
50. Haldar, A, Guran, A., and Ayyub, B.M., (Editors), 1997, Uncertainty Modeling in Finite Element, Fatigue, and Stability of Systems, World Scientific.
51. Hartford, D.N.D., *How Safe is Your Dam? Is it Safe Enough?,* B.C. Hydro, Maintenance, Engineering, and Projects, Burnaby, BC, 1995.
52. Helmer, O., 1968. "Analysis of the Future: The Delphi Method," and "The Delphi Method An Illustration," in J. Bright (ed.), Technological Forecasting for Industry and Government, Prentice Hall, Englewood Cliffs, NJ.
53. Henley, E. J., and Kumamoto, H., Reliability Engineering and Risk Assessment, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1981.
54. Horwich, P., 1987, *Asymmetries in Time: Problems in the Philosophy of Science*, MIT Press, Cambridge, MA.
55. Ishizuka, M., Fu, K. S., and Yao, J. T. P., 1981, "A rule-Inference Method for Damage Assessment," ASCE *Preprint 81-502*, ASCE, St. Louis, Missouri.
56. Ishizuka, M., Fu, K. S., and Yao, J. T. P., 1983, "Rule-Based Damage Assessment System for Existing Structures," *Solid Mechanics Archives*, 8, 99-118.
57. Itoh, S., and Itagaki, H., 1989, "Application of Fuzzy-Bayesian Analysis to Structural Reliability," *the proceedings of the Fifth International Conference on Structural Safety and Reliability,*

ICOSSAR, Volume 3, published by ASCE, Edited by A. H-S. Ang, M. Shinozuka and G. I. Schuëller, San Francisco, 1771-1774.
58. Kahn, H., 1960. *On Thermonuclear War*, Free Press, NY.
59. Kahn, H., and Wiener, A. J., 1967. *The Year 2000: A Framework for Speculation*, Macmillan, NY.
60. Kahneman, D., Slovic, P., and Tversky, A. (eds.). 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.
61. Kaneyoshi, M., Tanaka, H., Kamei, M., and Furuta, H., 1990, "Optimum Cable Tension Adjustment Using Fuzzy Regression Analysis," the third WG 7.5 working conference on reliability and optimization of structural systems, *International Federation for Information Processing*, University of California, Berkeley, CA, 11 p.
62. Kaufman, A. and Gupta, M. M., 1985, *Introduction to Fuzzy Arithmetic, Theory and Applications*, Van Nostrand Reinhold Co., New York.
63. Kaufmann, A., 1975, *Introduction to the Theory of Fuzzy Subsets*, Academic Press, New York, N.Y., (Translated by D. L. Swanson).
64. Klir, G. J., 1985, *Architecture of Systems Problem Solving*, Plenum Press, New York.
65. Klir, G. J., and Folger, T. A., 1988, *Fuzzy Sets, Uncertainty, and Information*, Prentice Hall, N.J.
66. Lai, K.-L., and Ayyub, B.M., 1994. "Generalized Uncertainty in Structural Reliability Assessment," Civil Engineering Systems, 11(2), 81-110.
67. Langer, E., 1975. "The Illusion of Control," J. of Personality and Social Psychology, Vol. 32, 311-328.
68. Lichtenstein, S., and J.R. Newman, 1967. "Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities," *Psychonometric Science*, *9*(10), 563-564.
69. Lindley, D., 1970. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, UK.
70. Linstone H. A., and Turoff, M., 1975. The Delphi Method, Techniques and Applications, Addison Wesley, MA.
71. Moore, R. E., 1966. *Interval analysis*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
72. Morgan, M. G., and Henrion, M., 1992. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press, New York.
73. Morris, J. M., and D'Amore, R. J., 1980. "Aggregating and Communicating Uncertainty," Pattern Analysis and Recognition Corp., 228 Liberty Plaza, Rome, NY.
74. Murphy A., and Daan, H., 1984. "Impact of Feedback and Experience on the Quality of Subjective Probability Forecasts: Comparison of the results from the first and second years of the Zierikzee Experiment," Monthly Weather Review, Vol. 112, 413-423.
75. Newman, J. R., 1961. "Thermonuclear War," Scientific America, March 1961.
76. Nuclear Regulatory Commission, 1975. Reactor Safety Study, WASH-1400, NUREG 751014.
77. Nuclear Regulatory Commission, 1997. Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Expert Use, prepared by the Senior Seismic Hazard Analysis Committee, NUREG/CR-6372, UCRL-ID-122160, Vol. 1 and 2, Washington, DC.
78. Paté-Cornell, E., "Uncertainties in Risk Analysis and Comparability of Risk Estimates," Society for Risk Analysis 1996 Annual Meeting, McLean, VA.

79. Ponce, V. M., 1989. *Engineering Hydrology Principles and Practices*. Prentice-Hall, Englewood Cliffs, NJ, 640 p.
80. Preyssl C, and Cooke, R., 1989. "Expert Judgment: Subjective and Objective Data for Risk Analysis for Space-flight Systems," Proceedings PSA 1989, Pittsburg, PA, April 2-7, 1989.
81. Ramsey, F, 1931. "Truth and Probability," in Braithwaite (ed.), *The Foundation of Mathematics*, Kegan Paul, London, 156-198.
82. Reichenbach, H., 1951. *The Rise of Scientific Philosophy*, University of California Press, 1968 edition.
83. Rowe, G., 1992. "Perspectives on Expertise in Aggregation of Judgments," in Wright, G., and Bolger, F. (Eds.), *Expertise and Decision Support*, Plenum Press, NY, 155-180.
84. Sackman, H., 1975. Delphi Critique: Expert Opinion, Forecasting and Group Process, Lexington Books, Lexington, MA.
85. Samet, M. G., 1975. "Quantitative Interpretation of Two Qualitative Scales Used to Rate Military Intelligence," Human Factors, 17(2), 192-202.
86. *Science*, 1983. Volume 222, No. 4630, p. 1293, December 23, 1983.
87. Shiraishi, N. and Furuta, H., 1983, "Reliability Analysis Based on Fuzzy Probability," *Journal of Engineering Mechanics*, ASCE, 109(6), 1445-1459.
88. Shiraishi, N. Furuta, H., and Sugimoto, M., 1985, "Integrity Assessment of Structures Based on Extended Multi-Criteria Analysis, "*Proc. of the fourth ICOSSAR*, Kobe, Japan.
89. Smithson, M., 1988, *Ignorance and Uncertainty*, Springer-Verlag, New York, NY.
90. Spetzler, C. S., and Stael von Holstein, C.-A. S., 1975. "Probability Encoding in Decision Analysis," Management Science, 22(3).
91. Thys, W., 1987. *Fault Management*. PhD Dissertation, Delft University of Technology, Delft.
92. Trauth, K. M., Hora, S. C., and Guzowski, R. V., 1993. Expert Judgement on Markers to Deter Inadvertent Human Intrusion into the Waste Isolation Pilot Plant, Report SAND92-1382, Sandia National Laboratories, Albuquerque, NM.
93. U.S. Bureau of Reclamation, *Policy and Procedures for Dam Safety Modifications*, USBR, Denver, 1989.
94. U.S. Army Corps of Engineers, 1965. Standard Project Flood Determinations. Civil Engineer Bulletin No. 52-8, Engineering Manual EM 1110-2-1411.
95. U.S. Army Corps of Engineers, 1982. National Program of Inspection of Nonfederal Dams, Final Report to Congress, Engineering Report ER 1110-2-106.
96. U.S. Army Corps of Engineers, 1997. Guidelines for Risk-based Assessment of Dam Safety, Draft Engineering Pamphlet EP 1110-1-XX, CECW-ED.
97. U.S. Interagency Advisory Committee on Water Data, Hydrology Subcommittee, 1982. Guidelines for Determining Flood Flow Frequency. Bulletin No. 17B, USGS, Reston, VA.
98. Wiggins, J., 1985. ESA Safety Optimization Study, Hernandez Engineering, HEI-685/1026, Houston TX.
99. Winkler, R.N and Murphy, A., 1968, "Good Probability Assessors," *Journal of Applied Meteorology*, Vol. 7, 751-758.

100. Yao, J. T. P. and Furuta, H., 1986, "Probabilistic Treatment of Fuzzy Events in Civil Engineering," *Probabilistic Engineering Mechanics*, 1(1), 58-64.
101. Yao, J. T. P., 1979, "Damage Assessment and Reliability Evaluation of Existing Structures," *Engineering Structures*, England, 1, 245-251.
102. Yao, J. T. P., 1980, "Damage Assessment of Existing Structures," *Journal of Engineering Mechanics Division*, ASCE, 106(EM4), 785-799.
103. Zadeh, L. A., 1965, "Fuzzy Sets," *Information and Control*, 8, 338-353.
104. Zadeh, L. A., 1968, "Probability Measures of Fuzzy Events," *J. of Math. Analysis*, 23, 421-427.
105. Zadeh, L. A., 1973, "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(1), 28-44.
106. Zadeh, L. A., 1975, "The Concept of Linguistic Variable and Its Application to Approximate Reasoning," Parts I, II and III, *Information and Control*, Vol. 8, pp. 199-249, 301-357, Vol. 9, pp. 43-80.
107. Zadeh, L. A., 1987, "Fuzzy Sets as a Basis for a Theory of Possibility," *Fuzzy Sets and Systems*, 1, 3-28.
108. Zadeh, L. A., Fu, K. S., Tanaka, K. and Shimara, J., 1975, *Fuzzy Sets and Their Application to Cognitive and Decision Processes*, Academic Press, New York, N.Y.

Appendix A. Occurrence Probabilities, Moments and Percentiles

A.1. Background

Knowledge, information, ignorance and uncertainty were discussed in great levels of detail in Chapter 1. In the expert-opinion elicitation process, terms, such as uncertainty, probability, unsatisfactory-performance rate, mean (or average) value, average time between unsatisfactory-performances, and median value are commonly used. Other terms such as dispersion (variability), variance, standard deviation, coefficient of variation and percentiles are also used. The objective herein is to provide definitions and background information on these terms. These definitions are partly taken from Ayyub and McCuen (1997).

Engineers must make decisions under conditions of uncertainty. It is common in engineering to use probabilistic analysis to deal with uncertainty. For example, engineers who have the responsibility of monitoring water quality in our nation's streams and bodies of water estimate pollution levels using samples collected from the water. The samples are then analyzed in a laboratory and the results are used to make a decision. Most sampling programs involve 10 or fewer measurements. Uncertainty arises because of the highly variable nature of pollution; that is, the concentration of a pollutant may vary with time, the degree of turbulence in the water, and the frequency with which wastes are discharged into the water. These sources of variation must be accounted for when the engineer makes a decision about water quality.

Traffic engineers must also make decisions under conditions of uncertainty. For example, intersections are frequently the sites of accidents. The traffic engineer knows that accidents can be reduced by installing stop signs or traffic lights. However, there is a cost associated with installing such hardware. Also, traffic controls can cause delay and inconvenience to those that must travel through the intersections. Thus, the traffic engineer must consider numerous factors in making a decision, including the likelihood and severity of accidents at the intersection and the traffic load in each direction. The frequency and severity of accidents can be assessed using data from accidents that have occurred at that intersection in the past. However, these are data of the past, and there is no assurance that they will accurately reflect accident rates in the future. Reduced travel due to an increase in the cost of gasoline may reduce the number of accidents. Data on the traffic volumes originating from each street entering the intersection can be obtained using traffic counters. However, these data may not completely characterize the traffic volumes that will take place in the future. For example, if the traffic volume data

are collected during the summer, the opening of schools may alter the relative proportion of traffic volumes on each street entering the intersection. Such sources of diversity introduce uncertainty into the decision-making process.

A.2. Definition of Probability

The concept of probability has its origin in games of chance. In these games, probabilities are determined based on many repetitions of an experiment and counting the number of outcomes of an event of interest. Then, *the probability of the outcome of interest can be measured by dividing the number of occurrences of an event of interest by the total number of repetitions*. Quite often, probability is specified as a percentage; for example, when the weather bureau indicates that there is a 30 percent chance of rain, experience indicates that under similar meteorological conditions it has rained 3 out of 10 times. In this example, the probability was estimated empirically using the concept of relative frequency expressed as

$$P(X = x) = \frac{n}{N} \tag{A-1}$$

in which n = number of observations on the random variable X that results in an outcome of interest x, and N = total number of observations of x. The probability of an event x in this equation was defined as the *relative frequency* of its occurrence. Also, probability can be defined as a *subjective probability* (or called *judgmental probability*) of the occurrence of the event. The type of definition depends on the underlying event. For example, in an experiment that can be repeated N times with n occurrences of the underlying event, the relative frequency of occurrence can be considered as the probability of occurrence. In this case, the probability of occurrence is n/N. However, there are many engineering problems that do not involve large numbers of repetitions, and still we are interested in estimating the probability of occurrence of some event. For example, during the service life of an engineering product, the product either fails or does not fail in performing a set of performance criteria. The events of unsatisfactory-performance and satisfactory-performance are mutually exclusive and collectively exhaustive of the sample space (that is the space of all possible outcomes). The probability of unsatisfactory-performance (or satisfactory-performance) can be considered as a *subjective probability*. Another example is the failure probability of a dam due to an extreme flooding condition. An estimate of such probabilities can be achieved by modeling the underlying system, its uncertainties and performances. The resulting subjective probability is expected to reflect the *status of our knowledge* about the system regarding occurrence of the events of interest. Therefore, subjective probabilities can be associated with degrees of belief, and can form a basis for Bayesian methods (Ayyub and McCuen 1997). It is important to keep in mind both definitions, so that results are not interpreted beyond the range of their validity.

An axiomatic definition of probability is commonly provided in the literature such as Ayyub and McCuen (1997). For a event A, the notation $P(A)$ means the probability of occurrence of the event A. The probability $P(.)$ should satisfy the following axioms:

1. $0 \leq P(A) \leq 1$, for any A that belongs to the *set of all possible outcomes* (called sample space *S*) for the system.
2. The probability of having *S*, $P(S) = 1$.
3. The occurrence probability of the union of mutually exclusive events is the sum of their individual occurrence probabilities.

The first axiom states that the probability of any event is inclusively between 0 and 1. Therefore, negative probabilities, or probabilities larger than one are not allowed. The second axiom comes from the definition of the sample space. Since the sample space is the set of all possible outcomes, therefore, one or more of these outcomes must occur resulting in the occurrence of S. If the probability of the sample space does not equal 1, this means that the sample space was incorrectly defined. The third axiom sets a basis for the mathematics of probability. These axioms as single entity can be viewed as a definition of probability, i.e., any numerical structure that adheres to these axioms will provide a probability structure. Therefore, the relative frequency and subjective probability meet this definition of probability.

The relative-frequency and subjective-probability concepts are tools that help engineers and planners to deal with and model uncertainty, and should be used appropriately as engineering systems and models demand. In the case of relative frequency, increasing the number of repetitions according to Eq. A-1 would produce an improved estimate with a diminishing return on invested computational and experimental resources until a limiting (i.e. long-run or long-term) frequency value is obtained. This limiting value can be viewed as the *true probability* although the absolute connotation in this terminology might not realistic and cannot be validated. Philosophically, a true probability might not exist especially when dealing with subjective probabilities. This, however, does not diminish the value of probabilistic analysis and methods since they provide a consistent, systematic, rigorous, and robust framework for dealing with uncertainty and decision making.

A.2.1. Linguistic Probabilities

Probability as described in the previous section provides a measure of the likelihood of occurrence of an event. It is a numerical expression of uncertainty; however, it is common for subjects (such as experts) to express uncertainty verbally using linguistic terms, such as likely, probable, improbable, …, etc. Although, these linguistic terms are somewhat fuzzy, they are meaningful. Lichtenstein and Newman (1967) developed a table that translates commonly used linguistic terms into probability values using responses from subjects. The Lichtenstein and Newman (1967) summary is shown in Table A-1 (Baecher 1998). The responses of the subjects show encouraging consistency in defining each term, however the ranges of responses are large. Moreover, mirror-image pairs sometimes produce asymmetric results. The term "Rather unlikely" is repeated in table as it was used twice in the questionnaire to the subjects at almost the start and at the end of the questionnaire to check consistency. It can be concluded from this table that verbal descriptions of uncertainty can be useful as an initial assessment, but other analytical techniques should be used to assess uncertainty; for example the linguistic terms in Table A-1 can be modeled using fuzzy sets (Haldar et al 1997, Ayyub et al 1997, Ayyub and Gupta 1997, Ayyub 1998).

Table A-1. Linguistic Probabilities and Translations (Lichtenstein and Newman 1967)

Rank	Phrase	No. of Responses	Mean	Median	Standard Deviation	Range
1	Highly probable	187	0.89	0.90	0.04	0.60-0.99
2	Very likely	185	0.87	0.90	0.06	0.60-0.99
3	Very probable	187	0.87	0.89	0.07	0.60-0.99
4	Quite likely	188	0.79	0.80	0.10	0.30-0.99
5	Usually	187	0.77	0.75	0.13	0.15-0.99
6	Good chance	188	0.74	0.75	0.12	0.25-0.95
7	Predictable	146	0.74	0.75	0.20	0.25-0.95
8	Likely	188	0.72	0.75	0.11	0.25-0.99
9	Probable	188	0.71	0.75	0.17	0.01-0.99
10	Rather likely	188	0.69	0.70	0.09	0.15-0.99
11	Pretty good chance	188	0.67	0.70	0.12	0.25-0.95
12	Fairly likely	188	0.66	0.70	0.12	0.15-0.95
13	Somewhat likely	187	0.59	0.60	0.18	0.20-0.92
14	Better than even	187	0.58	0.60	0.06	0.45-0.89
15	Rather	124	0.58	0.60	0.11	0.10-0.80
16	Slightly more than half the time	188	0.55	0.55	0.06	0.45-0.80
17	Slight odds in favor	187	0.55	0.55	0.08	0.05-0.75
18	Fair chance	188	0.51	0.50	0.13	0.20-0.85
19	Tossup	188	0.50	0.50	0.00	0.45-0.52
20	Fighting chance	186	0.47	0.50	0.17	0.05-0.90
21	Slightly less than half the time	188	0.45	0.45	0.04	0.05-0.50
22	Slight odds against	185	0.45	0.45	0.11	0.10-0.99
23	Not quite even	180	0.44	0.45	0.07	0.05-0.60
24	Inconclusive	153	0.43	0.50	0.14	0.01-0.75
25	Uncertain	173	0.40	0.50	0.14	0.08-0.90
26	Possible	178	0.37	0.49	0.23	0.01-0.99
27	Somewhat unlikely	186	0.31	0.33	0.12	0.03-0.80
28	Fairly unlikely	187	0.25	0.25	0.11	0.02-0.75
29	Rather unlikely	187	0.24	0.25	0.12	0.01-0.75
30	Rather unlikely	187	0.21	0.20	0.10	0.01-0.75
31	Not very probable	187	0.20	0.20	0.12	0.01-0.60
32	Unlikely	188	0.18	0.16	0.10	0.01-0.45
33	Not much chance	186	0.16	0.15	0.09	0.01-0.45
34	Seldom	188	0.16	0.15	0.08	0.01-0.47
35	Barely possible	180	0.13	0.05	0.17	0.01-0.60
36	Faintly possible	184	0.13	0.05	0.16	0.01-0.50
37	Improbable	187	0.12	0.10	0.09	0.01-0.40
38	Quite unlikely	187	0.11	0.10	0.08	0.01-0.50
39	Very unlikely	186	0.09	0.10	0.07	0.01-0.50
40	Rare	187	0.07	0.05	0.07	0.01-0.30
41	Highly improbable	181	0.06	0.05	0.05	0.01-0.30

A.2.2. Unsatisfactory-Performance Rate

Unsatisfactory-performance rate can be defined as the probability of unsatisfactory-performance per unit time or a unit of operation, such as cycle, revolution, rotation, start-up, etc. For example, a constant unsatisfactory-performance rate for an electronic device of 0.1 per year means that on *the average* the device fails once per 10 years. Another example that does not involve time is an engine with a unsatisfactory-performance rate of 10^{-5} per cycle of operation (or it can be in terms of mission length). In this case, the unsatisfactory-performance rate means that on the average the engine fails once per 100,000 cycles. Due to manufacturing, assembly and aging effects, unsatisfactory-performance rates can generally be variant with time (or other units of operation), therefore, requiring sometimes a statement of limitation on their applicability. Unsatisfactory-performance rates can be used in probabilistic analysis. There are analytical methods to convert unsatisfactory-performance rates into probabilities of some events of interest.

A.3. Central Tendency Measures

A very important descriptor of data is central-tendency measures. The central tendency can be measured using, for example, (1) the mean (or average) value, or (2) the median value.

A.3.1. Mean (or Average) Value

The average value is the most commonly used central-tendency descriptor. The definition of the mean (or average) value herein is based on a sample of size n. The sample consists of n values of a random variable X. For n observations, if all observations are given equal weights, then the average value is given by

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(A-2)}$$

where x_i = a sample point, and i = 1, 2, ..., n; and

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n \qquad \text{(A-3)}$$

Since the average value ($\overline{X}$) is based on a sample, it has statistical error due to two reasons: (1) it is sample dependent, i.e., a different sample might produce a different average, and (2) it is sample-size dependent, i.e., as the sample size is increased, the error is expected to reduce. The mean value has another mathematical definition that is based on probability distributions according to probability theory, which is not described herein.

A.3.2. Average Time Between Unsatisfactory Performances

The average time between unsatisfactory-performances can be computed as the average ($\overline{X}$), where x_i = a sample point indicating the age at unsatisfactory performance of a failed component, and i = 1, 2, ..., n. The average time between unsatisfactory performance is related to the unsatisfactory-

performance rate as its reciprocal. For example a component with a unsatisfactory-performance rate of 0.1 per year, has an average time between unsatisfactory-performances of 1/0.1 = 10 years. Similar to unsatisfactory-performance rates, the average time between unsatisfactory-performances can be constant or time-dependent.

A.3.3. Median Value

The median value x_m is another measure of central tendency. It is defined as the point that divides the data into two equal parts, i.e., 50% of the data are above x_m and 50% are below x_m. The median value can be determined by ranking the n values in the sample in decreasing order, 1 to n. If n is an odd number, then the median is the value with a rank of $(n+1)/2$. If n is an even number, then the median equals the average of the two middle values, i.e., those with ranks $n/2$ and $(n/2)+1$.

The advantage of using the median value as a measure of central tendency over the average value is its insensitivity to extreme values. Consequently, this measure of central tendency is commonly used in combining expert judgments in an expert-opinion elicitation process.

A.4. Dispersion (or Variability)

Although the central tendency measures convey certain information about the underlying sample, they do not completely characterize the sample. Two random variables can have the same mean value, but different levels of data scatter around the computed mean. Thus, measures of central tendency cannot fully characterize the data. Other characteristics are also important and necessary. The dispersion measures describe the level of scatter in the data about the central tendency location.

The most commonly used measure of dispersion is the variance and other quantities that are derived from it, such as, the standard deviation and coefficient of variation. For n observations in a sample that are given equal weight, the variance (S^2) is given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{X})^2 \qquad \text{(A-4)}$$

The units of the variance are the square of the units of the variable x; for example, if the variable is measured in pounds per square inch (psi), the variance has units of $(psi)^2$. Computationally, the variance of a sample can be determined using the following alternative equation:

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \qquad \text{(A-5)}$$

By definition the standard deviation (S) is the square root of the variance as follows:

$$S = \sqrt{\frac{1}{n\text{-}1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]} \qquad \text{(A-6)}$$

It has the same units as both the underlying variable and the central tendency measures. Therefore, it is a useful descriptor of the dispersion or spread of a sample of data. The coefficient of variation (*COV* or δ) is a normalized quantity based on the standard deviation and the mean, and is different from the covariance. Therefore, the *COV* is dimensionless, and is defined as

$$COV = \frac{S}{\overline{X}} \qquad \text{(A-7)}$$

It is also used as an expression of the standard deviation in the form of a percent of the average value. For example, consider $\overline{X}$ and *S* to be 50 and 20, respectively; therefore, *COV*(*X*) = 0.4 or 40%. In this case, the standard deviation is 40% of the average value.

A.5. Percentiles

A p-percentile value (x_p) for a random variable based on a sample is the value of the parameter such that *p*% of the data is less or equal to x_p. On the basis of this definition, the median value is considered to be the 50-percentile value.

Aggregating the opinions of experts sometimes requires the computation of the 25, 50 and 75 percentile values. The computation of these values depends on the number of experts providing opinions. Table A-2 provides a summary of the needed equations for 4 to 20 experts. In the table, X_i means the opinion of an expert with the i^{th} smallest value; i.e., $X_1 \geq X_2 \geq X_3 \geq \ldots \geq X_n$, where n = number of experts. In the table, the arithmetic average was used to compute the percentiles. In some cases, where the values of X_i differ by power order of magnitude, the geometric average can be used. Expert opinions should not be aggregated in this manner all the times, other aggregation methods as provided in Section C.4 might be more appropriate and should be considered.

Table A-2. Computations of Percentiles

Number of experts (n)	25 percentile		50 percentile		75 percentile	
	Arithmetic Average	Geometric Average	Arithmetic Average	Geometric Average	Arithmetic Average	Geometric Average
4	$(X_1+X_2)/2$	$\sqrt{X_1 X_2}$	$(X_2+X_3)/2$	$\sqrt{X_2 X_3}$	$(X_3+X_4)/2$	$\sqrt{X_3 X_4}$
5	X_2	X_2	X_3	X_3	X_4	X_4
6	X_2	X_2	$(X_3+X_4)/2$	$\sqrt{X_3 X_4}$	X_5	X_5
7	$(X_2+X_3)/2$	$\sqrt{X_2 X_3}$	X_4	X_4	$(X_5+X_6)/2$	$\sqrt{X_5 X_6}$
8	$(X_2+X_3)/2$	$\sqrt{X_2 X_3}$	$(X_4+X_5)/2$	$\sqrt{X_4 X_5}$	$(X_6+X_7)/2$	$\sqrt{X_6 X_7}$
9	$(X_2+X_3)/2$	$\sqrt{X_2 X_3}$	X_5	X_5	$(X_7+X_8)/2$	$\sqrt{X_7 X_8}$
10	$(X_2+X_3)/2$	$\sqrt{X_2 X_3}$	$(X_5+X_6)/2$	$\sqrt{X_4 X_5}$	$(X_8+X_9)/2$	$\sqrt{X_8 X_9}$
11	X_3	X_3	X_6	X_6	X_9	X_9
12	X_3	X_3	$(X_6+X_7)/2$	$\sqrt{X_6 X_7}$	X_{10}	X_{10}
13	$(X_3+X_4)/2$	$\sqrt{X_3 X_4}$	X_7	X_7	$(X_{10}+X_{11})/2$	$\sqrt{X_{10} X_{11}}$
14	$(X_3+X_4)/2$	$\sqrt{X_3 X_4}$	$(X_7+X_8)/2$	$\sqrt{X_7 X_8}$	$(X_{11}+X_{12})/2$	$\sqrt{X_{11} X_{12}}$
15	X_4	X_4	X_8	X_8	X_{12}	X_{12}
16	X_4	X_4	$(X_8+X_9)/2$	$\sqrt{X_8 X_9}$	X_{13}	X_{13}
17	$(X_4+X_5)/2$	$\sqrt{X_4 X_5}$	X_9	X_9	$(X_{13}+X_{14})/2$	$\sqrt{X_{13} X_{14}}$
18	$(X_4+X_5)/2$	$\sqrt{X_4 X_5}$	$(X_9+X_{10})/2$	$\sqrt{X_9 X_{10}}$	$(X_{14}+X_{15})/2$	$\sqrt{X_{14} X_{15}}$
19	X_5	X_5	X_{10}	X_{10}	X_{15}	X_{15}
20	X_5	X_5	$(X_{10}+X_{11})/2$	$\sqrt{X_{10} X_{11}}$	X_{15}	X_{15}

A.6. Statistical Uncertainty

Values of random variables obtained from sample measurements are commonly used in making important engineering decisions. For example, samples of river water are collected to estimate the average level of a pollutant in the entire river at that location. Samples of stopping distances are used to develop a relationship between the speed of a car at the time the brakes are applied and the distance traveled before the car comes to a complete halt. The average of sample measurements of the compressive strength of concrete collected during the pouring of a large concrete slab, such as the deck of a parking garage, is used to help decide whether or not the deck has the strength specified in the design specifications. It is important to recognize the random variables involved in these cases. In each case, the individual measurements or samples are values of a random variable, and the computed mean is also the value of a random variable. For example, the transportation engineer measures the stopping distance; each measurement is a sample value of the random variable. If ten measurements are made for a car stopping from a speed of 50 mph then the sample consists of ten values of the random variable. Thus, there are two random variables in this example: the stopping distance and the estimated

mean of the stopping distance; this is also true for the water-quality-pollutant and compressive-strength examples.

The estimated mean for a random variable is considered by itself to be a random variable, because different samples about the random variable can produce different estimated mean values; hence the randomness in the estimated mean. When a sample of n measurements of a random variable is collected, the n values are not necessarily identical. The sample is characterized by variation. For example, let's assume that five independent estimates of the compressive strength of the concrete in a parking garage deck are obtained from samples of the concrete obtained when the concrete was poured. For illustration purposes, let's assume that the five compressive strength measurements are 3250, 3610, 3460, 3380 and 3510 psi. This produces a mean of 3442 psi and a standard deviation of 135.9 psi. Assume that another sample of five measurements of concrete strength was obtained from the same concrete pour; however, the values were 3650, 3360, 3328, 3420, and 3260 psi. In this case, the estimated mean and standard deviation are 3404, and 149.3 psi, respectively. Therefore, the individual measurement and the mean are values of two different random variables, i.e., X and .

It would greatly simplify decision making if the sample measurements were identical, i.e., there was no sampling variation so the standard deviation was zero. Unfortunately, that is never the case, so decisions must be made in spite of the uncertainty. For example, let's assume in the parking garage example that the building code requires a mean compressive strength of 3500 psi. Since the mean of 3442 psi based on the first sample is less than the required 3500 psi, should we conclude that the garage deck does not meet the design specifications? Unfortunately, decision making is not that simple. If a third sample of five measurements had been randomly collected from other locations on the garage deck, the following values are just as likely to have been obtained: 3720, 3440, 3590, 3270, and 3610 psi. This sample of five produces a mean of 3526 psi and a standard deviation of 174.4 psi. In this case, the mean exceeds the design standard of 3500 psi. Since the sample mean is greater than the specified value of 3500 psi, can we conclude that the concrete is of adequate strength? Unfortunately, we cannot conclude with certainty that the strength is adequate any more than we could conclude with the first sample that the strength was inadequate. The fact that different samples lead to different means is an indication that we cannot conclude that the design specification is not met just because the sample mean is less than the design standard. We need to have more assurance.

The data that are collected on some variable or parameter represent sample information but it is not complete by itself, and predictions are not made directly from the sample. The intermediate step between sampling and prediction is the identification of the underlying population. The sample is used to identify the population and then the population is used to make predictions or decisions. This sample-to-population-to-prediction sequence is true for the univariate methods of this chapter.

The need then is for a systematic decision process that takes into account the variation that can be expected from one sample to another. The decision process must also be able to reflect the risk of making an incorrect decision. This decision making can be made using, for example, hypothesis testing as described by Ayyub and McCuen (1997).

Appendix B. Unsatisfactory-Performance Consequences

Risk analysis and risk-based decision making for maintaining the integrity of a facility requires estimating the likelihood of unsatisfactory performance, and unsatisfactory-performance consequences of the different components of a system. The objective of this section is to discuss consequence types, and provide methods for quantifying unsatisfactory-performance consequences. This section was adapted from Ayyub (1992 and 1993).

Unsatisfactory-performance consequences in this document are limited, for illustrative purposes only, to (1) production loss including delays, (2) property damage that includes repair, and (3) flood inundation. Other consequences such as loss of life, injuries, ecological effects, various types of environmental damage, and social and cultural impacts are not considered herein in detail. The assessment of consequences can be based on accident or unsatisfactory-performance reports, operational and production logs or files, analytical predictions of potential consequences due to different unsatisfactory-performance scenarios, and formal expert-opinion elicitation. In cases where the available time for performing risk analysis is limited, preference should be given to the use of risk analysis methods for all equipment within the system and for one or more consequence types (as many as can be accommodated within the available timeframe), rather than one piece of equipment with all consequence types.

B.1. Consequence Types

B.1.1. Production Loss

In analyzing a civil-work facility or process, such as a navigation lock, for risk-analysis purposes, an important unsatisfactory-performance consequence is production loss due to unavailability of some critical pieces of equipment for the process. The production loss can be due to a complete shutdown of the process or limited production. Also, it includes delays as a result of failures. The unsatisfactory performance of an equipment can have an impact on both the upstream and downstream ends of the process. Therefore, the assessment of this consequence type requires the definition of upstream and downstream production losses that should be considered in the assessment. The influence domain, therefore, needs to be defined. It is possible to consider only the immediate (or direct) unsatisfactory-performance consequence of an equipment failure; however, in this case care should be exercised in

interpreting the results of risk analysis. The production loss can be expressed in any convenient units, for example dollars or a volume unit such as cargo tonnage.

B.1.2. Property Damage

This consequence type includes the repair or replacement cost of a failed equipment, repair and replacement cost of other affected components by the unsatisfactory-performance, and damage to surrounding property within the facility. This consequence type does not include damages to other properties, business interruption of non-production nature, and legal expenses. The repair and replacement costs should include equipment transportation and installation using possibly new technologies, possible modification of linked components, and cost of expediting the repair or replacement. This type of consequence can be expressed in monetary value. However in certain applications, it might be more convenient to use other units, such as the retail value of the replaced piece of equipment. By normalizing this consequence cost by the failed equipment retail price, the effect of currency changes can be reduced, although technology advances can be a complicating factor in using such a unit.

B.1.3. Flood Inundation

Dam failure can have various consequences some of which can be significant that can include loss of life, injuries, property damage, and ecological effects, various types of environmental damage, and social and cultural impacts. The property damage can be to residential, commercial, industrial, and agricultural structures and systems. This section was taken and abbreviated from (USACE 1997). Each system failure that can arise has a consequence. A consequence from a failure can be many different things. A failure could cause economic damage such as loss of capital, loss of property, and adverse publicity. It could also result in more serious events such as environmental damage, injury or loss of human lives, or public endangerment. Consequence estimations are formed from either events in past history or on educated guesses. Consequence assessment is discussed in Appendix B.

A floodplain is defined by the American Geological Institute as the portion of a river valley adjacent to the river channel which is built of sediments during the present regimen of the stream and which is covered with water when the river overflows its banks at flood stages. The floodplain is a level area near the river channel. Clearly, the floodplain is an integral and necessary component of the river system. If a climate change or land use change occurs, then the existing floodplain may be abandoned and new floodplain construction begins. Sediment is deposited when the stream flow overtops the banks; this occurs approximately every 1.5 to 2 years in stable streams. The floodplain extends to the valley walls. In engineering, floodplains are often defined by the water surface elevation for a design flood, such as the 100- or 200-year flood.

Changes in the natural floodplain development are caused by changes in sediment loads or water discharge. Increases in both the sediment and water discharge are often caused by land use changes, typically urbanization. Other causes include changes to the channel itself, such as straightening or relocating. Climatic changes can cause the current floodplain to be abandoned; however, this is seldom a concern for engineering as the time scale is geologic rather than engineering.

There are a number of mathematical models that simulate a dam breach of an earthen dam by overtopping. Simulation of a breach requires flow over the dam, flow through the breach, and flow down the dam face. The flow over the dam is typically modeled as weir flow. The breach shape is assumed in all models, either as a regular geometric shape or a most efficient breach channel shape where the hydraulic radius of the breach channel is maximized similar to stable channel design. The initial breach grows by collapse of the breach slopes, due to gravity and hydrodynamic forces, and erosion of the soil, typically modeled using sediment transport equations which have been developed for alluvial river channels. The most detailed model for outflow from a breach is based on an implicit finite difference solution of the complete one-dimensional unsteady flow equations.

Inundation mapping is generally carried out by determining the extent of the flooding over the current topography. The water surface elevation or stage, determined in the breach outflow modeling, is extended to all topographic points with the same elevation to determine the extent of inundation. The most effective way to develop these maps is to use a GIS system based on reliable topographic maps, such as the U.S. Geological Survey quadrangle series for the United States.

B.1.4. Other Consequence Types

Other consequences include loss of life, injuries, ecological effects, various types of environmental damage, and social and cultural impacts. This section focuses only on the loss of life as a result of dam breach.

The number of people at risk in the event of capacity exceedence or other uncontrolled release depends on the population within the inundation area and the conditions of release. A variety of scenarios are defined by the planning team to represent a range of modes of failure, given overtopping and other potential conditions of breaching. The term *scenario* as used here means, "a particular situation, specified by a single value for each input variable" (Morgan and Henrion, 1992). For each scenario, specific characteristics of the release are defined, and quantitative characteristics of downstream effects are estimated for economic cost and loss of life.

For estimating the characteristics of downstream effects, a fluvial hydraulics model possibly combined with a dam breach analysis is used to forecast depths and extents of flooding. With this information, the economic affect on structures and facilities can be estimated, as can the environmental effect on downstream ecosystems. The number of people at risk, however, depends on additional considerations. These include the time of day and season of the year at which the release occurs, rate of water rise, available warning time and effectiveness of evacuation plans, and changes in downstream land use (e.g., Bowles, 1990). An empirical review of uncontrolled releases at other dams and of levee overtoppings provides an initial basis for estimating population at risk under the various scenarios. Nevertheless, the quantitative historical record of dam failures is small, and any particular project will have characteristics which differ in important ways from those of the database.

DeKay and McClelland (1993) present a quantitative expression for estimating loss of life in dam failures, based on statistical analysis of empirical data related to severe flooding (see also, USBR, 1989; and Hartford, 1995):

$$LOL = \frac{PAR}{1+13.277(PAR^{0.44})\exp\{0.750(WT) - 3.790(Force) + 2.223(WT)(Force)\}} \qquad \text{(B-1)}$$

in which, LOL = potential loss of life, PAR = population at risk, WT = warning time in hours, $Force$ = forcefulness of flood water (1 for high force, 0 for low force). The PAR is defined as the number of people within three hours travel time of the flood wave, and includes not just those exposed to "treacherous flood waters," but all risk of "getting their feet wet." The empirical equation is statistically valid only for PAR's less than 100,000. An example calculation is shown in Figure B-1. For an example dam, the following values are assumed: PAR = 100,000, WT = 3 hours, and $Force$ = 0 and 1. The resulting values for LOL are 0.3 and 5 persons for $Force$ = 0 and 1, respectively.

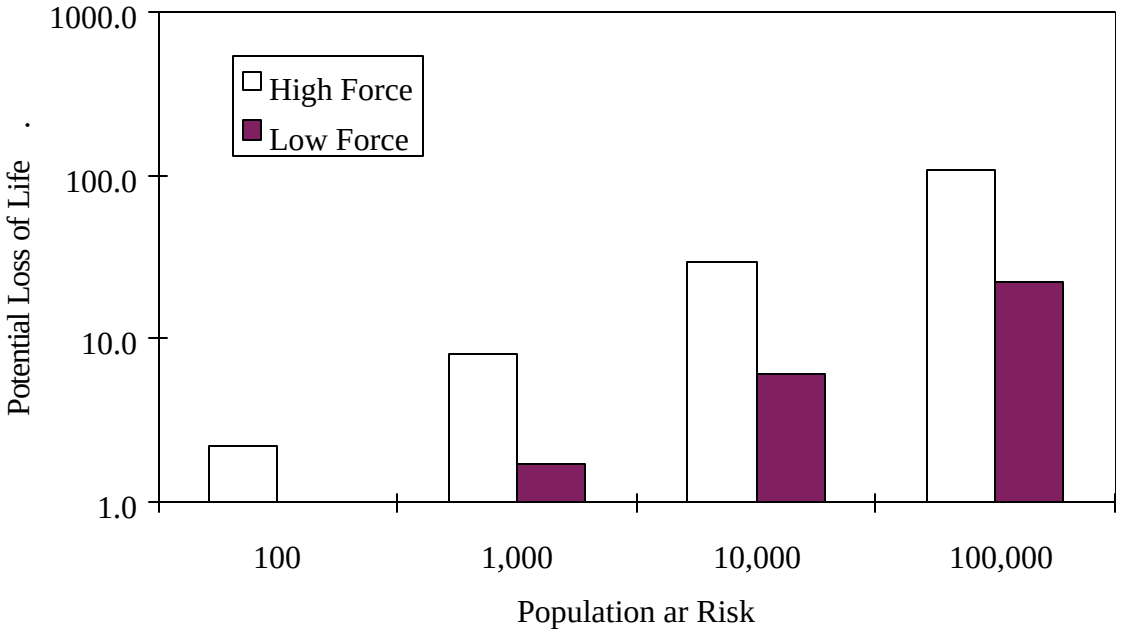

Figure B-1. Example Calculation of Potential Loss of Life for a Warning Time of One Hour.

The USBR (1989) suggests estimating population at risk (PAR) by applying an annual exposure factor to the number of residents in the flood plain. The annual exposure factor is the fraction of the year a typical individual spends at home. This factor ranges from about 0.6 to 0.8. The number of residents in the flood plain is estimated from census data, interviews with local planning officials, the number of homes in the area multiplied by the average number of residents per home, planning or cadastral maps, and house-to-house surveys. In most cases, the analysis must be augmented by consideration of facilities other than homes, such as schools, factories, and shopping centers.

The warning time (*WT*) in the above equation depends on the existing warning system. This is the time in hours before the arrival of flooding by which the "first individuals for each *PAR* are being warned to evacuate" (USBR, 1989). As a lower bound, warning time is sometimes taken as just the flood travel time (i.e., no warning is issued prior to loss of containment). This is thought appropriate for events such as earthquake induced failures, but conservative for hydrologically caused failures. The affect of warning time on loss of life also depends on the warning procedure (e.g., telephone chain calls vs. siren) and on the evacuation plan. Neither of these enter the above equation.

The forcefulness of flood waters (*Force*) in the above equation is treated as a dichotomous variable with value one for high force and zero for low force. "High force" means waters that are swift and very deep, typical of narrow valleys. "Low force" means waters that are slow and shallow, typical of broad plains. For cases in which the population resides in both topographies, the *PAR* is subdivided. DeKay and McClelland suggest that the *PAR* be divided into no more than two subgroups, because non-linearity in the above equation causes over estimation of loss of life as the *PAR* is subdivided.

B.2. Assessment of Consequences

Several methods can be used for assessing unsatisfactory-performance consequences. The methods include (1) unsatisfactory-performance and loss records, (2) unsatisfactory-performance databases, (3) cause-consequence diagrams, (4) event-tree analysis, (5) expert-opinion elicitation, and (6) questionnaires (Henley and Kumamoto 1981). In this section these methods are briefly introduced but not adapted for use in civil-work projects that is recommended for future work.

B.2.1. Unsatisfactory-Performance and Loss Records

Records of previous unsatisfactory performances should be examined to extract any useful consequence information. However since such records were possibly developed for other purposes than risk analysis, they might not contain the needed information or they might contain ill-defined and incomplete information. It is possible, in some cases, to utilize informed judgment to revise the consequence information from these records.

Information about the different types of unsatisfactory-performance consequences, e.g., production and property losses, might be available at different locations within an organization. Units within the organization that are responsible for production and administration aspects of the civil-work facilities need to be contacted for the purpose of soliciting consequence information.

B.2.2. Unsatisfactory-Performance Databases

Unsatisfactory-performance databases can be used to obtain information about consequences. Sometimes unsatisfactory-performance databases were not developed for risk analysis. Therefore, they might not contain the needed information or they might contain ill-defined and incomplete information for risk analysis. It is possible, in some cases, to utilize informed judgment to revise consequence information from them. Many organizational units are commonly interested in analyzing unsatisfactory

performances. Therefore, some of them might have such databases. If unsatisfactory-performance databases are not available in-house, unsatisfactory-performance databases developed by other civil-work facilities with similar functions or generic unsatisfactory-performance databases can be used. However, it is important to ensure that the collected consequence information is relevant to the process under investigation. If not, informed judgment can be utilized to revise the collected consequence information in order to make it more relevant.

Information on the different types of consequences, e.g., production and property losses, can be assessed using this approach if such databases are available. Since consequence assessment is a primary component to risk analysis, it is highly recommended to develop an unsatisfactory-performance database that include consequences in its fields, if such a database is not available.

B.2.3. Cause-Consequence Diagrams

Cause-consequence (CS) diagrams (Henley and Kumamoto 1981) were developed for the purpose of assessing and propagating the conditional effects of an unsatisfactory performance using a tree representation. The analysis according to CS starts with selecting a *critical event*. Critical events are commonly selected as convenient starting points for the purpose of developing the CS diagrams. For a given critical event, the consequences are traced using logic trees with event chains and branches. The logic works both backward (similar to fault trees) and forward (similar to event trees). The procedure for developing a CS diagram can be based on answering a set of questions at any stage of the analysis. The questions can include, for example, the following:

- Can this event lead to other unsatisfactory-performance events?
- What are the needed conditions for this event to lead to other events?
- What other components are affected by this event?
- What other events are caused by this event?
- What are the associated consequences with the other (subsequent) events?
- What are the occurrence probabilities of subsequent events or unsatisfactory-performance probabilities of the components?

The resulting CS tree can be used to compute the unsatisfactory-performance consequences for the possible unsatisfactory-performance scenarios (tree branches) with their occurrence probabilities. Then, the average unsatisfactory-performance consequence can be computed. Additional information about cause-consequence diagrams can be obtained from textbooks on reliability engineering such as Henley and Kumamoto (1981).

Event-tree analysis results in unsatisfactory-performance sequences (scenarios) with associated probabilities that can be useful in developing a cause-consequence diagram. The analysis is based on an inductive logic that moves forward in failing a system of interest. For example, starting with an initiating event questions such as "what might happen next and what are the associated probabilities" are asked. Therefore, the tree results by branching forward towards the unsatisfactory-performance of the system. The logic in event-tree analysis is similar to the cause-consequence diagrams, but without considering unsatisfactory-performance consequences. Also, it does not include any deductive (i.e., backward)

logic, whereas the cause-consequence analysis includes deductive logic by performing localized fault-tree analysis.

B.2.4. Formal Expert-Opinion Elicitation and Questionnaires

Expert-opinion elicitation is a formal process of obtaining information or answers to specific questions about certain issues that are needed to meet certain analytical objectives. The expert-opinion elicitation process is described in Chapter 2.

If expert-opinion elicitation is needed to assess unsatisfactory-performance consequences in risk analysis, a formal expert-opinion elicitation is highly recommended. However, sometimes a formal process is not possible due to a variety of reasons, such as the logistics of convening a meeting of all the experts at the same time. In this section, a procedure is suggested for performing expert-opinion elicitation through questionnaires. Additional information on construction and use of questionnaires based on social science is provided in Appendix C.

The main difficulty in designing questionnaires for the purpose of expert-opinion elicitation is that their design needs to ensure the following conditions:

- Communicating properly the statements of the questions of interest to the experts.
- Eliminating any ambiguity or vagueness in the statements of the questions and the anticipated responses.
- Eliminating any ambiguity or vagueness in how the responses should be expressed.
- Providing an efficient design that is complete, concise, clear and easy to follow.

Additional limitations on the use of questionnaires and experts are presented in Appendix C.

An approach similar to the formal expert-opinion elicitation process as described in Chapter 2, can be used for constructing and administering questionnaires. The needed steps are similar to the formal expert-opinion elicitation process with a primary difference being the design and testing of questionnaires. This step should be performed by a risk analysis team. For the selected issues, the questionnaires need to be designed so that each separately addresses a specific issue. However, similar issues can be addressed by the same questionnaire design with some changes in its contents. A questionnaire design should include the following components:

- Issue description
- Expert familiarization of the issue
- Aspects of the issue that should be considered in its assessment
- Aspects of the issue that should not be considered in its assessment
- Cause-consequence diagrams
- Anticipated response in content, units, presentation and style

The developed questionnaires need to be tested before their use in the expert-opinion elicitation process. The test group can be selected on the bases of their familiarity with the issues, the objectives of the study, availability and willingness to provide expedient responses.

Appendix C. Heuristics, Elicitation, Scoring and Aggregation

C.1. Introduction

The objective of this chapter to summarize expert-opinion elicitation methods, methods for combining expert opinion, and methods used in developing questionnaires in educational and psychological testing and social research.

C.2. Scientific Heuristics

The contemporary philosopher of science Hans Reichenbach (1951) made a distinction between "discovery" and "justification" in science. Discovery in science can be characterized as nonhomogenous, subjective and nonrational. It can be based on hunches, predictions, biases, and imaginations. It is the product of creativity that extends in the domain of the unknown knowledge to humankind. For example, everyone has seen the moon movement across the night sky and seen apples and other objects falling to earth, however, it took a Newton to realize the same physical laws underlay both phenomena. Newton's ideas were subjected to testing and validation using the scientific processes of justification. There is surely a difference between discovering ideas or phenomena and scientifically justifying them. The process of discovery and justification in science can be viewed as a rational consensus process that is based on empirical control (testing) and repeatability, i.e., the outcome of ideas should pass empirical testing by anyone, and should be repeatable by anyone. **Heuristics** is a process of discovery that is not necessarily structured.

Discovery is a form of scientific heuristics that does not entail a lot of structure and relies heavily on rules of thumb, subjectivity, and creativity. In order to be successful in its pursuit, it cannot approach issues at hand in orderly fashion, requiring a level of coherent disorder that must not reach to a level of disarray. However, subjectivity and disorder can lead to errors especially biases that are not intentional; although intentional or motivational biases can be present and should be targeted for elimination. Psychometric researchers such as Kahneman et al (1982) and Thys (1987) have studied this area extensively on the fundamental level and to understand its relation to expert opinions, respectively.

Heuristics are the product of four factors: (1) availability, (2) anchoring, (3) representativeness, and (4) control as shown in Figure C-1. For a given issue, **availability** is related to the ease with which individuals (including experts) can recall similar events or situations to this issue. Therefore, probabilities

of well-publicized events tend to be overestimated whereas probabilities of unglamorous events are underestimated.

Anchoring is the next factor in heuristics where subjects, i.e., individuals or experts, tend to start with an initial estimate and correct it to the issue at hand. However, the correction might not be sufficient. For example, high school kids guessed order of magnitude differences in estimating the product of the following two number sequences within a short period of time:

8x7x6x5x4x3x2x1 and 1x2x3x4x5x6x7x8

The differences can be attributed to performing the first few multiplications, establishing anchors, and estimating the final answers through extrapolation (Kahneman et al 1982).
Representativeness can affect conditional probability assessments. For example, individuals tend to intuitively evaluate the conditional probability P(A|B) by assessing the similarity between A and B. The problem with this assessment is that similarity is symmetric whereas conditional probabilities are not, i.e., the resemblance of A to B is the same as the resemblance of B to A; whereas P(A|B) does not equal P(B|A).

The **control** factor refers to the perception of subjects in that they can control or had control over outcomes related to an issue at hand. For example, Langer (1975) demonstrated that lottery ticket buyers demanded higher median prices for reselling their tickets to a ticket seeker if they had selected the ticket numbers than others who were given tickets with randomly selected numbers. The false sense of control contributed to increased believe in the value of their tickets.

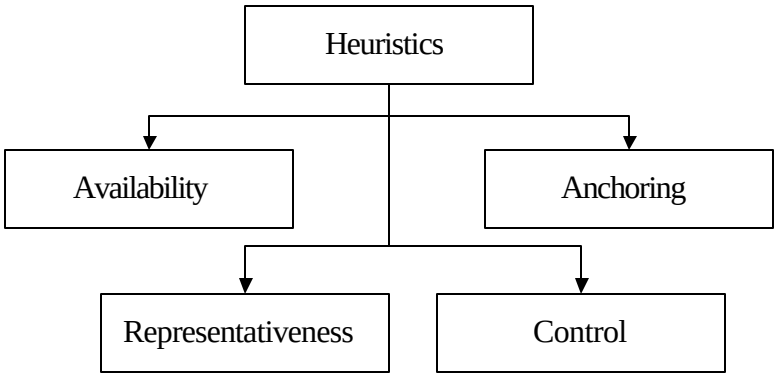

Figure C-1. Heuristics

Other sources of bias or error include (1) base-rate fallacy, (2) overconfidence. The base-rate fallacy arises as a result of using misguided, or misinformed subjects. A subject might rely on recent or popular information and unintentionally ignore the historic rate for an event of interest. The recent or popular information might make the subject biased towards a substantially different rate than the historic value. For example, a subject might assign a relatively large occurrence probability for a failure event of a component as a result of recent or highly popularized failure information despite a historically low failure probabilities for such components. The base rate which is low in this case should be combined with the new information (i.e., recent or highly popularized failure information) using Bayes' theorem resulting in relatively small change in the base rate. Overconfidence results in error and biases usually as a result of

poor calibration (Cooke 1991). Overconfidence especially common in assessing confidence intervals on an estimated value. Subjects tend to provide narrower confidence intervals compared to real intervals. Calibration can help in controlling overconfidence. Overconfidence also appears in assessing small (or large) probabilities, less than 0.01 or in some cases less than 0.1 (larger than 0.99 or in some cases less than 0.9). Subject calibration can help in reducing the effects of base-rate fallacy and overconfidence. A well-calibrated subject can be defined as an individual that would consistently produce an estimate that is in agreement with the corresponding true value. Subjects can be calibrated by providing them with feedback on their assessments in training-like sessions. Expert calibration was successfully performed for weather forecasting as was reported by Murphy and Daan (1984). The calibration process involves training subjects of probability concepts, error sources, biases, expectation, issue familiarization, aggregation methods, reporting and use of results (Alpert and Raiffa 1982, Murphy and Daan 1984, Winkler and Murphy 1968, and Ferrell, 1994).

Subjectively assessed probabilities should be examined carefully for any signs of error or inadequacy. Historically, such signs include (1) data spread, (2) data dependence, (3) reproducibility, and (4) calibration. It is common to have spread in subjectively assessed probabilities especially when dealing with low numbers (or large numbers for their complementary events). For example, the failure probability of a high-quality steel pipe (10-meter long) of a diameter at least 7.6 cm per hour was subjectively assessed by 13 experts (NRC 1975) as follows: 5E-6, 1E-6, 7E-8, 1E-8, 1E-8, 1E-8, 1E-8, 6E-9, 3E-9, 2E-9, 2E-10, 1E-10, and 1E-10. The Nuclear Regulatory Commission (NRC) used a value of 1E-10 with 90% confidence bounds of 3E-9 and 3E-12 in assessing the annual probability of core melt due to an earthquake. The following observations can be made based on these assessments:

1. The data have a spread of 5E-6 to 1E-10 which can be expressed as an upper-limit to lower-limit ratio of 10,000.
2. The adopted value corresponds to the smallest value in this spread.
3. The 90% confidence bounds contain only 5 value out of the 13 gathered values.

Data spread is common in dealing with low numbers. Data spread can be reduced by asking subjects with extreme views to justify their values, and re-eliciting the data to establish a consensus on a tighter spread; although, it is common that data spread is just reported and discussed in order to justify and adopt a value with associated confidence bounds.

Data dependence can arise as a result of pessimistic or optimistic subjects, i.e., consistently biased subjects that provides low or high values, respectively, in comparison to corresponding true (long-term) values. Statistical tests can be performed on data dependence as a result of using biased subjects as given by Cooke (1986).

Reproducibility of results can be examined by performing a bench mark study that would require several teams to perform independent analyses based on a common set of information about a systems. Then, the results of analyses are compared for spread in the form of ratios of maximum to minimum reported values by the teams. Several bench mark studies of this type were performed (for example, Amendola 1986, and Brune et al 1983). The Amendola (1986) study was structured in four stages using 10 teams

from different European countries to assess the failure probability of a feedwater system. The four stages are:

1. The first stage involved *blind*, independent evaluation by the teams as an initial probabilistic analysis without information sharing among the teams. The spread ratio in the resulting probabilities is 25 (i.e., 8E-4 and 2E-2).
2. Fault tree analysis was independently performed by the teams resulting in a spread ratio of 36. Afterwards the teams met to produce one fault tree, but could not agree on a common one.
3. In this stage, a common fault tree was assigned to the teams. The teams used their own data to produce the system failure probability. The spread ratio in the resulting probabilities is 9. This stage isolates the effect of data on the results.
4. In this stage, a common fault tree and data were given to the teams. The teams used their analytical tools to produce the system failure probability. The spread ratio in the resulting probabilities is about 1. This stage isolates the effect of the analytical tools.

Having data with small spread and without dependence, that have been reproduced by several teams, does not mean that the data are correct, it only increases our confidence in them. The process of calibration is closely tied to the process of result validation which is difficult since opinion elicitation is commonly associated with rare events that cannot be validated. Training of subjects, however, can be based on other events or issues in order to have calibrated subjects.

Example C-1. Information Communication for National Security Intelligence

The intelligence community is in the business of information collection. It is quite common that gathered information is marred with subjectivity, uncertainty, and perhaps irrelevance, and can be from non-reliable sources. The intelligence community is aware of and regularly deals with these problems. For example, the U. S. Defense Intelligence Agency (DIA) investigated uncertainty in intelligence information (Morris and D'Amore 1980), and provided a summary of various conceptual and analytical models for this purpose. The primary interest of the study was to assess uncertainty in projecting future force levels of the USSR. A secondary motive to the study was the failure to predict the fall of the Shah of Iran in 1979 (Cooke 1991).

The intelligence community widely used a reliability-accuracy rating system to communicate uncertainty as shown in Table C-1. However, Samet (1975) indicated that this system is not adequate since correspondents tend to emphasize information accuracy, and does not necessarily covey uncertainty attributed to source reliability. The DIA used the Kent chart as shown in Table C-2 to provide a quantitative interpretation of natural language expressions of uncertainty. As reported by Morris and D'Amore (1980), however, the Kent chart has been replaced by a direct use of probabilities.

Table C-1. Reliability and Accuracy Ratings in Intelligence Information (Morris and D'Amore 1980)

Source Reliability	Information Accuracy
A. Completely reliable	1. Confirmed
B. Usually reliable	2. Probably true
C. Fairly reliable	3. Possibly true
D. Not usually reliable	4. Doubtfully true
E. Unreliable	5. Improbable
F. Reliability cannot be judged	6. Accuracy cannot be judged

Table C-2. A Kent Chart (Morris and D'Amore 1980)

Likelihood order	Synonyms	Chances in 10	Percent
Near Certainty	Virtually (almost) certain, we are convinced, highly probable, highly likely	9	99 90
Probable	Likely We believe We estimate Chances are good It is probable that	8 7 6	60
Even chance	Chances are slightly better than even Chances are about even Chances are slightly less than even	5 4	40
Improbable	Probably not Unlikely We believe … not	3 2	10
Near impossibility	Almost impossible Only a slight chance Highly doubtful	1	1

Note: Words such as "perhaps," "may," and "might" will be used to describe situations in the lower ranges of when used without further modification, will generally be used only when a judgment is important but cannot be given an order of likelihood with any degree of precision.

C.3. Elicitation and Scoring Methods

This section provides a summary of various methods that can be used for elicitation of expert opinions. Also, methods for scoring or rating experts are presented. In order to increase the chances of success in using elicitation and scoring methods, Cooke (1991) provided suggested practices and guidelines. They were revised for the purposes of this study and are summarized as follows:

1. The issues or questions should be clearly stated without any ambiguity. Sometimes there might be a need for testing the issues or questions to ensure their adequate interpretation by others.
2. The questions or issues should be stated using appropriate format with listed answers, perhaps graphically expressed, in order to facilitate and expedite the elicitation and scoring processes.
3. It is advisable to test the processes by performing a dry run.
4. The analysts must be present during the elicitation and scoring processes.

5. Training and calibration of experts must be performed. Examples should be presented with explanations of elicitation and scoring processes, and aggregation and reduction of results. The analysts should avoid coaching the experts or leading them to certain views and answers.
6. The elicitation sessions should not be too long. In order to handle many issues, several sessions with appropriate breaks might be needed.

C.3.1. Elicitation Methods

C.3.1.1. Indirect Elicitation

The direct elicitation method is popular among theoreticians and was independently introduced by Ramsey (1931) and De Finetti (1937). The indirect method is based on betting rates by experts in order to reach a point of indifference among presented options related to an issue. The primary disadvantage of this method is the utility value of money is not necessary linear with the options presented to an expert, and the utility value of money is expert dependent.

Other indirect techniques were devised by researchers in order to elicit probabilities from probability-illiterate experts. For examples analysts have used time to first failure estimation or age at replacement for a piece of equipment as an indirect estimation of failure probability.

Example C-2. Betting Rates for Elicitation Purposes

Betting rates can be used to subjectively and indirectly assess the occurrence probability of an event A, called $p(A)$. According to this method, an expert E is hypothetically assigned a lottery ticket of the following form:

Expert E receives \$100 if A occurs. (C-1)

The interest hereafter is the value that the expert attaches to this lottery ticket. For an assumed amount of money \$$x$, that is less than \$100, the expert is asked to trade the ticket for the \$$x$ amount. The amount \$$x$ is increased incrementally until a point of indifference is reached, i.e., the lottery ticket has the same value as the offered \$$x$ amount. The \$$x$ position is called **certainty equivalent** to the lottery ticket.

Assuming the expert to be a rational and unbiased agent, the \$$x$ position which is certainty equivalent to the lottery ticket, provides an assessment of an expectation. The expected utility of the lottery ticket can be expressed as

$$\text{Expected utility of the lottery ticket} = \$100k\big(p(A)\big) \quad \text{(C-2)}$$

Where $p(A)$ = the occurrence probability of A, and k = a constant that represent the utility for money as judged by the expert. The utility of money can be a nonlinear function of the associated amount. At the certainty equivalent position, \$$x$ has a utility of k\$$x$ which is equivalent to the expected utility of the lottery ticket as shown in Eq. C-2. Therefore, the following condition can be set:

$$\$100k\big(p(A)\big) = k\$x \quad \text{(C-3)}$$

Solving for $p(A)$ produces

$$p(A) = \frac{\$x}{\$100} \tag{C-4}$$

The utility of money in the above example was assumed to be linear; whereas empirical evidence suggests that it is highly nonlinear. Galanter (1962) constructed Table C-3 by asking subjects the following question:

> *"Suppose we give you x dollars; reflect on how happy you are. How much should we have given you in order to have made you twice as happy?"*

The following utility function U was developed based on these data:

$$U(x) = 3.71x^{0.43} \tag{C-5}$$

It is evident that the willingness of people to run a risk does not grow linearly with an increased amount x. Similar tests were performed for losses of money and their relationship to unhappiness, but were inconclusive as subjects found the questions "*too difficult.*" Therefore, betting rates might not be suitable for failure probability assessment especially since such probabilities are commonly very small.

Table C-3. Money Required to Double Happiness (Galanter 1962)

	Twice as Happy	
Given *X*	Mean	Median
$10	$53.	$45
$20	$538	$350
$1000	$10,220	$5,000

C.3.1.2. Direct Method

This method elicit a direct estimate of the degree of belief of an expert on some issue. Despite its simple nature, this method might produce the worst results especially from experts who are not familiar with the notion of probability. Methods that fall in this category are Delphi method and the nominal group technique. The Delphi technique as described in detail in Chapter 1 allows for no interaction among the elicited expert before rending opinions. Variations to this method were used by engineers and scientist by allowing varying levels of interactions that range from limited interaction to complete consensus building as described in Chapter 2. The nominal group technique allows for a structured discussion after the experts have provided initial opinions. The final judgement is made individually on a second cycle of opinion elicitation and aggregated mathematically similar to the Delphi method (Gustafson et al 1973, and Morgan and Henrion 1992). Lindley (1970) suggested a method that is based on comparing an issue to other familiar issues with known answers. This comparative examination has been proven to be easier for experts than directly providing absolute final answers. For example, selected experts that are familiar with an event A and its occurrence probability $p(A)$ are used to subjectively assess the occurrence probability of event B. We are interested in assessing the occurrence probability of event B that is not of the same probability familiarity to the experts as $p(A)$. Experts are asked to assess the relative occurrence of B to A, say 10 times as frequent. Therefore, $p(B) = 10p(A)$.

C.3.1.3. Parametric Estimation

Parametric estimation is used to assess the confidence intervals on a parameter of interest such as the mean value. The estimation process can be in the form of a two-step procedure as follows (Preyssl and Cooke 1989):

1. Obtain a median estimate of a probability (m), and
2. the probability (r) that the true value will exceed 10 times the median value (m).

The m and r values can be used to compute the 5% and 95% confidence bounds as $\frac{m}{k_{0.95}}$ and $m(k_{0.95})$, respectively, where

$$k_{0.95} \approx \frac{\exp(-0.658)}{z_{1-r}} \qquad \text{(C-6)}$$

in which z_{1-r} is the $(1-r)^{th}$ quantile value of the standard normal probability distribution. Experts were found to like and favor two-step methods for dealing with uncertainty.

C.3.2. Scoring Methods

Scoring methods can be used to assess the information reliability (or quality) provided by experts through an expert-opinion elicitation process. The resulting scores can be used, for example, to determine weight factors for combining expert opinions if needed. Several methods can be used for this purpose as described in this section.

C.3.2.1. Self Scoring

According to this method, each expert provides a self assessment in the form of a confidence level for each probability or answer provided for an issue. The primary disadvantages of this method are bias and overconfidence that can result in inaccurate self assessments.

C.3.2.2. Collective Scoring

According to this method, each expert provides assessments of other experts in the form of confidence levels in their provided probabilities or answers related to an issue. The primary disadvantages of this method are bias and non-reproducibility.

C.3.2.3. Entropy and Discrepancy Measures

Experts can be asked to provide a probability mass function that is associated with all possible values for an issue of interest such as occurrence probability of an event. Assuming that there are m possible values, the probability assignment by an expert can be expressed as $p(i)$, $i = 1, 2, \ldots, m$, and the Entropy $H(P)$ measure can be computed as an uncertainty measure. The Entropy measure (Ayyub 1999, and Klir and Folger 1988) takes values from 0 to 1. Its value is zero if $p(i) = 1$, and one for equally likely outcomes of $p(i) = 1/m$ for all i. It is desirable to obtain an assessment from experts with the lease Entropy value from a set of experts with equal circumstances and conditions; although equal circumstances and conditions might not be attainable. The corresponding true values of the probability mass function can be expressed as $s(i)$, $i = 1, 2, \ldots, m$; therefore, a discrepancy measure can be

defined to account for circumstances and conditions that are not the same (Cooke 1991). The primary features of this method are its higher analytical complexity and information needs in comparison to previous methods. These features can hinder its use in some cases, and make a most suited method in other cases.

C.4. Combining Expert Opinions

In some applications, expert opinions in the form of subjective probabilities of an event need to be combined into a single value and perhaps confidence intervals for their use in probabilistic and risk analyses. Cooke (1991) and Rowe (1992) provided a summary of methods for combining expert opinions. The methods can be classified into consensus methods and mathematical methods (Clemen 1989, and Ferrell 1985). The mathematical methods can be based on assigning equal weights to the experts or different weights. This section provides a summary of these methods.

C.4.1. Rational Consensus

The use of expert opinions in engineering needs to be performed as a part of a rational consensus process. A rational consensus process should meet the following requirements (Cooke 1991):

1. Reproducibility. The details of collection, gathering and computation of results based on expert opinions need to be documents to a level that make them reproducibility by other expert peers. This requirement is in agreement with acceptable scientific research.
2. Accountability. Experts, their opinion and sources should be identified for reference by others as expert unanimity might degrade outcomes of consensus building and expert-opinion elicitation.
3. Empirical Control. Expert opinion should be susceptible to empirical control if possible at a minimum for selected practical cases. Empirical control can be performed by comparing results of expert-opinion elicitation with observations for selected control issues. This empirical control might not be possible in some situation, but it is in agreement with acceptable scientific research.
4. Neutrality. The method of eliciting, evaluating and combining expert opinions should encourage experts to state their true opinions. For example, the use of the median to aggregate expert opinions violates this requirements since the median rewards centrally compliant experts. Methods of using weighted averages of opinions based on self weights or weights by experts of each other have the same fallacy.
5. Fairness. The experts should be equally treated during the elicitation and for the purposes of processing the observations.

C.4.2. Consensus Combination of Opinions

A consensus combination of opinion is arrived at through a facilitated discussion among the experts to some agreeable common values with perhaps a confidence interval or outer quartile values. The primary shortcomings of this method are (1) socially reinforced irrelevance or conformity within a group, (2) dominance of strong-minded or strident individuals, (3) group motive of quickly reaching an agreement, and (4) group reinforced bias due to common background of group members. The

facilitator of an expert-opinion elicitation session should play a major role in reducing group pressure, individual dominance, and biases.

C.4.3. Percentiles for Combining Opinions

A p-percentile value (x_p) for a random variable based on a sample was defined in Appendix A as the value of the parameter such that p% of the data is less or equal to x_p. On the basis of this definition, the median value is considered to be the 50 percentile value. Aggregating the opinions of experts can be based on computing the 25, 50 and 75 percentile values of the gathered opinions. The computation of these values depends on the number of experts providing opinions. Table A-2 provides a summary of the needed equations for 4 to 20 experts. For example, 7 experts provided the following subjective probability of an event that are sorted in decreasing order:

1.0E-02 , 5.0E-03 , 5.0E-03 , 1.0E-03 , 1.0E-03 , 5.0E-04 , and 1.0E-04.

The median and arithmetic quartile points according to Table A-2 are respectively given by

Median = 1.0E-03,
25 percentile = 5.0E-03, and
27 percentile = 7.5E-04.

C.4.4. Weighted Combinations of Opinions

French (1985) and Genest and Zidek (1986) provided summaries of various methods for combining probabilities and example uses. For E experts with the i^{th} expert providing a vector of n probability values, p_{1i}, p_{2i}, …, p_{ni}, for sample space outcomes A_1, A_2, …, A_n, the E expert opinions can be combined using weight factors w_1, w_2, …, w_E that sum up to one using one of the following selected methods: weighted arithmetic average, weighted geometric average, weighted harmonic average, maximum value, minimum value, and generalized weighted average as provided in detail by Ayyub (1999).

C.4.5. Opinion Aggregation Using Interval Analysis, Fuzzy Numbers and Uncertainty Measures

Sometimes it might be desirable to elicit probabilities and/or consequences using linguistic terms as shown in Table A-1 for linguistic probabilities. Linguistic terms of this type can be translated into interval or fuzzy numbers. Intervals are considered as a special case of fuzzy numbers which are in turn a special case of fuzzy sets. Fuzzy arithmetic can be used to develop methods for aggregating expert opinions that are expressed in linguistic terms. This aggregation procedure returns the uncertainties in the underlying opinions by obtaining a fuzzy combined opinion. Also, uncertainty measures can be used to aggregate expert opinions based on principles of maximizing uncertainty as was demonstrated by Lai and Ayyub (1994). The needed analytical tools for this purpose are recommended for further development in future USACE studies (Ayyub 1999).

C.5. Methods of Educational and Psychological Testing, and Social Research

C.5.1. Standards for Educational and Psychological Testing

Credible behavioral testing and research adhere to the Standards for Educational and Psychological Testing (SEPT) published by the American Psychological Association (1985). The objective of this section is to summarize these standards, to determine how they relate to expert-opinion elicitation, and to identify any pitfalls in expert-opinion elicitation based on examining these standards.

Sacman (1975) from the RAND corporation provided a highly critical critique of the Delphi methods based on its compliance with the SEPT among other scientific and research practices. This critique is valuable, and is summarized herein since its applicability in some concerns goes beyond the Delphi methods to other expert-opinion elicitation methods.

Sacman (1975) found that conventional Delphi applications (1) often involve crude questionnaire designs, (2) do not adhere to proper statistical practices of sampling and data reduction, (3) do not provide reliability measures, (4) do not define scope, populations, and limitations, (5) provide crisply stated answers to ambiguous questions, (6) involve confusing aggregation methods of expert opinions with systematic predictions, (7) inhibit individuality, encourage conformity, and penalize dissidents, (8) reinforce and institutionalize early closure on issues, (9) can give an exaggerated illusion of precision, and (10) lack professional accountability. Although his views are sometimes overstated, they are still useful in highlighting pitfalls and disadvantages of the Delphi method. The value of the Delphi method comes from its initial intended uses as a heuristic tool, not a scientific tool, for exploring vague and unknown future issues that are otherwise inaccessible. It is not a substitute to scientific research.

According to the SEPT, a test involves several parties as follows: (1) test developer, (2) test user, (3) test taker, (4) test sponsor, (5) test administrator, and (6) test reviewer. In expert-opinion elicitation studies, similar parties can be identified. The SEPT provide a criteria for the evaluation of tests, testing practices, and the effects of test use. The SEPT provide a frame of reference to supplement professional judgement for assessing the appropriateness of a test application. The standard clauses of the SEPT are classified and identified as (1) *primary standards* that should be met by all tests, and (2) *secondary standards* that are desirable as goals but are likely to be beyond reasonable expectation in many situations. The SEPT consist of four sections as follows:

Part I. Technical Standards for Test Construction and Evaluation
Part II. Professional Standards for Test Use
Part III. Standards for Particular Applications
Part IV. Standards for Administrative Procedures

These SEPT parts are described in subsequent section as they relate to expert-opinion elicitation.

Part I. Technical Standards for Test Construction and Evaluation

Part I of the SEPT provides standards for test construction and evaluation that contain standards for validity, reliability, test development, scaling, norming, comparability, equating, and publication.

The validity consideration of the SEPT covers three aspects: (1) construct-related evidence, (2) content-related evidence, and (3) criterion-related evidence. The construct-related evidence primarily focuses on the test score appropriateness in measuring the psychological characteristic of interest. In these guidelines, expert-opinion elicitation deals with occurrence likelihood and consequences. The corresponding test scores can be selected as probabilities and consequence units such as dollars. The use of these scores does meet the validity standards of SEPT in terms of a construct-related evidence. The content-related evidence requires that the selected sample is representative of some defined universe. In the context of expert-opinion elicitation, experts should be carefully selected in order to meet the content-related evidence. The criterion-related evidence needs to demonstrate that the test scores are related to a criterion of interest in the real world. In the context of expert-opinion elicitation, the estimated occurrence probabilities and consequences need to be related to corresponding real, but unknown, values. This criterion-related evidence for validity is in agreement with the validation concept in the AIAA Guide for Verification and Validation of Computational Fluid Dynamics Simulations (AIAA 1998). The last consideration in validity is *validity generalization* that was reported in the form of the following two uses: (1) to draw scientific conclusions, and (2) to transport the result validity from one case to another. In the context of expert-opinion elicitation, validity generalization based these two uses might be difficult to justify. Selected primary validity standards, most related to expert-opinion elicitation are shown in Table C-4. They were taken from the 1997 draft revision of the SEPT is posted on the World Wide Web site of the American Psychological Association.

The reliability consideration of the SEPT deals with measurement errors due to two primary sources: (1) variations from one subject to another that are subjected to the same conditions and provided with the same background information, and (2) variations from one occasion to another by a specified subject. The tools that are needed to estimate the reliability of the scores, and test measurement errors are dependent on the error type. Statistical methods can be used for this purpose. In the context of expert-opinion elicitation, this reliability consideration requires aggregation procedures of expert opinions to include measures of central tendency, biases, dispersion, correlation, variances, standard error of estimates, spread of scores, sample sizes, and population definition.

Part I of the SEPT requires that tests and testing programs should be developed on a sound scientific basis. The standards puts the responsibility on the test developers and publishers to compile evidence bearing on a test, decide which information is needed prior to test publication or distribution and which information can be provided later, and conduct the necessary research.

The scaling, norming, comparability, and equating considerations in the SEPT deal with aggregation and reduction of scores. The documentation of expert-opinion elicitation should provide experts and users with clear explanations of the meaning and intended interpretation of derived score scales, as well as their limitations. Measurement scales and aggregation methods with their limitations, that are used for reporting scores, should be clearly described in expert-opinion elicitation documents. The documents should also include clearly defined populations that are covered by the expert-opinion elicitation

process. For studies that involve score equivalence or comparison and equating of findings, detailed technical information should be provided on equating methods or other linkages and on the accuracy of equating methods.

Administrators of a test should publish sufficient information on the tests in order for qualified users and reviewers to reproduce the results and/or assess the appropriateness and technical adequacy of the test.

Part II. Professional Standards for Test Use

Part II of the SEPT provides standards for test use. Users of the results of a test should be aware of methods used in planning, conducting and reporting the test in order to appreciate the limitations and scope of use of the test. Documented information on validity and reliability of test results as provided in Part I of the SEPT should be examined by the users for this purpose.

This part also deals with clinical testing, educational and psychological testing at schools, test use in counseling, employment testing, professional and occupational licensure and certification, and program evaluation. These standards have minimal relevance to expert-opinion elicitation.

Part III. Standards for Particular Applications

Part III of the SEPT provides standards for testing linguistic minorities and people with handicapping conditions. These standards have minimal relevance to expert-opinion elicitation.

Part IV. Standards for Administrative Procedures

Part IV of the SEPT provides standards for test administration, scoring, reporting, and rights of test takers. This part requires that tests should be conducted under standardized and controlled conditions similar to conducting experimental testing. Standardized and controlled conditions enhance the interpretation of test results by increasing the interpretation quality and effectiveness. Also this part deals with access to test scores, i.e., test security, and cancellation of test scores because of test irregularities.

Table C-4. Selected Validity Standards from the Standards for Educational and Psychological Testing (APA 1997)

1997 Draft SEPT Standard	Standard Citation	Relationship to Expert-opinion elicitation
1.1	A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.	Definition of issues for expert-opinion elicitation.
1.2	The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described.	Definition of issues for expert-opinion elicitation.
1.3	If validity for some common or likely interpretation has not been investigated, or is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.	Definition of issues for expert-opinion elicitation.
1.4	If a test is used in a way other than those recommended, it is incumbent on the user to justify the new use, collecting new evidence if necessary.	Definition of issues for expert-opinion elicitation.
1.5	The composition of any sample of examinees from which validity evidence is obtained should be described in as much detail as is practicable, including major relevant sociodemographic and developmental characteristics.	Selection of and training of experts.
1.7	When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented.	Selection of and training of experts, and definition of aggregation procedures of expert opinions.

C.5.2. Methods of Social Research

Social research concerns itself with gathering data on specific questions, issues, or problems of various aspects of society, and thus helps humans to understand society. Social study has evolved to social science especially in the field of sociology where there are three primary schools of thought (Bailey 1994): (1) humans have free will, and thus no one can predict their actions and generalize about them (the Wilhelm Dilthey school of the 19th century), (2) social phenomena are orderly and can be generalized, and they adhere to underlying social laws that need to be discovered through research similar to physical laws (the Emile Durkheim methods of *positivisim*), and (3) social phenomena are the product of free-will human volitional actions that are not random and can be predicted by understanding the human rational behind them (an intermediate school of thought of Max Weber). The stages of social research can be expressed in a circle of five stages as shown in Figure C-2 to allow for feedback in redefining the hypothesis in the first stage.

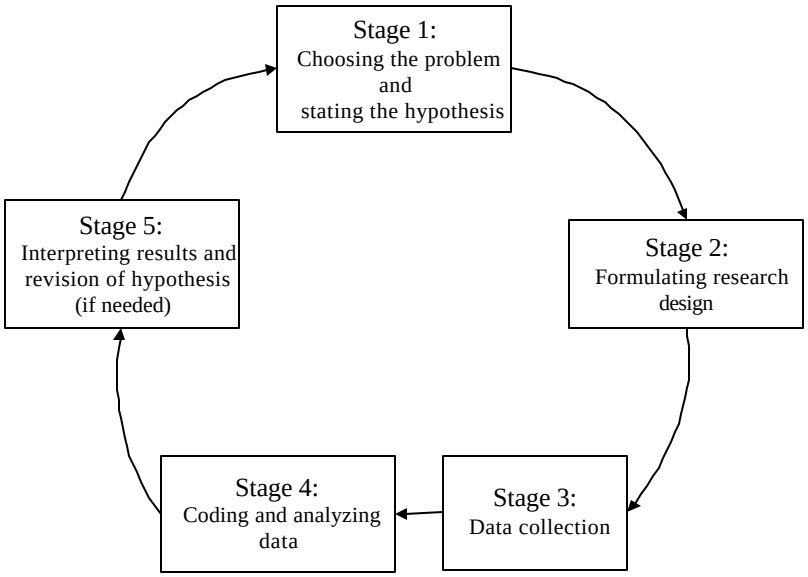

Figure C-2. Stages of Social Research

The construction and use of questionnaires is common and well developed in social research. Experiences from this field might be helpful to expert-elicitation developers, facilitators and administrators. The construction of a questionnaire should start by defining its relevance at the following three levels:

1. Relevance of the study to the subjects: It is important to communicate the goal of the study to the subjects, and establish its relevance to them. Establishing this relevance would make them stake holders and thereby increase their attention and sincerity levels.
2. Relevance of the questions to the study: Each question or issue in the questionnaire needs to support the goal of the study. This question-to-study relevance is essential to enhancing the reliability of collected data from subjects.
3. Relevance of the questions to subjects: Each question or issue in the questionnaire needs to be relevant to each subject especially when dealing with subjects of diverse views and backgrounds.

The following are guidelines on constructing questions and stating issues:

1. Each item on the questionnaire should include one question. It is a poor practice to include two questions in one.
2. Question or issue statements should not be ambiguous. Also, the use of ambiguous words should be avoided. In expert-opinion elicitation of failure probabilities, the word "failure" might be vague or ambiguous to some subjects. Special attention should be given to its definition within the context of each issue or question.
3. The level of wording should be kept to a minimum. Long questions should be avoided. Also, the choice of the words might affect the connotation of an issue especially by different subjects. The words should be selected carefully that meet the goal of the study in a most reliable manner.

4. The use of factual questions is preferred over abstract questions. Questions that refer to concrete and specific matters result in desirable concrete and specific answers.
5. Questions should be carefully structured in order to reduce biases of subjects. Questions should be asked in a neutral format, sometimes more appropriately without lead statements.
6. Sensitive topics might require stating questions with lead statements that would establish supposedly accepted social norms in order to encourage subjects to answers the questions truthfully.

Questions can be classified into *open-ended questions* and *closed-ended questions*. A closed-ended question limits the possible outcomes of response categories, can provide guidance to subjects thereby making it easier to the subjects, provides complete answers, allows for dealing with sensitive or taboo topics, allows for comparing the responses of subjects, and produces answers that can be easily coded and analyzed; but can be misleading, allows for guess work by ignorant subjects, can lead to frustration due to subject perception of inappropriate answer choices, limits the possible answer choices, does not allow for detecting variations in question interpretation by subjects, results in artificially small variations in responses due to liming the possible answers, and can be prone to clerical errors by subjects in unintentionally selecting wrong answer categories. An open-ended question does not limit the possible outcomes of response categories, is suitable for questions without known answer categories, is suitable for dealing with questions with too many answer categories, is preferred for dealing with complex issues, and allows for creativity and self expression; but can lead to collecting worthless and irrelevant information, can lean to non-standardized data that cannot be easily compared among subjects, can produce data that are difficult to code and analyze, requires superior writing skills, might not communicate properly the dimensions and complexity of the issue, can be demanding on the time of subjects, and can be perceived as difficult to answer and thereby discourages subjects from responding accurately or at all.

The format, scale and units for the response categories should be selected to best achieve the goal of the study. The minimum number of questions and question order should be selected with the following guidelines: (1) sensitive questions and open-ended questions should be left to the end of the questionnaire, (2) the questionnaire should start with simple questions and questions that are easy to answer, (3) a logical order of questions should be developed such that questions at the start of the questionnaire feed needed information into questions at the end of the questionnaire, (4) questions should follow other logical orders that are based on time-sequence or process related, (5) the order of the questions should not lead or set the response, (6) reliability-check questions that are commonly used in pairs (stated positively and negatively) should be separated by other questions, (7) questions should be mixed in terms of format and type in order to maintain the interest of subjects, and (8) the order of the questions can establish a *funnel* by starting with general questions following by more specific questions within several branches of questioning, this funnel technique might not be appropriate in some applications and its suitability should be assessed on case by case basis.

The final stage of developing a questionnaire is writing a cover letter or introductory statement, instructions to interviewers, subjects or facilitators, precoding, and pretesting. The introductory statement should provide the goal of the study and establish relevance. The instructions should provide guidance on expectations, completion of questionnaire, and reporting. Precoding assigns numerical

values to responses for the purpose of data analysis and reduction. Pretesting should be administered to a few subjects for the purpose of identifying and correcting flaws.

Some of the difficulties or pitfalls of using questionnaires, with suggested solutions or remedies, include the following (Bailey 1994):

1. Subjects might feel that the questionnaire is not legitimate and has a hidden agenda. A cover letter or a proper introduction of the questionnaire is needed.
2. Subjects might feel that the results will be used against them. Unnecessary sensitive issues and duplicate issues should be removed. Sometimes assuring a subject's anonymity might provided the needed remedy.
3. Subjects might refuse to answer questions on the basis they've done their share with questionnaires or tired of "being a guinea pig." Training and education might be needed to create the proper attitude.
4. A "sophisticated" subject that participated in many studies thereby developed an attitude of questioning the structure of the questionnaire, test performance, and result use might require a "sampling around" to find a replacement subject.
5. A subject might provide "normative" answers, i.e., answers that the subject thinks that they are being sought. Unnecessary sensitive issues and duplicate issues should be removed. Sometimes assuring a subject's anonymity might provided the needed remedy.
6. Subjects might not want to reveal their ignorance and appear perhaps stupid. Emphasizing that there are no correct or wrong answers, and assuring a subject's anonymity might provide the needed remedy.
7. A subject might think that the questionnaire is a waste of time. Training and education might be needed to create the proper attitude.
8. Subjects might feel that a question is too vague and cannot be answered. The question should be restated so that it is very clear.