NUREG/CR–5424
LA–11667–MS

Eliciting and Analyzing Expert Judgment

A Practical Guide

Prepared by M. A. Meyer, J. M. Booker

Los Alamos National Laboratory

Prepared for
U.S. Nuclear Regulatory Commission

AVAILABILITY NOTICE

Availability of Reference Materials Cited in NRC Publications

Most documents cited in NRC publications will be available from one of the following sources:

1. The NRC Public Document Room, 2120 L Street, NW, Lower Level, Washington, DC 20555

2. The Superintendent of Documents, U.S. Government Printing Office, P.O. Box 37082, Washington, DC 20013-7082

3. The National Technical Information Service, Springfield, VA 22161

Although the listing that follows represents the majority of documents cited in NRC publications, it is not intended to be exhaustive.

Referenced documents available for inspection and copying for a fee from the NRC Public Document Room include NRC correspondence and internal NRC memoranda; NRC Office of Inspection and Enforcement bulletins, circulars, information notices, inspection and investigation notices; Licensee Event Reports; vendor reports and correspondence; Commission papers; and applicant and licensee documents and correspondence.

The following documents in the NUREG series are available for purchase from the GPO Sales Program: formal NRC staff and contractor reports, NRC-sponsored conference proceedings, and NRC booklets and brochures. Also available are Regulatory Guides, NRC regulations in the *Code of Federal Regulations*, and *Nuclear Regulatory Commission Issuances*.

Documents available from the National Technical Information Service include NUREG series reports and technical reports prepared by other federal agencies and reports prepared by the Atomic Energy Commission, forerunner agency to the Nuclear Regulatory Commission.

Documents available from public and special technical libraries include all open literature items, such as books, journal and periodical articles, and transactions. *Federal Register* notices, federal and state legislation, and congressional reports can usually be obtained from these libraries.

Documents such as theses, dissertations, foreign reports and translations, and non-NRC conference proceedings are available for purchase from the organization sponsoring the publication cited.

Single copies of NRC draft reports are available free, to the extent of supply, upon written request to the Office of Information Resources Management, Distribution Section, U.S. Nuclear Regulatory Commission, Washington, DC 20555.

Copies of industry codes and standards used in a substantive manner in the NRC regulatory process are maintained at the NRC Library, 7920 Norfolk Avenue, Bethesda, Maryland, and are available there for reference use by the public. Codes and standards are usually copyrighted and may be purchased from the originating organization or, if they are American National Standards, from the American National Standards Institute, 1430 Broadway, New York, NY 10018.

DISCLAIMER NOTICE

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expresed or implied, or assumes any legal liability of responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights.

NUREG/CR--5424
TI90 006165

4

Eliciting and Analyzing Expert Judgment

A Practical Guide

Manuscript Completed: December 1989
Date Published: January 1990

Prepared by
M. A. Meyer, J. M. Booker

Los Alamos National Laboratory
Los Alamos, NM 87545

Prepared for
Division of Systems Research
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, DC 20555
NRC FIN A7225

Abstract

In this book we describe how to elicit and analyze expert judgment. Expert judgment is defined here to include both the experts' answers to technical questions and their mental processes in reaching an answer. It refers specifically to data that are obtained in a deliberate, structured manner that makes use of the body of research on human cognition and communication. Our aim is to provide a guide for lay persons in expert judgment. These persons may be from physical and engineering sciences, mathematics and statistics, business, or the military. We provide background on the uses of expert judgment and on the processes by which humans solve problems, including those that lead to bias. Detailed guidance is offered on how to elicit expert judgment ranging from selecting the questions to be posed of the experts to selecting and motivating the experts to setting up for and conducting the elicitation. Analysis procedures are introduced and guidance is given on how to understand the data base structure, detect bias and correlation, form models, and aggregate the expert judgments.

Contents

3 Background on Human Problem Solving and Bias 31

PART II--ELICITATION PROCEDURES

PART III--ANALYSIS PROCEDURES

List of Figures

(Figures listed in italic are not titled in the text.)

List of Tables

(Tables listed in italic are not titled in the text.)

List of Examples

Preface

In this book we describe how to elicit and analyze expert judgment. Expert judgment is defined here to include both the experts' answers to technical questions and their mental processes in reaching an answer. It refers specifically to data that are obtained in a deliberate, structured manner that makes use of the body of research on human cognition and communication.

The book was written at the request of the Nuclear Regulatory Commission. At the time, the Risk Analysis Division of the Nuclear Regulatory Commission was breaking new ground in gathering and using expert judgment in large probabilistic risk assessments. The book's content has been generalized to meet their needs and those of others using expert judgments.

Our aim in this book is to provide a guide for lay persons in expert judgment. These persons may be from the physical and engineering sciences, mathematics and statistics, business, or the military. Alternatively, they may be working in one of the fields that have traditionally relied on expert judgment, such as risk analysis, reliability analysis, decision analysis, operations research, or knowledge acquisition, a branch of artificial intelligence. To illustrate, people working in the sciences and the military have often remarked to us that they wish there was detailed information somewhere on how to gather or analyze expert judgment. Earlier, there was no source to provide the guidance that they, as lay persons, needed to design and conduct their own elicitations or analyses.

There are several reasons for there being little usable literature on how to elicit information from experts. First, the way in which these techniques are learned does not lend itself to publication. Interviewing techniques in anthropology, psychology, and sociology are usually taught in laboratory situations. Students in these fields typically learn by watching one of their professors and, then, by doing. Thus, even within these specialized fields, there are few sources on elicitation. Second, the sources that do exist are specialized for a particular discipline or situation and are not easily generalizable to others. Third, it is difficult to communicate elicitation techniques because the written medium is not well suited to conveying levels of information that are communicated through nonverbal means. Also, most of the mechanics of elicitation become automatic in the experienced practitioner and thus inaccessible for retrieval.

The prerequisites for understanding this book are minimal. Generally, the content is simple and procedural in orientation. For a few of the statistical sections, an understanding of the elementary concepts would be advantageous, but the procedures can be followed without this technical background. When jargon is used, it is defined. Also, a glossary is provided as an aid. The data sets used in the examples are referenced where appropriate. Those data sets that are not referenced have been artificially generated using data set structures and values similar to real data.

We gratefully acknowledge the Division of Risk Analysis, Office of Nuclear Regulatory Research, Nuclear Regulatory Commission for their financial support (FIN A7225) and encouragement of this effort. In particular, we are indebted to Dale Rasmussen for suggesting this work, James Johnson for overseeing the research, and

P.K. Niyogi for guiding the book's development. Special thanks also go to the many scientists at Los Alamos National Laboratory, Sandia National Laboratories, and Science Associates International Incorporated who participated in our studies of expert judgment. Without their contribution of time and expertise, this book would not have been possible. In addition, we extend our appreciation to our colleagues, Gary Tietjen and Thomas Bement, for their insightful reviews of the early drafts. We are also thankful for the constant support and encouragement provided by A. Juan. Finally, we are most grateful to Wilma Bunker, our expert in desk-top publishing, for her work in designing and editing the book.

M. A. Meyer
J. M. Booker

Los Alamos, New Mexico

PART I

INTRODUCTION TO EXPERT JUDGMENT

1

Introduction

In this book we provide guidance on how to gather and analyze expert judgment. Such a source has been lacking, particularly for those who are lay persons in the area of expert judgment. We have met many people working in the physical sciences, in government, or in the military services who were struggling to elicit or analyze expert judgment. Their jobs required that they perform these tasks, but there was little information available to assist them. This book is our response to their special needs. We describe elicitation and analysis procedures, when to use them, and how to perform them in a way that allows the lay readers to design methods suited to their own particular application. Those more experienced with expert judgment, typically those working in the fields of risk analysis, reliability analysis, operations research, decision analysis, or knowledge acquisition, may also find the book of interest because in it are mentioned techniques and options that could extend or improve their usual methods.

In this chapter we give a general introduction to expert judgment and to the situations in which it is used. We define expert judgment as it will be covered in this book and provide an overview of the methods that will be presented for eliciting and analyzing expert judgment. Lastly, we describe our philosophy of elicitation and analysis as background for understanding the methods presented.

What is Expert Judgment?

Expert judgment is data given by an expert in response to a technical problem. An expert is a person who has background in the subject area and is recognized by his peers or those conducting the study as qualified to answer questions. Questions are usually posed to the experts because they cannot be answered by other means. For instance, it may be impossible or impractical to measure the quantity of interest, such as the coal reserves in the United States, therefore a judgment is needed. Areas for expert judgment can vary from being an estimate of the number of homeless in the United States, to the probability of an occurrence of a nuclear reactor accident of a particular type, to an assessment of whether a person is likely to carry out a threatened act, to a description of the expert's thought processes in arriving at any of the above answers. Expert judgment has also been called *expert opinion, subjective judgment, expert forecast, best estimate, educated guess,* and most recently, *expert knowledge*. Whatever it is called, expert judgment is more than a guess. It is an informed opinion based on the expert's training and experience.

When Expert Judgment is Used

Expert judgment data has been used widely, especially in technical fields. It is a means of providing information when other sources, such as measurements, observations, experimentation, or simulation, are unavailable. Furthermore, it can be employed to supplement existing data when these are sparse, questionable, or only indirectly applicable. For example, in a new reactor-risk study called NUREG-1150 (U.S. NRC 1989), expert judgment was used where "experimental or observational data or validated computer models were not available or not widely agreed upon" (Ortiz, Wheeler, Meyer, and Keeney 1988:4).

Expert judgment has been gathered, specifically, to meet the following needs.

- **To provide estimates on new, rare, complex, or otherwise poorly understood phenomena.** Such phenomena have also been described as being *fuzzy* or of *high uncertainty*. One example would be estimates of the likelihood of the occurrence of rare reactor accidents. Another would be estimates of the safety of new automotive fuels as, for example, in the early eighties, when fuels such as liquid and compressed natural gas were being proposed for automotive use, but little was known about how safe they would be. To solve the problem a group of experts were convened to estimate the relative safety of the new fuels by considering their physical properties in combination with potential accident scenarios (Krupka, Peaslee, and Laquer 1983).
- **To forecast future events.** In general, when good actuarial data are unavailable, predicting future events or actions requires use of expert judgment. The experts are needed in order to adjust, sometimes radically, from the status quo or past patterns in making predictions. For instance, businesses often rely on expert judgment to forecast the market for their products. What the demand for various utilities will be in the United States may also come from experts' projections (Ascher 1978). A forecast of Soviet weapon capabilities for the year 2000 as a means of determining what the weapon needs of the United States will be for this same period can rely on the judgment of experts (Meyer, Peaslee, and Booker 1982).
- **To integrate or interpret existing data.** Expert judgment is frequently needed to organize qualitative information or mixtures of qualitative and quantitative data into a framework for making decisions. **Qualitative data** are any nonnumeric data, such as text on the expert's reasons for giving the answer, or the expert's answer encoded in descriptive categories or preference scales like poor, moderate, and good. **Quantitative data** are numeric data such as estimates of probabilities, physical phenomenon such as temperature, simple **ranks or ratings** (1-5), and error bounds on any such estimates of probability, physical phenomenon, or ranks or ratings, etc. (0.75 ± 0.25).

  For example, an expert might determine how his firm's quantitative data, such as on projected cost, and qualitative data, such as on market potential, should be modeled in order to make decisions about next year's product line. And this data might include other expert's judgments, such as on market potential.

Similarly, expert judgment might be employed to interpret existing data, even when that data is other expert judgment. For instance, decision makers often interpret expert judgment data. They receive multiple and differing experts' judgments and have to decide whether or how to use them.

Another situation in which experts interpret data is diagnosis. Medical specialists frequently must interpret differing test results in arriving at a decision.

- **To learn an expert's problem-solving process or a group's decision-making processes.** Often the experts do not know how they solve a problem or reach a decision because their thoughts have become automatic and, thus, difficult to recall. Yet, this information on their problem solving is needed to improve current practices, to train others, or to create systems that provide expert advice.

  One project involving learning the expert's problem solving focused on discovering the expert's procedures in reaching a decision. In this project, the experts were police officers who specialized in resolving hostage-taking events. The experts assessed information on the situation and reached decisions on how to proceed, such as whether to negotiate or resort to an assault (Vedder and Mason 1987).

  In another project, the goal was to provide guidance on how the organization would make future decisions on the export of munitions. The experts, members of different Army offices, divided the problem into parts and specified the type of input that they wanted each office to provide in making the larger decision (Meyer and Johnson 1985).

- **To determine what is currently known, what is not known, and what is worth learning in a field of knowledge** (Ortiz et al. 1988). In the reactor risk study, NUREG-1150, the experts exchanged the most up-to-date information in preparation for giving their answers to particular questions. As a result, they identified gaps in their field's state of knowledge and determined in which areas they would most like to see research. This type of information offers several benefits: it can serve as a complement to the current state of knowledge or as motivation for further study.

Often expert judgment is used to address more than one of the above-mentioned needs. Such was the case in the new reactor risk project, NUREG-1150 (Wheeler, Hora, Cramond, and Unwin 1989), where the expert judgment met all of the above-mentioned purposes. In addition, the gathering of expert judgment often provides side benefits: one of the most common benefits being the facilitation of communication. The experts are able to see how their judgments differ and relate to each other's views in an environment of openness and objectivity. We have noticed that the synergism of interexpert discussion stimulates results that would not have been achieved otherwise.

General Attributes of Expert Judgment

In general, expert judgment can be viewed as a representation, a snapshot, of the expert's knowledge at the time of response to the technical question (Keeney and von

Winterfeldt 1989). As Ascher (1978: 203) notes, "multiple-expert-opinion forecasts, which require very little time or money, do very well in terms of accuracy because they reflect the most-up-to-date consensus on core assumptions." The expert's judgment legitimately can and should change as the expert receives new information. In addition, because the judgment reflects the expert's knowledge and learning, the experts can validly differ in their judgments.

Frequently, expert's answers are given in quantitative form, such as probabilities, ratings, or odds. For instance, an expert's answer to the question could be respectively 0.10, 1 on a scale of 10, or 1 in 10 chances. Quantitative response modes are often requested because the numeric data are more easily analyzed than qualitative data.

Much of expert judgment is the product of high-level thought processing, also called knowledge-based cognition. By **cognition** is meant the mental activity that occurs when a person is processing information, such as for solving a problem. **Knowledge-based cognition** is the high-level interpretive or analytic thinking that we do when confronted with new and uncertain decision situations (Dougherty, Fragola, and Collins 1986: 4-2) Thus, knowledge-based cognition is often invoked by the situations for which expert judgment is sought.

The quality of expert judgment varies according to how the data are gathered, and the data can be obtained in a variety of ways ranging from unconscious to deliberate. Expert judgment can be gathered unconsciously, as often occurs in technical projects. Analysts typically make decisions in defining problems, establishing boundary conditions, and screening data without being aware that expert judgment (their own) has been used.

Expert judgment is also gathered deliberately, although even this type of gathering varies along a continuum of informal to formal. On the informal end of the continuum, experts are asked to provide judgments *off the top of their heads*. The informal means of gathering expert judgment has been a source of current controversies involving expert judgment. The most recent controversy involves psychologists and psychiatrists serving as expert witnesses in legal proceedings. Recent articles have proclaimed that these expert witnesses are no more accurate than lay persons, particularly in predicting an individual's propensity for future violence. These situations illustrate that "without the safeguards of the scientific method, clinicians are highly vulnerable to the problematic judgment practices and cognitive limitations common to human beings" (Faust and Ziskin 1988: 33).

Formal means of gathering expert judgment usually involve selecting experts according to particular criteria, designing elicitation methods, and specifying the mode in which the expert is to respond. The formal approach to elicitation has two advantages over its unconsciously or informally gathered counterparts. First, with the formal approach more time and care is taken in eliciting the judgments. Because the quality of expert judgment is often evaluated in terms of the methods used to gather the judgments, the greater time and effort associated with the formal approach is an advantage. Second, the formal approach is more likely to be documented than those that used unconsciously or less formally. That is, records may be kept of which elicitation methods were used and of how the experts arrived at their final judgments. Such a record allows the formal method and its results to be scrutinized. Thus the formal approach is more likely to advance through the process of reviews (Ortiz et al. 1988).

Expert Judgment Covered in This Book

Gathering expert judgment in a formal and structured manner is covered in this book. An additional focal point is gathering expert judgment according to those methods suggested by the research into human cognition and communication. Even the formal means of gathering expert judgment have not generally made use of the rapidly emerging body of literature on how best to obtain expert's judgments (i.e., how to avoid biases). After all, expert judgment is frequently needed, and its gatherers are not always familiar with the biases to which we, as humans, are prone. The result is that experts are asked to provide estimates without concern to bias or the benefit of methods shown to improve accuracy.

The research on how judgments should be elicited comes from three fields--psychology, decision analysis, and more recently, knowledge acquisition, a branch of artificial intelligence. Examples of works in the field are Hogarth (1980) for psychology, Spetzler and Stael von Holstein (1975) for decision analysis, and Gaines and Boose (1988) for knowledge acquisition.

Using the methods suggested by the research usually enhances the quality of the expressed judgments. Following are some examples of just a few of the ways that research can be applied to improving the gathering of expert judgment.

The use of one such method--breaking a problem into its component parts--has been shown to yield more accurate answers (Hayes-Roth 1980 and Armstrong, Denniston, and Gordon 1975).

Other research has shown that people have difficulty correctly translating their judgments into quantities, such as probabilities. In the state-of-the-art elicitations of expert judgment, either probabilities are not required or the experts are given lessons in their use (U.S. NRC 1989).

Similarly, people are known to be unable to consider more than approximately seven things at once (Miller 1956). To deal with this limitation, the experts may be asked to use scales that allow them to compare two, rather than seven or more, things at a time.

In addition to focusing on eliciting expert judgment as suggested by the relevant research, in this book we define expert judgment to include more than simply the expert's estimates or solutions. We have broadened the traditional meaning of expert judgment because we believe that all the data associated with the expert's answer are important to understanding his answer. Indeed, many gatherers of expert judgment have begun to document the expert's thoughts as well as their answers, perhaps because of the influence of artificial intelligence and expert systems. From this point on, expert judgment will be defined to include the following:

1. **Any of the expert's work in selecting or defining the scope of the problem.** For example, in the reactor risk study mentioned earlier (U.S. NRC 1989), the experts reviewed proposed problem areas--safety issues in reactor risk studies--and added or deleted issues. The selected issues that they chose and their criteria of selection are considered expert judgment, according to this book's definition.
2. **Any of the expert's work in refining the problem.** In many uses of expert judgment, the problem is depicted in some fashion, such as in scenarios, tree structures, grids of the pertinent factors, and/or pages of explanatory text.

In this book, all of these attempts to break the problem into parts and describe them are considered to be expert judgment. For example, in the reactor risk study, the experts took the questions to a more detailed level by breaking the issues into parts--scenarios by which the particular failure could occur. The problem--the check valve fails causing a loss-of-coolant accident-- was broken into three scenarios that could cause an accident occurrence. One failure scenario was for both valves to fail independently; a second was for one check valve to fail to reclose, and a third was for the valve to randomly rupture (Ortiz et al. 1988). In this same study, the key variables were operationally defined by the experts. For instance, the experts set *independent rupture* to mean a catastrophic leak, which in turn had been defined to mean a particular flow rate per unit time.

3. **Any of the expert's mental processes in arriving at a solution.** This aspect of expert judgment often involves the expert's sources of data, definitions, assumptions, and mental procedures for processing the information. For example, in the same reactor risk study, some experts used information from event trees as their primary source of data in solving the problem, others used information from experiments, and still others relied on output from computer models. In addition, the expert's definition of terms are considered part of his problem-solving processes. Often experts unconsciously assign their own meanings to terms. For this reason, in the reactor risk study, variables like *catastrophic leak* were given definitions by the panel of experts. Experts also utilize cognitive techniques for helping them process the problem information. One such technique, or shortcut, is to adjust up or down from a base line. For example, an expert could evaluate the risk posed by a rare accident by setting the frequency of a related but more common risk as his base line.

The above-mentioned three categories of data relating to the expert's solving of the problem will be referred to as **expert data.** In a risk analysis application, expert data is likely to include the expert's assumptions, his definitions, and his decomposition of the problem into its parts. In a knowledge acquisition (artificial intelligence) project, the expert data could be the expert's rules or procedures for reaching a solution. Expert data is a subset of an even more general class of information that is elicited from the experts. This more general class is called **ancillary data/information** and it includes data gathered on the educational background, work history, current job environment, or personality of the expert. Additionally, the term **estimates** will refer to answers such as probabilities or ratings that are given in quantified form. **Solutions** will refer to answers given in qualitiative form, such as descriptive text or diagrams. **Answers** will be used as a general term for both estimates and solutions. **Expert judgment** will be used as a cover term to refer to a combination of the expert answers and ancillary information.

How Expert Judgment is Elicited

How expert judgment has been elicited has differed widely even within a particular field, such as risk analysis or knowledge acquisition. The following factors affect how the expert judgment can be best gathered in particular situations:

- The type of information that is needed from the experts (answers only or ancillary expert data)
- The form (**response mode**) in which the expert's answers are needed for input into a model
- The number of experts available
- The interaction desired among the experts
- Difficulty of setting up the problems
- The amount of time and study needed by the experts to provide judgments
- The time and resources available to the study
- The methodological preferences of the interviewer or knowledge engineer, analyst, funder, and experts

Elicitation is the process of gathering the expert judgment through specially designed methods of verbal or written communication.

For example, in one situation, a large group of experts is convened for a week to interactively construct a problem representation and to provide the estimates. They produce a representation of all the factors involved in deciding whether to export an army-developed technology (Meyer and Johnson 1985). During the week, the experts separately weight the importance of these factors. They test their representation by applying it to an example and examining the outcome from the mathematical processing of their estimates through the decision framework.

In another situation, the experts are interviewed in depth, separately, to obtain their judgment on the performance of their computer code in modeling reactor phenomena (Meyer and Booker 1987b). They are asked to work the selected problem in the presence of an interviewer and to explain in detail their thinking as they work through it. They are requested to rate the computer's performance on a linear scale.

In a third situation, the experts are interviewed separately as well as convened for discussions. The experts are first sent seismic-tectonic information and asked to detail zones for the United States (Bernreuter, Savy, Mensing, Chen, and Davis 1985). They are also asked to provide estimates on the frequency and magnitude of earthquakes by zone. The experts are then convened and presented with a combined zone map and estimates of earthquake phenomena. They receive the judgments of the other experts after these have been made anonymous. They then have the opportunity to discuss this information together and to privately and confidentially make adjustments to their estimates.

The above examples illustrate some of the diversity in elicitation processes. Elicitation processes can also differ in terms of the following:

1. The degree to which the experts interact
2. The amount of structure imposed by a group moderator or interviewer on the elicitation process
3. The number of meetings

4. The time allotted for structuring the problem, eliciting the expert judgment
5. Who performs these tasks, the experts and/or the analyst
6. The response mode in which the expert estimates are elicited
7. Whether the expert's reasoning is requested or not
8. The level of detail in the expert judgment elicited
9. Whether the expert judgment undergoes some translation in a model and is returned to the experts for the next step
10. Whether all or some of the elicitation is conducted in person, by mail, or by telephone

Despite this diversity in the elicitation processes, there are only three basic elicitation situations and a general sequence of steps. Expert judgment can be elicited through the following:

- **Individual interviews** in which one expert is interviewed in a private, usually face-to-face situation, by an interviewer or knowledge engineer. This situation is suited to obtaining in-depth data from the expert, such as on his means of solving the problem, without having him distracted or influenced by other experts.

  The individual interview is also called the *staticized* or *nominal-group* situation when the experts' estimates which have been obtained in private are mathematically combined to form one *group* answer.
- **Interactive groups** in which the experts are in a face-to-face situation with both one another and a session moderator when they give their data. The participants' interactions with one another can be structured to any degree: (1) a totally unstructured group resembles a typical meeting; and (2) a highly structured group is carefully choreographed as to when the experts present their views and when there is open discussion to prevent some of the negative effects of interaction.
- **Delphi** in which the experts, in isolation from one another, give their judgments to a moderator. The moderator makes the judgments anonymous, redistributes them to the experts, and allows them to revise their previous judgments. These iterations can be continued until consensus, if it is desired, is achieved. This elicitation situation was developed by RAND as a means for countering some of the biasing effects of interaction.

The general sequence of steps in the elicitation process is as follows:

1. Selection of the question areas and particular questions
2. Refining of the questions
3. Selection and motivation of the experts
4. Selection of the components (building blocks) of elicitation
5. Designing and tailoring of the components of elicitation to fit the application
6. Practicing the elicitation and training the in-house personnel
7. Elicitating and documentating expert judgments (answers, and/or ancillary information)

Philosophy Guiding the Elicitation

The philosophy put forth in this book is that the elicitation be designed to fit the experts and the way that humans think rather than forcing the experts to adapt to the methods. We propose that the research on human limitations and tendencies toward bias be taken into account in selecting the methods. For example, if the interviewer does not consider people's limitation in comparing more than seven things at once in selecting elicitation methods, the resulting data is less credible. If in the former case the expert estimates are being used as inputs into a model or decision process, there is the danger of *garbage in, garbage out.*

Also eliciting as much of the information on the expert's problem-solving processes as possible is advocated in this book. We believe that this data is necessary to the understanding of the expert's answers. Expert's estimates have been found to correlate to the way that they solve the problem (Booker and Meyer 1988a, Meyer and Booker 1987b). Frequently, the definitions or assumptions that the expert used explain why that particular answer and not some other answer was reached. In addition, this type of data is valuable later if multiple expert's estimates are to be mathematically combined to form a single estimate. The expert data can guide the aggregation so that experts who construed the problem very differently will not have their answers combined inappropriately . In general, recording information on the expert's thinking allows the judgments to be more easily updated as new information becomes available.

Another aspect of the elicitation philosophy is to control for the factors that can enter into the elicitation process and influence the expert's problem solving. For example, the phrasing of the problem, the interviewer's responses, and other participant's responses can affect the answer an expert reaches. The elicitation methods, mentioned in chapters 7 and 8, are designed to control these influences. For those influences that can not be easily controlled, such as the expert's tendency to anchor to his first impression, we recommend gathering as much data as possible to analyze their effects.

Philosophy Guiding the Analysis

The analysis philosophy of this book complements that of the elicitation philosophy mentioned above. Just as the elicitation approach allows the experts' capabilities to shape the data-gathering methods, the analysis philosophy allows the data to dictate which analytic methods are used. Thus, the analyses are data driven. This analysis approach is used in the belief that it will produce the highest quality results.

As a part of the analysis philosophy, the analyses avoid assuming particular properties of expert judgment. For example, the analyses do not *a priori* assume that the expert judgment data is **normally distributed**, that the answers of multiple experts are independent, or that the experts are perfectly calibrated (unbiased). Instead, the analysis sections offer methods that either do not require these assumptions or that can test for the existence of such properties. For example, **nonparametric** statistical procedures and data-based **simulation** techniques are used because they do not depend on an assumed distribution of the data.

The analyses also avoid assuming that the expert's data are independent or dependent (i.e., conditioned on some common factor, such as the expert's education). Analysts have been driven into assuming independence in the past because they need to combine multiple estimates into a single representative one, and most aggregation schemes have required independence. Those aggregation schemes that have not required independent data are more complicated and require information on the structure of **dependence**, information which the researcher rarely has. As part of the analysis philosophy in this book, simple methods from recent research (Booker and Meyer 1988b; Meyer and Booker 1987b) are provided for testing for independence and for handling dependence, if it is found.

In addition, the analyses allow the reader to check for biases in the elicitation process or in the expert's judgment. Many users of expert judgment have been forced to assume that the data was unbiased because they had no way of analyzing biased data. To meet this need, the analysis section provides methods for investigating potential sources of bias in the data elicited earlier .

A variety of methods are used to address the multivariate nature of expert judgment data. The data is multivariate because it often includes answers to multiple problems, data on the experts themselves, and mixtures of qualitative and quantitative data. **Multivariate analysis** techniques allow the simultaneous consideration of two or more variables of interest (Tietjen 1986). Thus, they can be used to investigate some of the more important properties of the data, such as the dependence of experts. Several multivariate techniques, such as **cluster analysis, discriminant analysis,** and **general linear models (GLMs)**, are used because no single one is universally applicable to the structure of expert judgment data. In addition, methods are given for transforming qualitative data into quantitative forms.

How to Use this Book

This book is divided into three parts--*Part I: Introduction to Expert Judgment; Part II: Elicitation Procedures;* and *Part III: Analysis Procedures*. Part I consists of three chapters. In chapter 1, *Introduction*, we have presented a description of what expert judgment is and our philosophy for its elicitation and analysis. Chapter 2, *Common Questions and Pitfalls Concerning Expert Judgment,* includes many questions asked about expert judgment and the hidden traps encountered in eliciting or analyzing it. In the third and last chapter, *Part I: Background on Human Problem Solving and Bias,* information on how we as humans think and some of the consequences of those processes, such as common biases, is provided. The background provided here is necessary to setting up, selecting, and tailoring the elicitation method so as to obtain the best quality data possible. Part II, chapters 4 through 10, containing sections on eliciting expert judgment, and Part III, chapters 11 through 18 on analyzing results, are outlined by chapter in a flow chart below. This chart gives guidance for the best use of these parts of the book, pointing out the most efficient sequencing to use in differing situations. Appendices A through D document the computer programs we have used. And finally, a glossary is provided as a quick reference for the terms that have been highlighted in bold in the text. We expect the

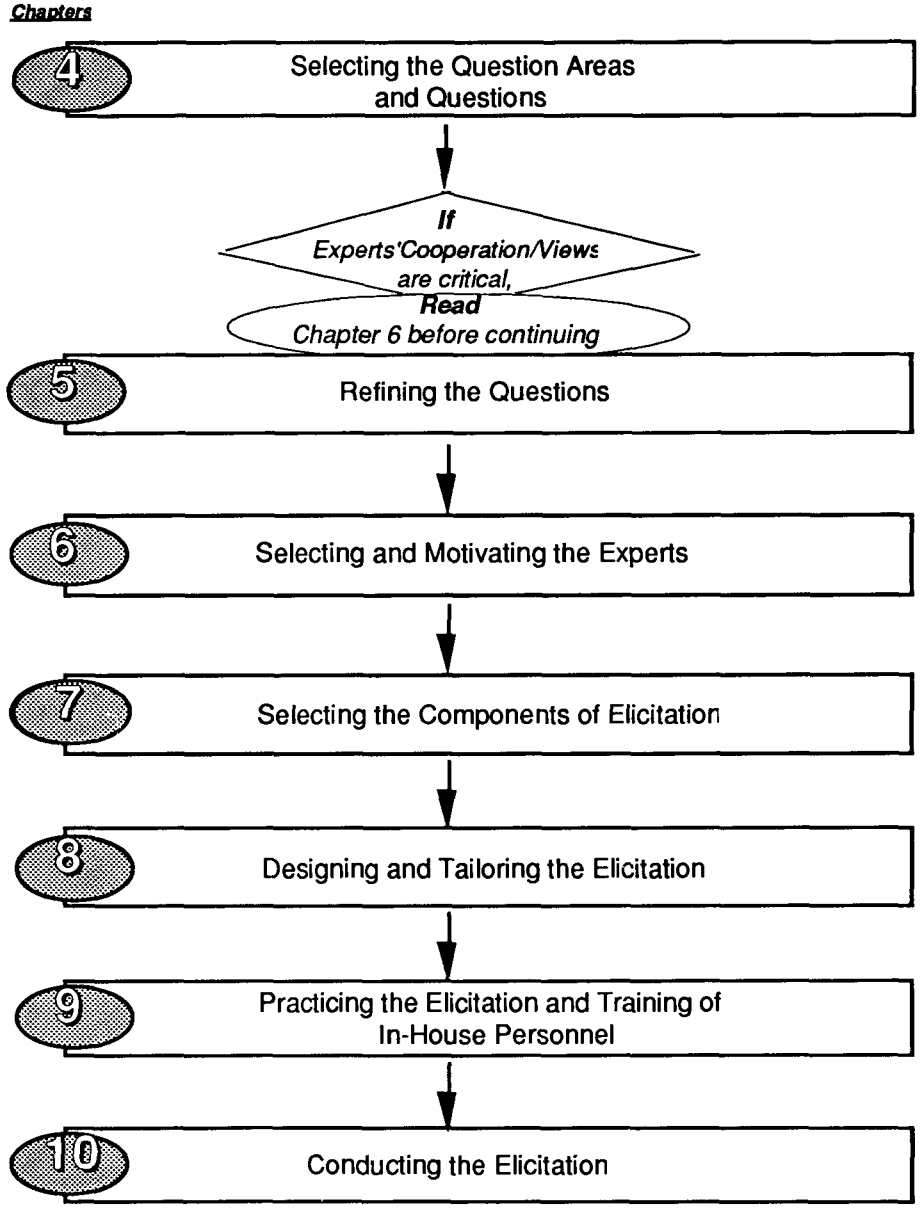
glossary to be a major aid to those readers who are unfamiliar with the book's terminology and recommend that they refer to the glossary heavily at first.

There are several ways for the reader to use this book, depending on needs and situation. The book proceeds sequentially from the selection of the problem (Part II) to the elicitation procedures and the analysis of results (Part III). The reader can read about each phase as he or she is ready to execute it. Reading portions in retrospect will give an understanding of the strengths and weaknesses in how the phases were conducted. Most of the chapters on the elicitation contain sections on common difficulties and means for resolving them.

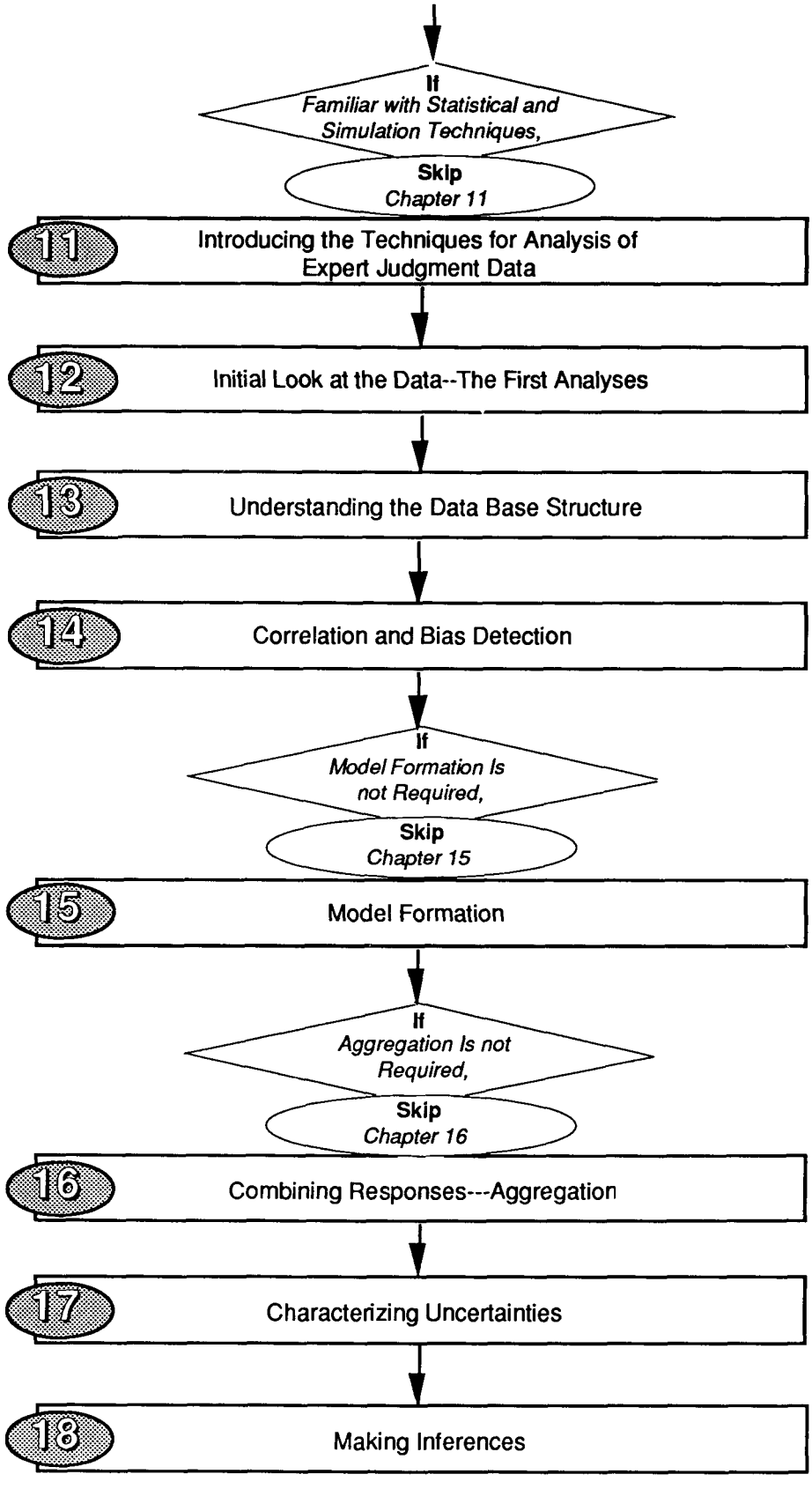
FLOW CHART FOR USE OF BOOK

(Parts II and III)

Part II--ELICITATION



Part III--ANALYSIS



2

Common Questions and Pitfalls Concerning Expert Judgment

In this chapter many of the common questions, often arising from misconceptions concerning expert judgment, are addressed. In addition, information on those aspects of elicitation and analysis that have typically caused problems are discussed under the designation *pitfalls*. We hope that the information on pitfalls will alert the user and prevent his falling into the difficulties described.

For the reader's convenience, the information on both the common questions asked and the common pitfalls encountered are summarized in the lists below; later questions/misconceptions and pitfalls are addressed at length.

Common Questions:

1. **What does it mean when the experts disagree?** When the experts disagree, as they do with irritating frequency, it can mean that they interpreted the question differently (i.e., took it to mean different things) or that they solved it in using different methods (e.g., used different algorithms or data sources). If the data indicates that the expert interpreted the question differently, one option is to have the expert, who essentially answered a different question, readdress the question as it has been defined by the other experts. His previous response may be discarded as being invalid.
2. **Is expert judgment scientific?** We propose that the type of expert judgment advocated in this book *is* scientific. The methods for gathering and analyzing it are based on research that has followed the tenets of science--observation, hypothesis formation, and experimentation.
3. **Are experts Bayesian?** People, in general, do not naturally follow the philosophy developed from the application of Bayes Theorem. This philosophy assumes that existing information is updated to account for new information as it becomes available. In laboratory and real-life settings, experts may fail to sufficiently adjust their estimates in view of new information, to grasp the effects of sample size and variability, or to follow the axioms of probability theory.

4. **Do experts give better data?** Whether experts give more accurate or better quality data depends on the type of data that they are asked to give. In providing predictions, humans are notoriously poor, and the experts have not consistently been shown to be better than nonexperts. Experts are considered better than nonexperts in providing data on the state of the knowledge in their field, on how to solve problems, and on the certainty of their answers.
5. **Can experts be calibrated?** Experts cannot yet be fully calibrated as some technical instruments and processes are. Calibration means the comparison of unknown instruments or processes to known or correct ones to adjust the unknown ones until they match the known. To improve expert's calibration, the feedback from the known procedures must be immediate, frequent, and specific to the task (Lichtenstein, Fischhoff, and Phillips 1982), such as in weather forecasting. This type of information is unavailable for the majority of problems requiring expert judgment, at least in the risk and reliability fields.

Common Pitfalls:

1. **Interviewers, knowledge engineers, and analysts can introduce bias.** These persons can unintentionally introduce bias--that is cause an altering of the expert's thinking or answers--in two ways. First, interviewers or knowledge engineers, are likely to misinterpret the expert's data by perceiving it to be the type of data that their training ideally equipped them to handle (training bias). Second, analysts are likely to misrepresent the expert data by forcing it to fit the models or analytic methods with which they are most comfortable (tool bias).
2. **Experts are limited in the number of things that they can mentally juggle.** The limit to the amount of information that humans can process in their short-term memory is seven plus or minus two (Miller 1956). This information-processing limit is something to consider in designing the elicitation situation.
3. **The level of detail in the data affects the analyses.** The **granularity** is the level of detail in a chunk of information (Waterman 1986). An example of coarse granularity could be the basic functions of a nuclear power plant; an example of finer granularity could be the subfunctions of one such function. We have found that the granularity used in gathering the data influences the results of the analyses.
4. **The conditioning effect poses difficulties in gathering and analyzing expert data.** The data that the expert gives can be conditioned on a variety of factors, ranging from the wording of the problem to the expert's internal mood. The conditioning effect poses problems because many factors can intrude without the data gatherer's awareness, overlap with other variables, and complicate the analyses.

Questions

What Does It Mean When the Experts Disagree?

We are all acquainted with instances when the experts have disagreed. This disagreement has led some persons to question the credibility of expert judgment. The reaction, although natural, is based on a misconception about expert judgment. The misconception is that expert judgment is reproducible--that experts, given the same data, will reach the same conclusion, otherwise their judgment is questionable. This concept comes from experiments in the *hard* sciences where reproducibility of results is used as a yardstick. However, this concept is inappropriate and misleading when applied to expert judgment for two reasons.

First, experts do not possess the same data, even if they are given the same briefings and information as background to the problem. The expert's body of knowledge, the primary reason for consulting him, is something which has been created over time through the expert's professional experience (education, on-the-job-training, and exposure to data bases or dramatic events in the field). Each expert's knowledge is, therefore, different. In addition, what the experts do with their information, how they consider the separate pieces, is likely to differ. For example, if two experts possessed the exact same information, they would probably differ in their use of the data (e.g., one might consider a datum highly relevant to the case in point and use it whereas the other might dismiss it from consideration). Because experts have different bodies of knowledge and approaches to solving problems, their answers are likely to differ. The relationship between the expert's problem-solving approach and the answer reached has been verified by research (Booker and Meyer 1988a, 1988b; Meyer and Booker 1987b).

Second, expert judgment is frequently sought in situations where there are not clear standards or well-developed theories as there are in many areas of the physical sciences. For example, expert judgment is often sought for prediction, such as the likelihood of an individual performing a threatened act, the chances of occurrence of a seismic disturbance of a particular magnitude, or the probability of occurrence of a rare sequence of reactor events. In the engineering sciences, where many standards are clearly defined, the experts have guides to follow in reaching a decision. Such standards tend to reduce expert variability in problem-solving approaches. However, in the fields where expert judgment is usually elicited, such standards are not in place and therefore do not lead to greater uniformity among the experts' solutions.

Thus, when the experts disagree, a valid interpretation would be that they have interpreted and solved the problem in differing manners. If records have been kept as to their problem-solving processes, one can pinpoint where they differed (e.g., in their assumptions, definitions, or algorithms) and justify the later inclusion or exclusion of a particular expert's answers. It may be that some of the expert's approaches are better approximations of the reality (of the problem being posed) and therefore more likely to be accurate, as proposed by the Brunswik lens model (Hogarth 1980:8) However, it is difficult to select which approach is better when the *right* answer or approach is not known. For this reason, differences among the experts can be interpreted in a positive light as evidence that the different perspectives on the problem are being represented. In fact, studies have shown that mathematically combining the experts' differing answers

provides a better chance of covering the right answer than does the use of single expert's answers.

Is Expert Judgment Valid Data?

Expert judgment has been called *subjective data*, *subjective judgments*, *expert estimates*, and *expert* (judgment) *data*. If the data are in the form of probabilities, then terms such as *subjective probabilities* and *probability estimates* have been used. Other times expert judgment is called *qualitative data* even though the data itself may be in the form of numbers rather than words. (The correct use of the term *qualitative data* is for non-numeric data, regardless of its source.)

The term *data* has different meanings for different people. To some, data refers exclusively to measured or observed numerical values (cardinal or real numbers). To others, data is used in a less restrictive context and refers to information. Historically, however, data has been cardinal (real-numbered values), ordinal (numeric or verbal ranks), categorical (numeric or verbal classes or categories), or descriptive (words, phrases, sentences). In this book the term *data* is used to refer to information either in a qualitative or quantitative form.

Some have questioned whether expert judgment is data, or *good* data, given its source. These individuals consider expert judgment to be lower in quality than *hard* data that has been measured or obtained from observation or from instruments.

In this book, the hypothesis is that expert judgment data is comparable to any other data. All data are an imperfect representation of the object they are supposed to represent. Data from instruments are not a perfect representation for many reasons: random noise, equipment malfunction, operator interference, data selection, or data interpretation. Expert judgment data is no less representative of the underlying truth than data from instruments or any other source of data.

Expert judgment data, like any other data, must be carefully gathered, analyzed, and interpreted. The guidelines given in this book are designed to facilitate the careful handling of expert judgment data.

Is Expert Judgment Scientific?

There are differing views as to whether the gathering of expert judgment can be considered scientific. To arrive at an answer, one can consider the definition of science: "a process or procedure for making inquiries of our world" (Hirely, 1989:25), with the basic tenets of observation, hypothesis formation, and experimentation. Expert judgment studies follow all these tenets. The large body of research on human judgment and problem solving illustrates this point. For example, these topics have been the focus of tightly controlled experiments: judgmental processes (Hogarth 1980; Kahneman and Tversky 1982), effects of problem decomposition (Hayes-Roth 1980; Armstrong, Denniston, and Gordon 1975), memory functions (Ericsson and Simon 1980), effects of group dynamics (Zimbardo 1983), effects of question phrasing (Payne 1951), probability estimation (Spetzler and von Holstein 1975; Hogarth1980), and sources of interexpert correlation (Booker and Meyer 1988a).

In conducting the third tenet of science, experimentation, expert judgment studies have faced challenges beyond those characteristic of the physical sciences. These challenges stem from the use of human subjects. Human subjects have the capability of manipulating the experiment while the typical subject of physical science studies do not. For example, a chemist does not have to cope with an element suddenly changing its behavior, such as its mass, because it favors a particular outcome in the experiment. In addition, with human subjects, the data is generally conditioned on multitudes of unknown and uncontrollable factors. An expert's judgment or descriptions of it can be conditioned on attributes of the expert (assumptions and algorithms used in solving the problem, training, past experiences, and so on), the interviewer's presence, the phrasing of the question, and the responses (verbal and nonverbal, real and imagined) of the interviewer or others. Factors such as the expert's mood prior to elicitation and aspects of his background are among those that generally cannot be controlled.

In addition, the expert data gathered is often of a predictive nature which cannot be objectively verified, at least not in real time. These judgments cannot be evaluated by performing additional measurements of some physical property or by referring to some authoritative text.

These aspects of expert judgment studies pose problems in experimental design. Typically, experimental design involves the planning of a study in terms of what will be observed for measurement and how it will be observed. In expert judgment studies and applications, the investigator cannot totally control the entry of factors for observation because the subject brings a variety of unknown ones to the elicitation. Then too, the small sample sizes of expert judgment studies do not allow the data to rise above their effect. In addition, the type of data gathered makes testing difficult, especially in a scientific tradition that holds that "the world is objectively knowable and that deductions about it can be tested" (Denning 1988).

In sum, we would argue that the field of expert judgment follows the tenets of science and produces scientific studies. Furthermore, it is through the careful gathering, use, and examination of expert judgment that this field will make further progress as a science.

Are Experts Bayesian?

In the mathematical community, there are many analysts and theorists who advocate the Bayesian philosophy of analysis. In the decision analysis and reliability assessment communities, this philosophy has come to mean the evaluation of gathered data as they are conditioned on other events or circumstances (variables). Given that the data are conditioned on other variables, the Bayesian philosophy implies that as these conditions change, the data changes. In other words, data are updated with changing conditions. However, there is a major problem in applying Bayesian philosophy to expert judgment--experts are not naturally Bayesian (Kahneman and Tversky 1982). Human cognitive processes do not naturally follow Bayesian philosophy.

Humans are not Bayesian for a variety of reasons demonstrated in both laboratory settings and actual applications. The above studies of Kahneman and Tversky have shown that that experts fail to change or adjust their estimates in view of new information. Mathematically, the failure to update estimates means that $P(A|B)=P(A|C)$, i.e., the

probability of A is not altered when the conditions governing A are changed from B to C. This equation would only be true if P(A) was independent of any conditioning, i.e., P(A|C)=P(A) and P(A|B)=P(A). However, in estimating probabilities it is unlikely that any event would be so totally independent of conditions.

Other characteristics of human cognition prevent humans, including experts, from being Bayesian. Some of these characteristics are the inability to grasp the effects of sample size, the frequencies of truly rare events, the meaning of randomness, and the effects of variability (Hogarth 1975).

These same failings also contribute to human difficulties in estimating probabilities in general (Kahneman and Tversky 1982). The human brain does not follow the axioms (rules) of probability, such as all probabilities lie in the [0,1] interval and the sum of mutually exclusive event probabilities must be 1. The *probabilities* elicited from a human are not representative of a true, mathematical, probability measure. One of the purposes in writing this book is to provide guidance on which inferences and interpretations can be applied to expert judgment data.

Do Experts Give Better Data?

The literature differs on whether experts give better quality data than nonexperts. Some reasons for this disparity in the literature might be the following:

- Real populations of experts are rarely used for these comparative studies
- Expert-level questions are not usually asked because of the difficulties in evaluating the responses when there are no known answers
- A number of factors intervene in the comparison (e.g., the type of data the expert is being asked to provide and in what form, or to whom the expert is being compared, novices or persons with training in the field)

In other words, the answer to the question "Do experts give better data?" is that it may depend on the type of data that they are asked to give. The expert can be asked: (1) to make predictions, (2) to provide information on the field, (3) to show how to solve the problem, or (4) to assess the accuracy of his responses.

It has been claimed that "there are no experts in forecasting change" (Armstrong 1981:89). Certainly studies in the social sciences and medicine have not consistently shown experts to be better at predictions than lay persons (Armstrong 1981). Most of these studies have aimed at predicting human behaviors, such as patients' lengths of stay in mental institutions, or the winning scores of sports teams, or economic trends. For instance, one highly publicized study of expert witness psychiatrists and psychologists found expert prediction of future violence to be highly inaccurate and no better than that of lay persons (Faust and Ziskin 1988). Faust and Ziskin noted (1988:33) that experts are dependent on the state of their science, which in this case "lacks a formalized, general theory of human behavior that permits accurate prediction." They proposed that both the lay persons and the clinicians had resorted to using common cultural stereotypes and assumptions about potentially violent people in making their predictions. Thus, their predictions were similar and inaccurate.

It should also be noted that prediction questions have been worded (without technical jargon) so that they can be understood by novices and answered by guesses, as opposed to problem-solving questions that a novice would find difficult to understand. Thus, the wording of the questions may minimize the difference between the performance of the expert and the novice.

Another reason why the experts do not significantly outperform nonexperts in prediction may be because of the response modes that are used. Typically, subjects are asked to give their predictions in the form of probabilities. While the **experts** may be knowledgeable in the field of study, this is no guarantee that they will be expert in assigning probabilities. (Being knowledgeable in the field of study is sometimes referred to as **substantive expertise** and being knowledgeable in the use of the response mode as **normative expertise**.) In general, people do not estimate probabilities in accordance with statistical principles, as mentioned in the previous pitfall.

It is generally thought that experts are both better on knowing the state-of-the-art and at providing data on how the problem can be solved. In more detail, this data could encompass formulating the problem, interpreting it, determining what additional information is needed to solve it, knowing whether and where this data is available, knowing how to solve the problem, providing the solution, and estimating how much confidence can be placed in the solution. Typically, all of this knowledge is used in an artificial intelligence application and has been documented in recent probabilistic risk assessments (U.S. NRC 1989). In these uses of expert judgment, the expert's judgment is not as frequently encoded in a response mode and, therefore, may more directly reflect their substantive expertise.

Armstrong (1981), who has questioned the value of experts in prediction, considers problem solving to be a proper area for the use of experts. This view seems to be held even more strongly in the expert systems community, where the knowledge engineer is likely to be advised to "be sure to pick an expert highly skilled in the target domain" (Waterman 1986:192). Use of a true expert is given as the means for "extracting high quality data" and for avoiding great difficulties.

There is some evidence that experts think differently than nonexperts when solving problems in their area of expertise. Best documented is the expert's ability to recall greater amounts of relevant visual information. For example, skilled electronics technicians were able to recall more of briefly shown circuit diagrams than the novices (Egan and Schwartz 1979). It is proposed that experts are able to code, for memory, chunks or groups of information that are conceptually related. It is also commonly proposed that experts are more abstract, pattern types of thinkers than nonexperts (Denning 1986; Doughterty et al. 1986). For example, the Dreyfuses (1986) characterize a novice as knowing basic facts and context-independent rules; an expert as having little conscious awareness of these rules but an ability to visualize and manipulate whole sets of objects and situations.

Another type of data that subjects can give is an assessment of their own accuracy or confidence in their answer. The process of trying to assess or improve the accuracy of the expert judgment is sometimes called calibration. These assessments are frequently given as probabilities. In general, people are very poor at assessing the accuracy of their own answers and are usually overconfident; that is, most individuals would estimate the chances of their solution being correct more highly than warranted. However, Lichtenstein and Fischhoff (1977) have found that those knowledgeable in the field are less prone to this

overconfidence bias than those who are not. An individual's calibration improves with knowledge (as measured by percentage of correct answers) until the individual has over 80% of his answers correct. Then, those with over 80% of their answers correct become less well calibrated because they tend toward underconfidence. It may be that those who are very knowledgeable are more aware of the dangers of estimation and thus increasingly tend to underestimate their accuracy.

In sum, the quality of expert judgment may depend on the area of questioning, the wording of the question, the form in which the expert responds, and the person to whom the expert is being compared. However, the current view is that experts provide better data in situations requiring their insights into the problem, such as in solving a problem or assessing their own accuracy. These situations generally occur in the building of expert or knowedge-based systems.

It should also be noted that there are additional benefits, beyond the quality of the data, from using experts. Using experts motivates other experts to participate in the study and thus increases the study's credibility.

Can Experts Be Calibrated?

The concept of calibration is a basic part of the scientific method. Calibration is used here to mean the comparison of an unknown (instrument or process) with a known, defined standard or a correct procedure in order to adjust the unknown until it matches the standard. Until recently, calibration was applied only to measuring devices or processes for which standards or known quantities were available or defined. Thus, the concept of calibrating experts seemed a reasonable approach for getting better expert judgment data.

The conclusions from experimental studies indicate that experts cannot yet be fully calibrated. Studies by many such as Lichtenstein, Fischhoff and Phillips (1982) show that feedback on the outcome of events can reduce, but not eliminate, the biases which hamper calibration. In order for feedback to be effective as a calibration tool, it must be immediate, frequent, and specific to the task. Such feedback cannot be given for problems where the outcomes are unknown, as is often the case in risk and reliability assessments.

While this situation of uncalibrated experts and unknown outcomes may seem problematic, many users of expert judgment, such as decision makers, do not worry about biases arising from their experts. Instead, they tend to have faith in experts because they perceive them as being very knowledgeable (Morris 1986).

This faith in the expert's judgment is not to imply that the calibration problem is being ignored by researchers. On the contrary, many decision analysts are focusing on the problems arising from the expert decision-maker interaction in view of calibration issues. In many applications, calibration of the expert cannot be defined independently of the decision maker (French 1986) because the decision maker factors the expert's thinking into his own in reaching a final decision. Ideally, the experts should be calibrated to the problem and to the decision maker. (This dependent relationship between the experts and the decision maker is further discussed in the section on different application environments in chapter 16.)

The decision maker also affects calibration through the evaluation of his own and the expert's calibration. For example, a decision maker who sees himself as miscalibrated can induce additional biases and misconceptions by overcompensating for his calibration.

He may not be able to perceive independence when it actually exits (Harrison 1977). Thus, awareness of miscalibration and overcompensation for it, just as ignorance of it, can exacerbate calibration problems.

In sum, calibration enters into many aspects of expert judgment and its use. However, the means for measuring the degree of miscalibration or preventing it requires further research.

Pitfalls

Interviewers, Knowledge Engineers, and Analysts Can Introduce Bias

The interviewer, knowledge engineer, or analyst can unintentionally introduce bias into expert judgment. The bias referred to here is motivational bias: an altering of the expert's responses due to the influence of the interviewer, knowledge engineer, or analyst. Specifically the data gatherers and analysts can cause bias through **misinterpretation** or **misrepresentation** of the expert data.

While the data is being gathered, the interviewers and knowledge engineers can bias the expert's data by misinterpreting it. When interviewers or knowledge engineers listen to experts and record their thoughts, they are likely to be influenced by what they already know or believe, their training, and their experience. For example, when an engineer, economist, and decision analyst met initially with military experts on a manufacturing matter, each interpreted the information in terms of her own training. The engineer perceived the problem to be an engineering one, the economist a cost/benefit one, and the decision analyst a multiattribute decision-theory one. Each questioned the experts to obtain the additional information that they needed to apply their orientation in greater depth. Each interviewer believed that she had received information that confirmed the applicability of her training to treating the matter. For this reason, we also refer to this source of bias as **training bias**.

In the later stages of an expert judgment project when the data must be represented, modeled, or analyzed, there is the potential for misrepresentation. In performing analyses, we have the tendency to force the data to fit the models or methods with which we are most comfortable or familiar. For this reason, misrepresentation is also referred to as **tool bias.** It is as if we had one tool, such as a wrench, and tried to use it on all problems--tightening bolts, pounding in nails, and removing nails. This favored tool would work well in some tasks and inadequately in others, such as in pounding in nails. However, it is likely that in continued use of the tool we would be intent on its use and unaware of its shortcomings or of a better alternative. For example, an analyst may wish to use a model that requires either independence or a particular distribution. She will probably assume that the data meets these requirements (or hope for robustness) so that the model can be used. The validity of these assumptions may not be questioned by the analyst owing to some of the social and psychological mechanisms discussed below.

It should be noted that the training and the tool bias are connected. They are connected because inherent in our fields are values that predispose us toward particular

approaches and methods. For example, artificial intelligence (Henrion and Cooley 1987) and cultural anthropology have valued the expert's knowledge and viewed it as the *gold standard* to be extracted and emulated. By contrast, the fields of decision analysis, statistics, and operations research have viewed particular mathematical and statistical rules as the standard (Henrion and Cooley 1987). Expert data is valued if it exhibits these standards, such as the axioms of probability and Bayesian philosophy. The methods that these two orientations use reflect their values. Artificial intelligence and cultural anthropology favor methods that are designed to obtain and represent the expert's natural way of thinking. The approaches of decision analysis and statistics correct for what they consider to be limitations in human information processing.

Why are we, as humans, prone to these subtle but pervasive biases? Why do we selectively take in data that supports what we already know, and believe that it can be handled by the approaches, models, or methods that we prefer? First, it should be noted that all of human perception is selective and learned. Our perceptions of reality, of what *is*, are conditioned at a cultural, societal, and individual level.

At the cultural level, meaning and structure are imposed and then taken for reality by members of that culture. For example, members of this western scientific culture would take the color spectrum, such as in a rainbow, and divide it into four to six colors--violet, blue, green, yellow, orange, and red. In another culture, the people would not see the segmentation that we do. Instead, they might have been conditioned to view the spectrum as consisting of wet and dry colors. The members of both of these cultures have been conditioned to see color in a particular way and to believe that it objectively exists as they perceive it.

At the societal level, our training leads us to define and structure problems in particular ways, to use our field's methods, and to value special types of data. Yet, we forget that these are learned values and tend to proceed as if they were simply truths that were revealed through our learning experiences. For example, many of the *hard* scientists believe that the only true data are the quantitative measurements gathered by instruments during physical experiments.

At the individual level, our desire to be able to handle the problem leads us to use those tools that we know best, and then believe that they worked. There is a psychological mechanism that allows us to avoid becoming aware of when our beliefs and perceptions do not match, such as when the use of a favored method was inappropriate. The psychological theory of cognitive dissonance (Festinger 1957) predicts that when we have either two beliefs or a belief and a perception in conflict, the conflict will be resolved unconsciously. Many tests have shown that people selectively pay attention to information that confirms their beliefs and discount that which could cause conflict (Baron and Byrne 1981). Scientists are not immune (Armstrong 1981; Mahoney 1976). For example, scientists tend to notice the data that confirms their hypotheses and either miss or discount the negative evidence (e.g., the data must be noisy, the equipment probably malfunctioned, or there could have been operator interference).

How can we guard against our own tendencies to introduce bias? First, we can strive to remain aware of this tendency. Second, we can use elicitation methods that are designed to minimize the role of, and hence the opportunity for interpretation of, the interviewer or knowledge engineer. These methods (chapter 7) used for obtaining data on the expert's answer and/or problem-solving processes, place the emphasis on the expert.

Because the focus is on learning the expert's thoughts and words and using them to pursue questioning, there is less room for the views of the data gatherer to intrude. In addition, the data gatherer can adopt the goal of being like a blank slate to avoid translating the expert's data into her own concepts. (See the section in chapter 3, *Countering or Reducing Bias--More Art Than Science.*) Last, we can use analysis methods that require the making of minimal assumptions, as described in the section *Philosophy Guiding the Analysis* in chapter 1.

Experts are Limited in the Number of Things that They Can Mentally Juggle

There are limits to the amount of information that we can process in solving problems. The classic paper by Miller (1956) identifies the number of things that people can accurately discriminate. In these studies, the subjects were differentiating things on the basis of one attribute, such as the volume of the sound. For example, when subjects were played a sound at varying levels of loudness, they could accurately discern about seven levels. Experiments were also conducted on differentiating the size of drawn squares, on the saltiness of various solutions, and on musical notes. From many such experiments, Miller determined that seven is the limit of our processing capacity because the number of errors increases greatly after that point.

The number seven is not a strict limit because, under particular conditions, we exceed it. We can go beyond the limit when we consider multidimensional data, when we perform relative rather than absolute comparisons, and when we make several absolute judgments in a row. Multidimensional data is the input that we receive simultaneously from our five senses and assess and act on as functioning human beings. As in our daily judgments, the limit of seven was exceeded in experiments using multidimensional attributes. For example, in one experiment which produced combinations of six acoustical variables, subjects were able to discern, without error, about 150 different categories. While we are able to judge more things using multidimensional attributes, this capacity also has its limits. In particular, when the total capacity of our information processing is increased, our accuracy on any particular item is decreased. In other words, when making multidimensional judgments, "we can make relatively crude judgments of several things simultaneously." (Miller 1956:88)

We can also exceed the limit of seven when we perform relative comparisons. Relative comparisons allow individuals to judge things with respect to one another and are frequently done on two things at a time. For example, A could be compared to B, B to C, C to A, and so on.

When several absolute judgments are made in a row, the information must be stored in short-term memory. Memory has its own limitations, such as the number of things that can be retained for short-term consideration. Memory limits can be expanded because humans have the capability of grouping or organizing more information per thing. This principle is called *chunking*. For example, a person learning radiotelegraphic code, first hears each dit and dat as separate chunks. Later, this person can organize letters, words, and even phrases into chunks. Experts have been found to be much more proficient at chunking data than novices. For example, skilled electrical technicians, in contrast to

novices, can briefly view a circuit diagram and immediately reconstruct most of it from memory (Egan and Schwartz 1979).

The information mentioned in this section has several implications for expert judgment. At the very least, an interviewer would not want to create an elicitation situation where the experts had to mentally juggle more than seven items at a time. Miller's research suggests that rating scales with seven or less gradations are the most useful because in larger samples finer discriminations are lost. If the project demands that a high number of distinctions be made simultaneously, the experts will judge these more crudely than if they had considered them separately. In contrast, methods requiring subjects to compare two items at a time will avoid the limit of seven and produce more precise judgments. In addition, experts, as opposed to nonexperts, may be more capable of receiving and handling larger magnitudes of information because of their ability to chunk it.

The Level of Detail in the Data (Granularity) Affects the Analyses

The term *granularity* has its origins in fields such as numerical analysis and artificial intelligence. In artificial intelligence, granularity is defined as "the level of detail in a chunk of information" (Waterman 1986). An example of coarse granularity might be the basic functions of a nuclear power plant; an example of finer granularity could be the subfunctions of one such function. In numerical analysis, granularity refers to the computational grid size used for defining the level at which the computations are made. Granularity is the level of detail at which the data is gathered, processed, and interpreted. Therefore, this level establishes the framework of operation for the problem.

The granularity, or level of detail, is an inherent part of the experimental design of a study. In most applications, this level is dictated by some limiting aspect of the problem, such as the goals of the study or the complexity of the questions asked. Thus, in most problems, the selection of the level is done implicitly and not as a separate, conscious decision. For example, in the design of a simple voter poll, the goal of the problem defines the granularity. If the goal is to determine for which party an individual voted, it would not be necessary to gather information on the voter's property holdings. If, however, one of the election issues was to determine who would support an increase in property taxes, this finer level of property information might become important. The latter goal is at a more specific level, and the information required must be correspondingly more detailed. Generally, providing data to answer the question *why* requires that a finer granularity of data be gathered.

The level of detail is also dependent upon the complexity of the problem. On simpler questions, such as those whose answers can be verified (e.g., almanac questions), the subject's problem-solving processes tend to be more structured and detailed. Thus, they are easier for the interviewer to record and for the analyst to model in full detail. On complex problems, the information tends to be more plentiful and less structured or clear. The subject and the interviewer may encounter the limitations to information processing mentioned in the previous section, *Experts are Limited in the Number of Things that They Can Mentally Juggle*. The subject is driven to using heuristics to simplify the problem solving. The subject struggles to report these complex processes, usually simplifying them or leaving out parts in the translation. In attempting to follow the interviewee's account,

the interviewer is likely to further screen and abstract the information. As a consequence, even though there is a fine granularity of data associated with solving complex questions, this level of detail is not as easy to extract or document as it is on simpler problems.

The level of granularity greatly affects model formation and interpretation and the conclusions reached. For example, different models can be formed depending on the chosen level of granularity. Typically, the analyst must construct a model whose level of detail is dictated by the data content of the subject who has provided the least or the most general information.

Granularity is also an issue in the interpretation of the data. The analyst *sees* data from her own perspective, which is not necessarily the same perspective as that of the subject from which it was gathered. When the analyst screens, transforms, and constructs problem-solving models, the granularity becomes a function of the analyst's thinking. The analyst is led, often unconsciously, to force the data into the desired level for fitting a pre-conceived model or hypothesis (See also the first pitfall under *Common Pitfalls: Interviewers, Knowledge Engineers, and Analysts Can Introduce Bias.*). Thus, the analyst's preconceptions can affect the way in which the data is represented. This pitfall is especially likely to occur when the data is highly qualitative, with high uncertainties, as is often the case with expert judgment.

An example of how granularity affects conclusions can be seen in studies of interexpert correlation (Booker and Meyer 1988a, Meyer and Booker 1987b). In the first study, (Booker and Meyer 1988a), where the problem-solving of expert statisticians were being studied, the technical questions asked of the experts were of simple construction. Very specific problem-solving features could be modeled and the statisticians were compared using standard general linear models. The conclusion was that experts using similar rules of thumb and assumptions reached similar solutions. Therefore, correlation among the experts appeared to exist at the detailed level of their problem-solving models. In the second study (Meyer and Booker 1987b), the technical questions asked of nuclear engineers were more complex in structure. The specific heuristics and assumptions that they used were so varied that the design matrix was prohibitively sparse for use in standard models. Thus, the problem-solving models had to be constructed at a more general level. When these models were constructed at a more general level, which mirrored the ways that the experts processed the magnitudes of information, the answers were found again to correlate with the expert's problem-solving techniques. If conclusions for the second study had been drawn at the same level as in the first study, there would have been no evidence for any interexpert correlation. This effect occurs because finding correlation depends on having the right data-to-noise ratio, something that the level of granularity determines. (Glen Shafer, originator of the Dempster-Shafer theory of belief functions, currently at the University of Kansas, is credited with calling this relationship to our attention.) In sum, conclusions can differ depending on the granularity of the models chosen.

The Conditioning Effect Poses Difficulties in Gathering and Analyzing Expert Data

The data that the expert gives can be conditioned on a wide variety of factors, ranging from the wording of the problem, the elicitation site and reference materials that it contains, the expert's internal state at the time of questioning, the expert's method of

solving the problem, the interviewer's or other's responses to the expert's data, to the expert's skill at articulating his thoughts. The authors believe that expert data is more highly conditioned than other kinds of data and that this attribute complicates the study of expert data.

The conditioning effect poses problems for the elicitation and analysis of expert judgment. Many of the factors are ones that the researcher has little or no control over. For example, in an elicitation session, the interviewer has little control over the state of mind that the expert brings to the session, particularly if that state has been affected by some event in the expert's private life. Furthermore, the factors often overlap and cannot be separated for analysis of their effects on the data (Meyer and Booker 1987a).

The conditioning effect relates to the problem of bias in expert data. Some of the conditioning effects could be labeled as causes of bias. That is, they lead to an altering of the expert's responses, or they lead to judgments that do not obey mathematical and statistical standards. For example, the interviewer's negative response to some aspect of the expert's problem solving could alter, or bias, the expert's subsequent problem solving. Then too, the expert's use of a shortcut in problem-solving (heuristic), such as using the present as a baseline from which to estimate future patterns, could bias his answer (Hogarth 1980).

A two-step approach is recommended for handling the conditioning effect and its offshoot, bias: (1) control those factors which can be controlled, and (2) gather as much data as possible on those factors that cannot be controlled so that the data may be analyzed later for their effect. For example, factors that relate to the question or the elicitation situation (e.g.,the wording of the question, its timing, the elicitation method, response mode, and dispersion measures) are under the discretion of the project personnel and can be designed with the conditioning effect in mind. Other factors, such as the expert's internal state, personality attributes, and professional background are not as easily controlled by project personnel. However, data can be gathered on them through a series of demographic questions administered before or after the expert solves the problem. In these two ways, the effects of conditioning can be examined, if not reduced.

3

Background on Human Problem Solving and Bias

In this chapter we provide a general background on how humans solve problems. In addition, we discuss the effects of bias and propose a program for handling its occurrence.

Why Is It Necessary to Have an Understanding of Human Problem Solving?

It is a premise of this book that awareness of how people solve problems or of the causes of bias is necessary to optimally designing an elicitation method. For example, if the interviewer were not aware that an expert's reasoning in solving a problem was only briefly available in his short-term memory, she might ask the expert for this information hours after it was gone. Similarly, if the researcher were not aware of the potential for bias, she would not design measures to detect or counter it. Bias is just beginning to be addressed as a problem affecting the quality of expert judgment.

What Is Involved in Solving Problems and Responding?

Frequently, those new to interviewing are unaware of the magnitude of the cognitive tasks that they are demanding of the expert in problem solving. After all, they reason, the experts solve problems every day. However, problem solving is not a simple task, and quite often the elicitation methods place additional cognitive burdens on this natural process like requesting the expert to answer using difficult response modes, such as logarithms.

The Four Cognitive Tasks

The expert is likely to perform four cognitive tasks during elicitation (Mathiowetz 1987):

1. Understanding the question
2. Retrieving the relevant information
3. Making judgments
4. Formulating and reporting an answer

The first task can be described as involving the expert's comprehension of the wording and context of the question. The interviewer may use a different nomenclature than the expert knows or she may use terms familiar to the expert but the terms may mean something different. The expert must also determine the aims of the question and limit his analytic frame to focus within that context.

The second task consists of the expert's retrieval of relevant information for answering the question. To retrieve this information, the expert must have previously received it and stored it in memory.

It should be noted that humans do not perceive and store all the data that is available to them. Instead, we tend to selectively notice data that supports information which we already possess. This failing is part of the reason why humans are not Bayesians (for further information, see *Are Experts Bayesian?* in chapter 2). For example, studies have shown that we pay attention to data that supports our hypotheses but ignore data that conflicts (Armstrong 1981). In addition, it is not raw data that is assimilated by a person but data that is interpreted in light of what the person has learned either individually or as a member of a particular culture. Thus, the expert retrieves data that has had at least one level of interpretation above that of objective reality.

Not all information is stored to be later accessed. Then too, mistakes are made in accessing memories. Often, the association used to access the memory can impact on what is retrieved. For example, if the expert accessed the information through a time frame, a different reconstruction of the memory could result than if he accessed it by way of a key word. In addition, the expert may not be able to distinguish between similar or related events. He may combine memories of separate events, confusing their characteristics and the time when they occurred.

The third task, making judgments, involves processing the information. Typically, people use mental shortcuts, heuristics, to assist in integrating and processing the information (Tversky and Kahneman 1974; Hogarth 1980). In simplifying the processing, these heuristics often skew the answer reached. For example, one heuristic is termed **anchoring and adjustment**. This heuristic is defined by Tversky and Kahneman (1974) as occurring when an individual reaches a final answer by starting from an initial value and adjusting from it. The initial value can be supplied with the question, or it can be determined by the expert through his impressions or computations. Usually, the final answer reflects, or is skewed toward, the initial value. For example, if an expert were trying to evaluate how well a complex computer code predicted experimental results, his first impression might be that it did a *good* job. After considering the problem in more depth, he might find a number of places where the code failed to adequately predict the

experimental results. Yet, the expert would be likely to give a final rating that was closer to his first impression than to his second thoughts (Meyer and Booker 1987b).

The fourth task, formulating a response, requires that the subject report an answer. If the expert is to use a particular response mode, this task includes his translation of his internal answer into the response mode. The expert may use some algorithms in selecting the response option that best expresses his concept of the answer. Often, response modes, such as probabilities, are governed by logical rules. Thus, the expert may also consider these rules. For example, in putting his answer in the desired form he may say, "my probabilities need to be values between 0.0 and 1.0."

A Simple Mechanistic Model of Human Information Processing

These cognitive tasks can be described mechanistically using the simple model of Ericsson and Simon (1980). We chose this model because it makes minimal assumptions about these relatively unknown processes.

The expert's thought processes can be described as involving a central processor and several types of memory that possess differing capacities and capabilities. For example, the short-term memory (STM) is of limited capacity and intermediate duration. The long-term memory (LTM) is of large capacity and relatively permanent duration. Information recently acquired (by the central processor) relating to the problem is kept in the STM for further processing. Only the most recently heeded information is accessible in the limited storage of the STM. Thus, in solving a problem, information is moved back and forth from the STM to the LTM. The expert's LTM can contain information from previous experiences, or it can have by-products of his current efforts in solving the problem. For example, if the expert needs the solution to an equation as a step in solving the problem, he may pull the equation from LTM. He may solve the equation and proceed with its product, relegating the intermediate variables and equation back to LTM. The STM has pointers to information in LTM. A portion of STMs are fixated in LTM before they are lost and can sometimes be retrieved from LTM.

According to this model, the optimal time for eliciting the expert's thinking would be while these thoughts were still in STM. Later, only a portion of what was in the expert's STM will have been fixed in his LTM and only a portion of that fixed in the LTM might be successfully accessed and retrieved. Thus, instead of the expert giving a simple report of his problem-solving processes at the time, the expert will have to recall what he did as a separate problem.

Bias

Another aspect of how people solve problems is bias. **Bias** is a skewing of the expert's judgment from what it is thought that it *should* be. There are reference points on what expert judgment should be: (1) the expert's thinking or answers; and (2) data which follows particular norms or standards. These reference points form the two definitions of bias--a skewing from the expert's natural way of thinking or from objective standards. The

two views on what constitutes bias come from the literature on expert judgment, judgment theory, decision analysis, and knowledge acquisition.

Two Views of Bias

The first view of bias, sometimes termed **motivational bias**, proposes that bias occurs when the expert's reports of his thoughts or answers are altered by the elicitation process. Thus, if the expert gives a different response than he believes because of comments from the knowledge engineer, this constitutes bias. This view of bias comes from the soft sciences, particularly psychology and ethnology (cultural anthropology). Proponents of this view consider the expert's thinking to be the *gold standard* that they wish to capture through the elicitation or the building of a knowledge-based system (Henrion and Cooley 1987).

The second view of bias, sometimes termed **cognitive bias,** defines bias as occurring when the expert's knowledge does not follow normative, statistical, or logical rules. To illustrate, if an expert would give probability estimates on all outcomes to a problem (previously defined as being mutually exclusive) and these probabilities did not sum to 1.0, this data would be considered biased. This view of bias comes to knowledge acquisition from the fields of decision analysis and statistics (Mumpower, Phillips, Renn, and Uppuluri 1987). The goal of this position is not to mimic the expert's thinking but to improve on it (Henrion and Cooley 1987), such as by bringing the bias to the expert's attention for correction.

Potential Impact of Bias

Bias can degrade the quality of the data, whether the bias is judged from the standard of the expert's problem-solving processes or from the standard of statistics and logic. To illustrate how motivational bias can affect the data, suppose that the interviewer asks the expert if he mentally models the problem using a decision tree structure. The expert may be led to answer "yes" even if he did not use this means of modeling. The interviewer than faces the difficulty of resolving the discrepancies that are likely to arise between the expert's claim and his answers (Meyer, Mniszewski, and Peaslee 1989). Cognitive bias has been found to affect expert judgment and to result in solutions that are not mathematically optimal. For instance, in making judgments people often use simplifying heuristics that skew the answer reached (Tversky and Kahneman 1974, Hogarth 1980).

Because expert data is often used as input to important decisions and computer models, bias can contribute to the problem of *garbage in, garbage out* The same is true in building expert systems.

The problem of bias also needs to be addressed because of its impact on the credibility of a project. Regardless of whether or not bias was expected to pose problems, a study is open to criticism if it has failed to address bias through experimental design. Bias needs to be monitored or controlled and analyzed for its impact.

The topic of expert bias has recently come into vogue. In particular, many have become aware of how the interviewer or others can lead the expert's thinking. For this reason, we have found that outside reviewers are particularly critical of methods where

there has been no attempt to control for these influences. For instance, one of the first criticisms a review panel (Kouts, Cornell, Farmer, Hanauer, and Rasmussen 1987:7) made of the expert judgment methods used in a Nuclear Regulatory Commission project, the 1987 draft version of NUREG-1150, was that "each expert should be free to make this characterization [of the problem areas] independently of decisions by others."

Causes of Bias

While the basic cause of both types of bias is the human being, the exact mechanisms by which bias is induced differ. Motivational bias is caused by our needs, such as for acceptance; cognitive bias is induced by the way in which we process information.

Motivational bias

Motivational bias can occur as a result of the following circumstances: (1) the expert does not report what his solutions or thought processes actually were, (2) the interviewer or knowledge engineer misinterprets the expert's report, or (3) the analyst misrepresents the expert's knowledge.

In the first aspect of motivational bias, **altering of the expert's reporting**, the phrasing of the interview questions could cause the expert to change his descriptions of his thinking. For example, if the interviewer asks the expert if he used subgoal x in solving the problem, the expert may answer "yes," even if he did not, and then he may begin using subgoal x on future problems. The mode in which the expert is asked to respond can also bias the expert's answer if he cannot accurately encode his final judgment in that mode. Some common response modes are probability distributions, continuous linear scales, Saaty's paired comparisons, ratings, and rankings. Then too, the interviewer's verbal or nonverbal responses can influence the expert's thinking. For instance, if the interviewer leans forward, displaying intense interest in something that the expert is saying, the expert may unconsciously respond by exaggerating his statements on this topic. Other experts, in interaction with the expert, can have similar effects on the expert's thinking. Furthermore, experts' descriptions of their thinking can be affected by their perception of how those not physically present, such as clients or supervisors, might view their responses. For example, if an expert judged that his response might irritate his supervisors, he might let that fact influence his reporting of his data.

There are several reasons why the expert's thinking can be influenced by others. First, most people have an emotional need to be accepted and to receive approval (Zimbardo 1983). Second, people are generally unaware of how they make decisions (Hogarth 1980:ix), such as in solving problems. Yet, it is difficult for them to admit ignorance because western scientific tradition assumes that the problem-solving process can be precisely stated (Denning 1986:345). Thus, in elicitation situations, the expert is likely to be responsive to what he believes the interviewer, knowledge engineer, or other experts wish to hear. Their expectations or wishes are communicated, often unconsciously, through their questions, responses to the expert, and body language. The expert is likely to acquiesce unconsciously to suggestions of an acceptable answer or to means by which he might have solved the problem.

The expert's thinking can be biased by another source--**the interviewer or knowledge engineer's interpretation of his thinking.** We, as humans, tend to perceive and interpret incoming information in a selective way that supports what we already believe. Sometimes, this tendency leads to a misinterpretation, or biasing, of the information. For instance, when the expert mentions a new term, we tacitly assume a meaning based on our experience with similar sounding words. Likewise, we may think that we hear the solution that we expect. Knowledge engineers can be particularly prone to this bias because of a method that they have used for learning the expert's field. This method is to study the expert's domain, build mental constructs representing the knowledge, and then to understand the expert's knowledge structure by comparing to one's own. This comparative means of learning the expert's knowledge structure would seem prone to bias because the knowledge engineers interpret the expert's thoughts through the filter of their own.

The expert's thinking can also be altered by the representation of it. Expert data is often modeled for analysis or represented in a computer program. The individual performing these tasks makes many tacit decisions and assumptions about the data's appearance and performance. For example, when an **analyst** selects a model for the data, the model assumes particular properties of the data, such as its distribution. These assumptions may not be valid. Then too, an analyst may have to aggregate multiple and differing expert judgments to provide one input value. Mathematical aggregation schemes often require assumptions, and different ones, like the mean versus the weighted average, can produce different answers. The knowledge engineer also makes decisions about the organization of the knowledge in the system and its representation that determine how the data are implemented.

Cognitive bias

In the definition of cognitive bias, bias is a consequence of the way in which we think. The following are some characteristics of the way that we think.

Humans model their world to understand, predict, and control it (Clancy 1989:11). It is important to note that there cannot be a perfect match between the model that the expert uses in problem solving and the reality being modeled. We selectively take in information, often perceiving those data that support rather than contradict our beliefs.

Probably the human mind is limited in how much information that it can process and in how much it can remember Hogarth (1980:9). In order to reduce the cognitive burden, people tend to take short cuts when solving a complex problem. Thus they start with a first impression and integrate the information in a sequential manner, making only a few minor adjustments. Later, if additional information is received, they probably will not adjust their initial impression to give a more accurate judgment. In other words, if an individual who has already reached an initial solution is given contradictory data, he will probably not take this data sufficiently into account when generating a final answer. In particular, this sequential means of integrating information handicaps us in making predictions where large or sudden changes are involved. This limiting effect is called anchoring or anchoring bias.

The human mind has limited memory capacity for information processing. As Miller (1956) has noted, most individuals can not discriminate between more than seven things at a time. This limitation in information processing causes people to be inconsistent

in working through the problem. For instance, people commonly forget an assumption made earlier and contradict it, thus causing inconsistency bias.

Then too, some data is more easily recalled than others. For instance, data involving catastrophic, familiar, concrete, or recent events may be easier to recall (Cleaves 1986, Gold 1987). This effect, termed availability bias, can lead to the overestimation of the frequency of some events.

List of Selected Motivational and Cognitive Biases

The following is a brief list of the biases that we have commonly encountered during the process of elicitation and analysis. For convenience, they have been separated into two categories, cognitive and motivational. Motivational biases, such as those produced by social pressure, have as their source the emotional needs and wishes of the expert. Cognitive biases have as their source the workings of the human mind. As mentioned earlier, experts and people in general can unconsciously conform to other's views because of their need to be accepted and receive approbation. [For a more thorough catalogue of biases, particularly of the cognitive variety, see table 9.2 of Hogarth's *Judgement and Choice* (1980)].

Motivational biases

SOCIAL PRESSURE is the altering of the expert's descriptions of his thoughts arising from the desire to be accepted and to see himself in the most positive light. This altering can take place consciously or unconsciously. The social pressure can come from those physically present, such as the interviewer or other experts, or from the expert's internal evaluation of others' reactions.

When the social pressure comes from other experts in a face-to-face group situation, the resulting bias is termed **group think**. Group think occurs when an individual alters his thoughts or his reporting of his thoughts to conform to the group judgment or to the judgment of someone respected in the group. For example, Janis' study of fiascoes in American foreign policy (1972) illustrates how presidential advisors often silently acquiesce rather than critically examine their own thoughts or those that they believe to be the group judgment. Group think is more likely to be a problem if the members of the group have worked together before, if they have qualms about bringing up conflicting points of view, or if there is a dominating leader (Meyer 1984). The tendency toward group think has also been called the *bandwagon* or the *follow-the-leader* effect.

When the source of the social pressure is the interviewer or persons who are not physically present, the bias may be given the general label of **impression management.** The interviewer's verbal and nonverbal responses can lead the individual. Even when the interviewer is not physically present, the individual may try to answer in such a way as to bring the most approbation (e.g., from the person who has written the

questions). Then too, he may try to respond in such a way as would be acceptable to his employer or to society in the abstract.

MISINTERPRETATION is the altering of the expert's thoughts as a result of the methods of elicitation and documentation. (See chapter 2, *Pitfalls: Interviewers, Knowledge Engineers, and Analysts Can Introduce Bias.*) While this effect is prevalent, it has not received much attention. Misinterpretation occurs when the elicitation is done from the interviewer's, rather than the expert's, viewpoint. For example, we have all had the frustrating experience of trying to force fit our views into the limited response options of a mail survey. Additionally, misinterpretation frequently occurs as a result of the response mode. If the expert can not adequately translate his judgment into the response mode, misinterpretation will result. We have noticed that experts seem to have more difficulty with the response modes of probability distributions, ranks, and percentiles.

MISREPRESENTATION is the altering of the expert's thoughts or answers as a result of modeling or analyzing this data. The person who is modeling the expert data for entry into a computer program or for analysis makes tacit assumptions about the data's appearance and performance. For example, the analyst might assume that the expert data were normally distributed. If these assumptions are not warranted, the expert data will be misrepresented.

WISHFUL THINKING is caused when the expert's hopes or involvement in the area on which he is being questioned influence his response (Hogarth 1980). For example, people frequently give overly optimistic estimates of what they can accomplish in a given amount of time because of wishful thinking (Hayes-Roth 1980). The wishful thinking effect is strongest when the subjects are personally involved or would somehow gain from their answers. Hence, this bias is also called *conflict of interest.* For example, conflict of interest could occur if an expert was asked to evaluate the services provided by several companies, one of which employed him or had offered him money for a good evaluation.

Cognitive biases

INCONSISTENCY is the inability to be consistent in solving a problem, especially through time. Of all of the biases mentioned here, this is the most common. Individuals often unintentionally change definitions, assumptions, or algorithms that they meant to hold constant throughout the problem. These inconsistencies may result in answers that do not make logical or Bayesian sense. For example, a series of answers proposing that factor A was more critical than B, B more than C, and C more than A would not make logical sense. Similarly, if an expert gave the same probability of A for two situations, one of which involved an influential factor C and one which did not, his answers would not be coherent from the Bayesian viewpoint.

ANCHORING is the failure to adjust sufficiently from a first impression in solving a problem. For example, members of a group of experts often discuss the issue before giving their final estimates. A member's last

estimate is likely to be closer to his initial impression than it would be had he fully taken into account the factors discussed.

AVAILABILITY refers to the differing ease with which events can be retrieved from LTM. Data involving catastrophic, familiar, concrete, or recent events may be easier to recall (Meyer 1986). Availability bias affects people's ability to accurately estimate frequencies and recall other aspects of the event. For example, the incidence of severe accidents in reactors tends to be overestimated, in part, because of its catastrophic and newsworthy nature.

UNDERESTIMATION OF UNCERTAINTY is the failure to account for the actual amount of uncertainty in the answers given. For example, when people are asked to put a range around an answer such that they are 90% sure that the range encompasses the correct answer, their ranges only cover 30 to 60% of the total (Capen 1975). Even when people are given quizzes and feedback on their performance, they cannot break the barrier of covering only 70% (Capen 1975:846). A popular explanation for this effect is that we are uncomfortable with the amount of uncertainty in life, and thus minimize it. In particular, we may avoid confronting the large uncertainties in our judgments.

In summary, it is difficult to judge the impact of these and other biases on expert data because there are few bases of comparison. One cannot compare the expert's data to what it would have been before the bias occurred. Similarly, one cannot judge the degree of bias by comparing the expert's judgment to the *right* answer because generally the *right* answer is not known. While the impact of such biases may not be discernible, the relative frequency of particular biases is more obvious. The biases most likely to be encountered are those resulting from the expert's *inconsistencies*. The next most common bias, in our experience, has been that of *anchoring*. We have observed experts using the anchoring and adjustment heuristic to allow them to solve complex problems. By contrast, we have not found *group think* bias to be as common in our research as the literature would have led us to expect. Similarly, *wishful thinking* has only emerged in a few projects where the experts were managers having to forecast whether their projects would reach various milestones on schedule (Meyer et al. 1981).

Countering or Reducing Bias--More Art Than Science

Approaches to handling bias are rare and in their early stages. They are perhaps as much art as they are science. While research on judgment has drawn attention to the presence of expert bias, little has yet been done to deal with its occurrence during elicitation (Cleaves 1986:9-9). Typically, practitioners have developed their own means for dealing with the biases they commonly encounter. The program proposed here is no exception. This section and discussions in later chapters on handling bias should be viewed as reflecting the authors' experiences.

One reason that there are not more programs for handling bias is that bias is a difficult topic to study. Studying, much less trying to counter bias, is complicated by not having a readily available baseline by which to determine the direction and magnitude of the

bias. For the questions that expert knowledge is elicited, there are frequently no known single *right* answers or empirical data. Thus, the expert's data cannot be simply compared to an answer looked up in a reference or to the result of modeling data. Measurement of motivational bias is further complicated by the absence of objective standards for comparison. At least with cognitive bias there are traits of the expert's answers that can be compared to standards. For example, do the expert's exclusive but dependent probabilities sum to zero as they should? With motivational bias, the baseline of the expert's knowledge, had its description not been altered, is difficult to determine, especially because expertise is not static. As Rosenfield (1988:194) argues in his book on memory, higher mental functions are not fixed procedures but subject to constant reconstruction. While we recognize that both biases are difficult to detect, we believe that for progress to occur programs like this one must be proposed and applied to expert judgment.

Our approach differs from the one presented by Cleaves (1986). First, Cleaves' proposal focuses on cognitive bias; second, Cleaves tries to anticipate biases by the judgment processes in which they are likely to occur--namely, hypothesis and solution generation, decision rule articulation, uncertainty assessment, and hypothesis evaluation. While we agree that biases occur during these processes, we try to anticipate the biases by the elicitation components that are likely to exhibit them. We assume that many readers will be lay persons in the areas of human cognition and that this approach may be easier for them, at least as a starting point.

Another major difference in our program is its real-time emphasis. Given the evolving nature of expertise, bias is best detected when it is being elicited. This program stresses monitoring for bias, particularly motivational bias, in real time rather than mathematically compensating for it afterwards. It is much more difficult to determine that motivational bias has occurred after the elicitation because the baseline--the expert's knowledge--is likely to have changed. For this reason, we consider each elicited datum to be a snapshot that can be compared to a snapshot of the expert's state of knowledge at the time of the elicitation.

Steps in a program for handling bias

Our program (Meyer and Booker 1989) consists of these general steps:

1. **Anticipate the biases to which the planned elicitation is prone and redesign the elicitation, as needed.** We have provided lists of selected biases and the situations in which they are likely to occur (see chapter 8). These biases were selected because they represent a range of bias within the two definitions. Certainly these are not the only sources of bias, but they are ones that we have most frequently encountered.

   To anticipate bias, the interviewer determines the parts of her planned elicitation and searches the *Index of Selected Biases* (see the table at the end of chapter 3) for the biases to which they are prone. For instance, if the researcher planned to use the interactive group method, she would see that it was prone to the following biases: social pressure from group think, wishful thinking, and inconsistency. The interviewer then turns to the section following the table *Definitions of Selected Biases* to look up the definitions and causes of the

selected biases. The definition section can further be used to redesign the elicitation, if the researcher, as a result of anticipating bias, wishes to do so.

Through the process of looking up these biases in the *Index* and *Definitions*, the data gatherers will become aware of the biases existence and of their own tendencies to introduce or acerbate them. In addition, the section entitled *Pitfalls: Interviewers, Knowledge Engineers, and Analysts Can Introduce Bias* (chapter 2) can be read to enhance the project personnel's awareness of bias.

2. **Make the experts aware of the potential for introducing bias and familiarize them with the elicitation procedures.** The experts need to be informed (as described in chapter 10) about the biases that they are likely to exhibit given the elicitation situation. In particular, they need to know the definitions and causes of these biases. Without this information, the experts will not be able to combat their own tendencies toward bias. The interviewers can use the *Index* and *Definitions* provided for step 1 as a base for informing the experts about bias.

   It should be noted that making the experts aware of the biases helps but does not completely alleviate the problem. In some cases, the cause of the bias, such as with the underestimation of uncertainty, is too ingrained to be completely overcome. In other cases, the experts will not make the needed effort to counter the natural tendency toward bias. People typically believe that others, not themselves, will suffer from the biases described. With some biases, such as anchoring and underestimation of uncertainty, the experts can participate in tests designed to evoke the bias. Frequently, almanacs are used to construct test questions, such as: *How much rain fell in St. Paul, Minnesota in 1987?* While the experts will not know the answers to such questions, the interviewer can look up the correct answer. The interviewer can read the answers and allow the experts to score their own. (This procedure is described further in chapter 10, *Introducing the Experts to the Elicitation Process: For an Interactive Group Situation.*) Such a demonstration is often necessary to convince the expert that he too is prone to the bias.

   The experts also need to be made aware of the elicitation procedures. If they are confused about how and when they are to respond, the data gathered as well as the expert's cooperativeness is negatively affected. One aspect of elicitation that is often confusing is the response mode, if the expert is not accustomed to using it. The use of unfamiliar response modes should be rehearsed by the experts during the training session. Information on how to familiarize the expert with the elicitation procedures is given in chapter 10.

3. **Monitor the elicitation for the occurrence of bias.** Prior to the elicitation sessions, the data gatherer looks up the signs that the biases may be occurring in *Signs of Selected Biases* below. For instance, if group think bias was anticipated, the data gatherer would look up this bias in the *Signs* section and read about indications of its presence. One sign of group think is that the experts appear to defer to another member of the group or to each other. The interviewer, knowledge engineer, or a trained observer then watches for this sign of group think during the elicitation. In general, monitoring biases, as

described here, requires that the experts verbalize their thoughts and answers. Without this feedback, we have found the monitoring to be much more difficult.

4. **Adjust, in real time, to counter the occurrence of bias.** In this step, the interviewer looks up the suggestions for preventing a particular bias in *Suggestions for Countering Selected Biases*. These suggestions vary because we have used two approaches: (1) controlling those factors contributing to a particular bias, or (2) employing the opposite bias. The first approach involves controlling the factors that contribute to the bias. For instance, fatigue is a factor that leads to increased inconsistencies in expert judgment. The interviewer can stop the elicitation sessions or schedule breaks before the experts become fatigued as a means of controlling this contributor to inconsistency. The basis of the second approach, fighting bias with bias, comes from Payne (1951), the grandfather of survey design. Payne believed that all interviewing was biased and that one should therefore aim for equal but opposite biases. An example of this technique is to try to have experts anchor to their own judgments in attempts to counter a group-think situation. Having the experts write their judgments encourages them to form and keep their own opinions even when they hear the opinions of others.
5. **Analyze the data for the occurrence of particular biases.** (Suggestions on how to test for bias are given in chapter 16.) If step 5 is the only step of the program being followed, the analysis will necessarily be simpler than if step 4 were also followed. If the steps 3 and 4 were followed, they would provide the additional data needed for performing more complex analyses. In general, adequately testing for one of the motivational biases requires this more complex testing. Occurrence of a cognitive bias, such as the underestimation of uncertainty, can often be determined by simple mathematical tests. For example, we analyzed the expert's ranges on their likelihoods of reaching particular magnetic fusion milestones. Their ranges were within one standard deviation of their pooled answers. This result indicated that the experts thought that they were adequately accounting for all of the uncertainty when they were only accounting for about 60% or less of it (Meyer et al. 1982).

Determining which steps to apply

The steps of the program above can be applied in sequence or singly, depending on the needs of the project. For example, if information on bias was not important to the project, none of the steps or only step 5, analyzing for bias, would be necessary. If on the other hand, the interviewer wished to follow some but not all of the steps, she could perform steps 1 and 5, or 1, 2, and 5, or 1, 3, and 5, or 1, 2, 3, and 5. We suggest that step 5 always be done regardless of the other steps because it provides a general check on the expert data.

To pick steps for use in a project, consider the following: (1) the reason for addressing the problem of bias; which view of bias, motivational or cognitive, will be employed; and (3) which biases are of special interest.

The reason for focusing on bias

The reason for focusing on bias can provide a criteria for determining which steps of the program to implement. For example, if the project personnel's interest in bias stems from a desire to avoid having reviewers criticize aspects of the project, their selection of steps would probably differ from those whose goal is to study the occurrence of bias. In the first example, all the steps might be used. In the second, steps 1, 3, and 5 might be used to anticipate which biases are likely, to design the study around the factors likely to affect their occurrence, to monitor the elicitation sessions for their appearance, and to analyze the data for their occurrence.

The steps are suited to accomplishing different aims. For instance, steps 1, anticipate the biases, and 2, make the experts aware of the potential for bias and familiarize them with the elicitation procedures, accomplish educational goals. Performing them gives the data gatherers and the experts a preview of how the data will be elicited and an understanding of why it will be conducted in a particular manner. Thus, these two steps, especially 2, are often used as practical steps for introducing the participants to the elicitation process and making them comfortable with it. Step 3, monitoring for the presence of bias, does not interfere with the elicitation or the results, so it suited to researching the presence of bias, as is step 5, analysis. In contrast, step 4, adjusting in real time to counter bias, affects the elicitation and, thus, is more appropriate to situations where the intent is to control, rather than study, bias.

The selection of the view of bias, motivational or cognitive

The selection of the view of bias, motivational or cognitive can also identify which set of steps will be most effective.

We ask the reader to *select one view of bias* because, while both views are equally valid definitions of bias, one way of construing bias may be more useful for a particular project. For example, if the focus of the project is learning the expert's problem solving in order to emulate it, the motivational definition of bias would be more appropriate. If the project involves estimating the likelihood of future events, the cognitive definition would be a natural choice. People are inaccurate in making predictions and the cognitive view of bias would help to combat this weakness. We suggest that the reader select and use only view of bias at a time to avoid being contradictory. For example, use of the cognitive definition would propose that a mathematically incorrect judgment be modified. This act would cause a misrepresentation of the expert's data, a bias, according to the motivational definition of bias.

As a general rule, if the motivational definition of bias is selected, steps 1 and 2 will be most helpful. Following step 1, anticipating the biases and redesigning the elicitation, would allow the project personnel to tailor elicitation methods (as described in chapter 8) to reduce their tendency to *lead* the expert or to misinterpret his data. Use of steps 1 and 2 (making the experts aware of the potential for bias and familiarize them with the elicitation procedures) would inform the data gatherers and experts about bias, and hopefully make them less prone to it. In this book, we have focused on presenting methods of elicitation and analysis which we believe minimize influencing the experts and force fitting their data. Thus, just using the methods suggested in this book, regardless of any program for handling bias, should provide some protection from motivational bias.

If the cognitive definition of bias is selected, step 5, analysis, is particularly effective. Analysis is generally more effective with cognitive than motivational bias because cognitive bias can be determined mathematically or statistically. Cognitive bias can usually be measured because it is defined as a violation of logical or statistical standards. Thus, one of the cognitive biases, underestimation of uncertainty, can be analyzed using the experts' ranges on their estimates (as described in chapter 17).

Interest in particular sources of bias

Interest in particular sources of bias will favor the use of some steps over others. The reader will have to identify which biases he or she is most interested in and then select those steps that address these biases. For example, if social pressure by the interviewer was a concern, use of steps 1 and 2 would be helpful. In step 1, the interviewer would be led to select elicitation methods (as described in chapter 7) that were nondirective, like the verbal protocol, to minimize her tendency to influence the expert. In step 2, the expert would be made aware of this bias so he would be able to guard against it. Another bias that people commonly worry about is wishful thinking. The step that we have used to deal with wishful thinking is step 1. In anticipating this bias, we have redesigned the elicitation to select those experts less likely to exhibit wishful thinking, those who had less at stake in the judgments. Additionally, we required that the experts explain their reasoning for their estimates to make it more difficult for them to give highly optimistic estimates. A third bias that often draws attention is group think. All of the steps would be helpful in countering group think. However different steps could be used depending on the project personnel's reason for focusing on this bias. If the intent was to study group think bias, the following steps might be used: step 1 to design the experiment, step 3 to monitor the presence of the bias, and step 5 to analyze its occurrence. On the other hand, if the purpose was to try to eliminate group think bias, more steps could be used. For instance, use of step 1 could lead the project personnel to redesign the elicitation situation (as described in chapter 8) to preclude expert interaction. In a less extreme case, step 1 might simply be used to anticipate the occurrence of this bias in an interactive group setting, step 2 to alert the experts to this danger, step 3 to look for the signs of its occurrence, and step 4 to take the suggested actions.

Index of Selected Biases

Elicitation Component	View of Bias	Source
Elicitation Situations:		
Individual Interview	Motivational	Social pressure from interviewer
	Motivational	Wishful thinking
	Cognitive	Inconsistency
Delphi	Motivational	Wishful thinking
	Cognitive	Inconsistency
	Cognitive	Anchoring
Interactive Group	Motivational	Social pressure, group think
	Motivational	Wishful thinking
	Cognitive	Inconsistency
Response Modes		
Complex ones such as probabilities, Bayesian updating, and uncertainty measures	Motivational	Misinterpretation by expert
	Cognitive	Inconsistency
	Cognitive	Underestimation of uncertainty
Aggregation		
Behavioral Aggregation	Motivational	Social pressure, group think

Mode of Communication	View of Bias	Source
Face-to-Face	Motivational	Social pressure from interviewer
	Motivational	Wishful thinking
	Cognitive	Underestimation of uncertainty
Telephone	Motivational	Social pressure from interviewer
	Motivational	Wishful thinking
	Cognitive	Availability
	Cognitive	Anchoring
	Cognitive	Underestimation of uncertainty
Mail	Motivational	Social pressure, impression management
	Motivational	Wishful thinking
	Motivational	Misinterpretation by analyst
	Cognitive	Inconsistency
	Cognitive	Availability
	Cognitive	Anchoring
	Cognitive	Underestimation of uncertainty

Definitions of selected biases

Motivational Bias--Altering of the expert's thoughts through social pressure, wishful thinking, or misinterpretation.

Social pressure. Social pressure is the altering of the expert's thought processes or descriptions of those thoughts arising from the desire to be accepted and be seen in the most positive light possible. This altering can take place consciously or unconsciously. The social pressure can come from those physically present, such as the interviewer or the other experts, or from the expert's own internal evaluation of others' reactions.

Social pressure from the interviewer is most likely to occur in those methods where the interviewer is meeting in a face-to-face situation with the experts, such as in the individual interview and the interactive group. In face-to-face situations, the interviewer can intentionally or unintentionally influence the expert through body language, facial expression, intonations, and choice of words. We expect this source of bias to be weaker in telephone conversations and weaker still in communications by mail. These last two modes do not allow some of the above-mentioned means of expression that the face-to-face mode does. In addition, social pressure bias is more pronounced when the interviewer is asking leading questions. Thus, it is weaker when the interviewer is using the verbal protocol, verbal probe, or ethnographic technique. The verbal protocol avoids leading the expert because it does not involve questioning him; the verbal probe uses general, nonleading phrases; and the ethnographic techniques uses the expert's own words in formulating questions.

Social pressure from others in the group induces individuals to slant their responses or to silently acquiesce to what they believe will be acceptable to their group (Meyer 1986: 89). The psychologist Zimbardo (1983) explains that it is due to the basic needs of people to be loved, respected, and recognized that they can be induced or choose to behave in a manner that will bring them affirmation. There is abundant sociological evidence of conformity within groups (Weissenberg 1971). Generally, individuals in groups conform to a greater degree if they have a strong desire to remain a member, if they are satisfied with the group, if the group is cohesive, and if they are not a natural leader in the group. Furthermore, the individuals are generally unaware that they have modified their judgment to be in agreement with the group.

Group think. One mechanism for this unconscious modification of opinion is explained by the theory of cognitive dissonance. **Cognitive dissonance** occurs when an individual finds a discrepancy between thoughts that he holds or between his beliefs and actions (Festinger 1957). For example, if an individual holds an opinion that conflicts with that of the other group members and he has a high opinion of the other's intelligence, cognitive dissonance will result. Often the individual's means of resolving the discrepancy is by unconsciously changing his judgment to be in agreement with that of the group (Baron and Byrne 1981). For example, Janis' study of fiascoes in American foreign policy (1972) illustrates how presidential advisors often silently acquiesce rather than critically examine what they believe to be the group judgment. This phenomena has also been called *group think bias*, the follow-the-leader or bandwagon effect.

Group think is only likely to be a concern in an interactive group situation. It is further likely to occur in situations where behavioral aggregation is used because this type of aggregation requires that pressures toward conformity be encouraged.

Impression management. Another type of social pressure can occur as the expert imagines the reactions of those not physically present. This effect can occur in any elicitation situation. However, it seems to be more noticeable when it is not covered by other effects, such as social pressure caused by the interviewer. For this reason, its occurrence is most noted in mail surveys. The individual may try to answer in such a way as to bring the most approbation, such as from the person who has written the questions. Then, too, he may try to respond in a way that would be acceptable to his employer or to society in the abstract. For this reason, this source of social pressure has been termed impression management (Goffman 1959). Payne (1951) has found evidence of individuals giving the responses that they perceived to be the most socially acceptable rather than those which accurately portrayed their thoughts or actions. For example, on surveys, people claim that their educations, salaries, and job titles are better than they are. Often there is a 10% difference between what is claimed for reasons of prestige and what objectively is (Meyer 1986: 90).

Wishful thinking. Wishful thinking occurs when an individual's hopes influence his judgment (Hogarth 1980). What the subject thinks should happen will influence what he thinks will happen. To illustrate, presidential election surveys show that people predict the winner to be the candidate that they expect to vote for (Armstrong 1981: 79). The above instance is one where the subjects stand to gain very little personally from their answer. The wishful thinking effect is stronger where the subjects are personally involved or would gain from their answers. Hence, this bias is also called conflict of interest. In general, people exhibit wishful thinking about what they can accomplish in a given amount of time: they overestimate their productivity (Hayes-Roth 1980).

Wishful thinking is not particular to any elicitation method. Instead it relates to selection of experts and the assignment of them to specific questions or problems. If they have a special interest in the answer, wishful thinking is likely to occur whether the individual interview, interactive group or Delphi elicitation method is used or whether the communication is face-to-face, by telephone, or mail.

In general, wishful thinking effects will be most pronounced when the expert does not have to explain his reasoning. The experts' highly optimistic responses are checked by having him disaggregate the problem and explain his problem solving. For example, Hayes-Roth (1980) found that having people break down the tasks that they had earlier thought they could accomplish in a given time led to more realistic estimates.

Misinterpretation. Misinterpretation is the altering of the expert's thoughts as a result of the methods of elicitation and documentation. While this effect is prevalent, it has not received much attention. (For further information, see the pitfall, *Interviewers, Knowledge Engineers and Analysts Can Introduce Bias*). Frequently misinterpretation occurs as a result of the response mode. If the expert can not adequately translate his judgment into the response mode, misinterpretation will result. We have noticed that experts seem to have more difficulty with the following response modes: probability distributions, ranks, and percentiles.

Misinterpretation is also more likely with elicitation and documentation methods that are written from the interviewer's, rather than the expert's, viewpoint. For example, we have all had the frustrating experience of trying to force fit our views into the limited response options of a mail survey.

Cognitive bias--Data failing to follow mathematical and logical standards because of inconsistency, anchoring, availability, or underestimation of uncertainty.

Inconsistency. Inconsistency is the inability to be consistent in solving of problems, especially through time. Of all of the biases mentioned here, this is the most common. Individuals often unintentionally change definitions, assumptions, or algorithms that they meant to hold constant throughout the problems. Inconsistency in an individual's judgment can stem from his remembering or forgetting information during the elicitation session. For example, the individual may remember some of the less spectacular pieces of information and consider these in making judgments later in the session, or the individual may forget that particular ratings were only to be given in extreme cases and begin to assign them more freely toward the end of the session.

As Dawes, Faust and Meehl (1989:1671) have noted, such factors as fatigue, recent experience, or seemingly minor changes in the ordering of the information or in the conceptualization of the task "can produce random fluctuations in judgment. Random fluctuation decreases judgmental reliability and hence accuracy." These inconsistencies may result in answers that do not make logical or Bayesian sense. For instance, a series of answers proposing that factor A was more critical than B, B more than C, and C more than A would not make logical sense. Similarly, if an expert gave the same probability of A for two situations, one of which involved an influential factor C and one which did not, his answers would not be coherent from a Bayesian viewpoint.

The natural tendency toward inconsistency is acerbated by several conditions such as memory problems, confusion, and fatigue. During elicitation sessions of more than 30 minutes, people often forget the instructions, definitions, or assumptions that they were requested to follow. For example, the experts may forget that a rating of nine meant a near certainty and assign it more easily than the definition specified. Thus, unstructured elicitations, which do not have periodic reviews of the question information, are more likely to have high inconsistency. This inconsistency can be between experts' answers (e.g., the experts meant different things by the same numerical answer) or within an expert's answer (e.g., sometimes the expert gave a specific rating more easily than at other times). Also, situations where the expert's understanding through time cannot easily be monitored are more prone to inconsistency. These situations include the Delphi or mail survey.

Confusion can also lead to inconsistency. Thus, any of the more complicated response modes, such as probability distributions and percentiles, are more prone to this problem. This confusion is why training the expert in the use of these modes is recommended. In addition, if the experts must mentally juggle more than five to seven things, such as in rating them, they are likely to become confused and inconsistent. It is for this reason that the Saaty paired-comparison mode is used even though it is more time consuming than some of the other response modes.

Anchoring. Anchoring is the failure to adjust sufficiently from one's first impression in solving a problem. We would rate it next to inconsistency in terms of frequency of occurrence. Sometimes this tendency is explained in terms of the Bayesian philosophy as peoples' failure to adjust a judgment in light of new information in the manner specified by Bayes Theorem (Meyer 1986:88). Spetzler and Stael von Holstein (1975) and Armstrong (1981) describe how people tend to anchor to their initial response, using it as the basis for later responses. Ascher (1978) has found this problem to exist in

forecasting where panel members tend to anchor to past or present trends in their projection of future trends. Ascher determined that one of the major sources of inaccuracy in forecasting future possibilities, such as markets for utilities, was the extrapolation from old patterns that no longer represented the emerging or future patterns. Another example of anchoring occurs when a member of a groups last estimate is closer to his initial impression than it would be had he fully taken earlier group discussions into account.

Anchoring is most prevalent in situations where the expert is not likely to experience the opposite bias of being influenced by the interviewer or the group, such as in the Delphi method. In addition, those modes, such as mail or telephone communications, where the expert's thoughts cannot be easily monitored by having the expert think aloud, are prone to this bias. In addition, we have noted that experts are more likely to stick with their anchor if they have either described it orally or in writing and fear losing face for changing their mind.

Availability. Availability bias arises from the differing ease with which events can be retrieved from long-term memory. Data involving catastrophic, familiar, concrete, or recent events tend to be easier to recall. Availability bias affects people's ability to accurately estimate frequencies and recall other aspects of the event. For example, the incidence of severe accidents in reactors tends to be overestimated in part because of their catastrophic and newsworthy nature.

Availability bias is more common when the expert does not receive any information from others and, thus, does not have a chance of triggering other, less accessible, memory associations. For this reason, the individual interview is the most prone to availability bias, and the interactive group, the least. With individual interviews, a series of different scenarios is often used to help the expert enlarge on the sample of things contributing to his final answer.

Availability bias is also more common with telephone and mail modes of communication because the expert is usually not given much background before being asked point blank for the answer. A structured hierarchical presentation of the information, such as from the general to the specific, can alleviate this weakness.

Underestimation of Uncertainty. People will underestimate the amount of uncertainty in the answers that they give. For example, when people are asked to put a range around an answer such that they are 90% sure that the range encompasses the correct answer, their ranges only cover 30-60% of the dispersion (Capen 1975). Even when they are given quizzes and feedback on their performance, they cannot break the barrier of covering only 70% (Capen 1975:846). A popular explanation for this effect is that we are uncomfortable with the amount of uncertainty in life, and thus, minimize it. In particular, we may avoid confronting the large uncertainties in our judgments.

Although this effect is very widespread, Martz, Bryson, and Waller (1985:72) have noted that it is more pronounced with probability and chance estimates than with some of the other response modes. Chance estimates, also called odds, are given as 1 chance in a total, such as 1 in 1000.

Signs of selected biases

Group think. There are several signs that a group-think situation may be developing. Generally, no difference of opinion is voiced, and the experts appear to defer to another member of the group or to each other (Meyer 1986: 95).

Wishful thinking. Wishful thinking is indicated if the experts were previously judged to have something to gain from their answers and if the answers were given quickly with very little thought.

Inconsistency. A number of signs can indicate inconsistency. The interviewer can hear many of these, if the experts are verbalizing their thoughts and answers. In particular, she can detect when a response mode or rating is being applied more easily through time (Meyer 1986:94). Experts tend to apply the extremes of a rating scale more easily as they become fatigued. The interviewer can also hear when the expert is contradicting an assumption that he made earlier. For example, a tank expert chose two very different routes through the mapped terrain because the second time he unconsciously assumed that his company was the main effort and had to push hard.

Inconsistency can also be monitored by the use of Bayesian-based scoring and ranking techniques. During the elicitation, the expert's judgments can be entered into a scoring and ranking program, such as that of Saaty's Analytical Hierarchical Process (1980), to obtain a rating of their consistency. Then, if the inconsistency index from this method is too high, indicating significant inconsistency, the experts can redo their judgments as described in step 4.

Availability. A potential problem with availability bias is indicated if the expert does not mention more than one or two considerations in giving his answer. If the expert only considers a few things, these were probably the most easily remembered and the answer is likely to be skewed to reflect these few.

Anchoring. Anchoring bias can be suspected if the experts receive additional information from experts or other sources during the elicitation but never waiver from their first impression. For example, reactor code experts were asked to compare the performance of their computer codes to plots of experimentally generated data. Often they commented on their first impression. When they examined the plots more closely, they typically found places where the computer code did not capture the experimental phenomena. However, the experts usually simply adjusted upward or downward of their initial assessment rather than revising it completely (Meyer and Booker 1987b).

Suggestions for countering selected biases

Group think. Social pressure from group think can be countered using techniques from two approaches (Meyer 1986:95-96). Using the first approach, the interviewer can try to prevent those factors that contribute to group think. For instance, the interviewer can stop the elicitation and warn the group members about group think. If there is an official or even a natural *ex officio* leader in the group, that individual can be asked to give his responses last, or privately, so as not to influence the other group members. In addition, if someone other than the interviewer has been leading the group meeting, he can be encouraged to be nondirective during the meetings. An explanation of the group-think phenomena usually convinces the leader that better discussions and data will result from their avoiding leading.

The other approach is to try to counter the effects of group think with an opposite bias--anchoring. One technique for fostering anchoring is to require the group members to write down their judgments and reasoning. In this way, they are more likely to anchor to their own judgments rather than silently acquiesce to someone else's. If the experts are to discuss their judgments, each person can record and report his before the floor is opened to

discussion. Once individuals have publicly announced their view, they are unlikely to spontaneously modify it. (They will still modify their view if someone raises a valid point that they had not previously considered.)

Wishful thinking. The tendency toward wishful thinking can be countered by making it more difficult for the expert to indulge in it. If the expert must explain his answer in detail, it will become apparent whether there was any objective basis for his response.

Inconsistency. Inconsistency can be reduced by using two techniques.

The first technique is to address the aspects of the elicitation that are contributing to the inconsistency.

As mentioned earlier, fatigue is a contributor to inconsistency. If the interviewer has noted that the experts are becoming more inconsistent with time, she can quickly end the meeting or schedule a break. In general, two hours is the maximum amount of time that experts participate in discussion or problem solving before becoming tired. (Experts often signal their fatigue either by briefer responses or by leaning way forward or back in their chairs.)

Another contributor to inconsistency is faulty memory. If at the beginning of every session the statement of the question, definitions, assumptions, and response mode are reviewed, the experts will be more consistent in their judgments (Meyer 1986: 96). They will be more consistent between and within themselves. In addition, if there is much time between this first review and when the experts' judgments are requested, the question can be worded to include some of the above information. For example, *What rating would you give to the importance of element X over Y to the reaching of goal Z?* If they are using a response mode, in this case a Saaty paired comparison, they will need to have the definitions of the scale available in front of them.

A second technique for reducing inconsistency is to have the group members monitor their own consistency (Meyer 1986:96). This techniques was successfully used in a simple interactive group elicitation where the experts were able to watch the interviewer's monitoring of inconsistency and then mimic it. (Meyer, Peaslee, and Booker 1982). The experts were given copies of a matrix of the elements being judged, the criteria on which these elements were being judged, and their past judgments. When experts monitor their own consistency they may wish to change an earlier judgment to be in line with their current thinking. If their reasoning does not violate the logic of the model or the definitions, they can be allowed to make the change. Often in this process the expert may discover that he had forgotten to include some pertinent information. After this addition, some of the judgments may need to be redone.

If Saaty's Analytic Hierarchy Process (1980) had been used and its results indicated high inconsistency, the experts could review and redo the affected judgments.

Availability. Availability bias can be countered by stimulating the expert's memory associations. In general, group discussion will cause the expert to think of more than just the first readily accessible information. In addition, free association can be introduced to single experts or those in groups. (Free association is having the expert or experts quickly generate any and all elements that might have bearing on the question (Meyer 1986:94)). Free association is similar to brainstorming or the Crawford Slip method (Boose and Gaines 1988:38). The experts are asked to refrain from being critical in order to generate the widest possible pool of ideas. (The number of ideas is later

narrowed to those judged to be most pertinent to the question.) A related technique is to hierarchically structure the presentation of question information so that it flows from the general to the specific. In this way, the expert is able to consider the pertinent information before having to reach a solution. Again this strategy is to fire as many memory associations as possible so that the maximum number of relevant ones will enter into the expert's final judgment.

Anchoring. Techniques similar to those used to counter availability bias are used to counter anchoring. In particular, giving the expert input from other experts as in a Delphi situation or an interactive group makes it more difficult for the expert to anchor to his first impression. Another technique is to ask the expert for extreme judgments before getting his likely ones (Cleaves 1986:9-10).

PART II

ELICITATION PROCEDURES

4

Selecting the Question Areas and Questions

The process of selecting questions that will be asked of the expert is a slow and evolutionary one. But, as the question list evolves, the selection process seems clearer in retrospect. The purpose in this chapter is to illustrate the steps involved in selecting the questions: (1) definition of the project's purpose, (2) selection of the general question areas, and (3) identification of the specific questions. In particular, information is given on which persons--clients, data gatherers (interviewers and knowledge engineers), analysts, or experts--can help with each of these steps and how. In addition, in this chapter we cover when selecting and motivating the external experts need to be done in parallel with the steps mentioned below. Following on with the process, how to refine the questions is the subject of chapter 5.

Steps Involved in Selecting the Questions

Selecting the questions to be asked of the expert is one of a sequence of steps where the information from one step is needed to accomplish the next, more detailed step. The steps are summarized as follows.

Step 1: Defining the project's purpose or goals

The project's purpose is simply what the project is to accomplish. For instance, the purpose of the reactor risk project, NUREG-1150, was to examine the risk of accidents in a selected group of nuclear power plants. The purpose of the project is not always as clear as the one stated in the above-mentioned project. Sometimes, the persons in charge of the project have only a vague idea as to the project's aims, or they are unable to express what they envision the project accomplishing.

Step 2: Selecting the general question areas

A **question area** is a specific issue for investigation. For example, from the above-mentioned reactor risk project, nine question areas were formed. These areas were internal events that could lead to core damage, such as the failure of the emergency core cooling system due to venting or containment failure.

Question areas are developed by considering such information as the goal of the project, the client's directives, and the practicalities of gathering expert judgment on this topic (e.g., whether experts exist and whether their expert judgment would be considered proprietary information). Of the question areas initially considered, only a few may emerge as the final areas.

Step 3: Identifying the questions

Questions are concrete, detailed points within question areas that the experts are asked to answer. The test for whether something qualifies as a question is if the expert finds it sufficiently specific to be answered. If an expert cannot address the question in its present form, it probably resembles a question area more than a question. To illustrate, the question area on an emergency core cooling system failure can be broken into different accident scenarios, each of which could lead to a cooling system failure. The experts can answer the specific questions on the probability of one of the scenarios occurring within a particular nuclear plant.

Sometimes the questions are technical problems that the expert is to solve to allow the data gatherers to examine his problem-solving processes. For example, experts in statistics could be asked to judge whether lists of numbers are random or not as a means of learning their mental rules for determining randomness.

The reader needs to assess which of the above three steps has already been accomplished. Frequently, the defining of the project's goals has already been made by the person who is sponsoring the project. Then too, the reader may have previously decided on the question areas. If one or more of the above steps has been completed, the reader may wish to skip ahead to the next step, *Executing the Steps with the Assistance of Clients, Project Personnel, and Experts*. The section below illustrates the possible variety in project goals, question areas, and questions.

Illustrations of the Variation in Project Goals, Question Areas, and Questions

The project goals, question areas, and specific questions can vary tremendously. The three examples below--on reactor risk, sources of interexpert correlation, and army exports--illustrate what the project goals, question areas, and questions could be in different projects. In particular, these examples can provide assistance in formulating the goals, question areas, and questions for a particular application.

In the first example, the reactor risk project (NUREG-1150) mentioned above, the goal was to perform risk analyses of five different U.S. light water reactors to provide data on the likelihoods of severe accidents and their consequences. The data requirements for this application were large and complex. The areas selected for receiving expert judgment were a reduced set that met the following criteria: (1) they were within the scope of NUREG-1150; (2) they were areas of significant importance to the estimation of risk or the uncertainty of risk; and (3) there were no other sources of data available (Wheeler, Hora, Cramond, and Unwin 1989). Questions were developed for each area. For example, an area of investigation in the risk project was the failure of a Westinghouse reactor coolant pump's shaft seals under station blackout conditions. A question asked in this area: "What

is the failure probability per year for severe seal leakage?" This question included definitions of *severe leakage* and the sequence of events leading to this failure.

In the second example, a study of interexpert correlation, the goal was to determine if expert's answers correlated, and, if so, what caused the correlation. Given this general goal, the question areas, even the field of expertise from which the experts could be drawn, were completely open. The areas selected were those that would be comprehensible to the researchers, have readily available experts, and have nonproprietary data. One of the areas selected included the type of questions encountered in walk-in statistical consulting situations. For example, one question gave the sample correlation coefficient, r, between two measurements of geologic core samples as 0.70 and asked for the sample size at which this value of r would be significant at the 5% level for a one-tailed test.

In a third example, an army export project, the object was to extract from experts those factors that impacted on decisions to transfer militarily critical army technologies, services, or data to foreign countries or persons. The purpose was to represent these factors in a structured manner that would promote better, more defensible decisions. The question areas were the perspectives of the various army offices (intelligence, political, military, and technical) in viewing potential technology transfers. For example, one question area was the factors that the intelligence office representatives would consider in making their decisions. The first question in the intelligence area was to evaluate whether the requesting country was vulnerable to having the particular technology compromised. This question included a definition of the concept of technology compromise and a description of the technology being considered.

Sources of Variation

The major sources of variation in projects' purposes, question areas, and questions are noted below to allow the reader to compare his or her situation to other situations. The reader is asked to do the following:

- **Determine whether the objective is to gather the experts' answers or to gather their problem-solving processes.** While both the expert's answers and problem-solving processes are frequently gathered for an application, one is considered to be of the first priority. For example, for most applications, especially in risk or reliability analyses, obtaining the expert's answer is the primary aim. There may be some attempt to document the expert's reasoning, but this is done to support the answer and is not usually the main focus. Obtaining problem-solving data is more common to artificial intelligence projects or to research into human cognition. While the experts' answers are usually gathered in these studies, they are considered only part of the problem-solving data.
- **Compare the complexity of the areas and the questions.** For instance, the questions of the reactor risk application mentioned above were more complicated than those classical statistical problems asked in the study of correlation because of the former's subject matter--reactor phenomenology. The reactor risk questions were extremely complex because they involved a variety of physical systems, components, and things happening to these

systems. As a general rule, complex question areas require more honing to form questions than simpler areas do.

- **Assess the magnitude of the data required.** For example, in the reactor risk application, a tremendous amount of data was gathered because the complex questions had to be broken into numerous scenarios, and the questions, the experts' reasoning, and their answers had to be documented.
- **Further assess the level of detail that will be needed in each chunk of data.** For the most part, applications where the goal is to gather problem-solving data (such as for building an expert system) require more detailed information than their answer-gathering counterparts. In general, gathering more detailed data coincides with fewer experts and longer elicitation sessions. In addition, the level of detail varies according to how difficult the problems are to solve. For example, in the second example above of the study of interexpert correlation, the solving of the statistical questions was straightforward compared to the solving of the questions in the follow-on study where the experts were asked to evaluate how well computer-modeled results matched experimentally obtained results. The expert's *interpretation* was involved to a greater degree in answering the questions on the computer-modeled results than in answering the statistical questions of the first study. The follow-on study yielded more detailed data on the expert's problem-solving processes than the first study. The level of detail in questioning affects later analyses, especially involving the detection of correlation and bias (chapter 14), the aggregation of experts' answers (chapter 16), and the drawing of conclusions (chapter 18).
- **Evaluate the scope of the application as illustrated by the number of experts that are likely to be used, the personnel available to elicit, and the amount of time needed to produce the product.** For instance, the reactor risk study, NUREG-1150, had the largest scope of any of the expert judgment applications that we have encountered; it had the greatest number of question areas (over 15), of questions per area (3 or more), of experts (50), and of project personnel from different organizations (40 or more persons).

Executing the Steps with the Assistance of Clients, Project Personnel, and Experts

The generic roles of persons who are likely to be working on expert judgment projects and who could assist in developing the questions follow.

Clients are the persons requesting the gathering of expert judgment. The client may also be the person funding the project or a decision maker who may eventually use the project results.

Project personnel include the in-house managers, data gatherers, and analysts. The data gatherers may be **interviewers** or **knowledge engineers**. Interviewers are sometimes referred to as elicitors.

Advisory experts are outside consultants or in-house personnel who are considered expert in the subject matter. They can assist the project personnel in the design of the elicitation methods. Generally, they do not serve as the experts for the final elicitation but assist in developing the elicitation methods by helping to select question areas, create the questions, and test each step for possible difficulties.

Experts, sometimes called the *external experts* to distinguish them from the advisory experts, are the ones who will later answer the questions. The external experts can fulfill the function of the advisory experts by selecting and refining those questions that they will later answer. Whether or when the external experts will help in developing the questions is a critical decision. For more information on making this decision, see the later section *Determining in Which Steps the Advisory or External Experts Will Assist.*

Recognize that these categories of persons are not mutually exclusive. For example, if the project were self-instigated, the client could be one of the project personnel. Then too, the advisory expert could be from in house and thus, a member of the project personnel.

Step 1: Defining project purpose

The client determines what, in general, needs to be investigated and should know what information will be needed from the experts and what resources can be provided. For example, the client may be able to state what is expected as a final project, when it is due, and what level of funding will be available. In addition, the client may be able to provide direction on the scope of the project (number of experts, question areas, time frame), the data to be gathered (primarily answers or problem-solving), and the level of detail needed in the data.

Step 2: Selecting question areas

Project personnel generally work with the client, advisory experts, and external experts in selecting the question areas. Occasionally, the project personnel are experts in the question areas and could forego receiving the input of the client and experts. However, even if this is the case, we recommend the involvement of the client and the experts for two reasons. First, with more persons working on selecting the areas, there is less chance of overlooking an area or having the areas reflect one narrow viewpoint. Second, people who were involved in the selection process are more likely to be supportive of the final selection than those who were not. Indeed, for this reason the external experts should participate in question selection, whenever this is possible. Sometimes this option is not possible because the experts cannot be selected until the areas of expertise as delineated by the question areas are decided.

Sometimes the client is not able to help in this step because he has not thought as far as the question area or has difficulty articulating his ideas for question areas. If this is the case, the project personnel can interview the client to determine what the areas should be (e.g., what is the purpose of this project, what are its constraints in terms of time and funding, and so on).

In our experience, the project personnel have usually set the criteria for the selection of question areas either by themselves or in combination with the client. Some examples of criteria were given in the earlier section *Steps Involved in Selecting the Questions*. Regardless of whether time is specified as a criteria, we have noticed that time limitations are often responsible for paring down the list of areas. The advisory and external experts can also assist in determining the question areas. The following are critical points best addressed by the experts.

- The potential question areas possible, given the purpose of the project.
- The approximate number of those who are experts in each area.
- Ideas on what would motivate the external experts to participate in the project.
- How much the question areas would need to be broken into their parts to become questions later answerable by the experts.
- Whether the question area has been sufficiently defined to proceed to the next phase--creation of its component questions.

While the advisory experts often address any of the above points, the external experts usually only assist on the last two. This difference exists because the external experts typically enter the project later than the advisory experts. When the question areas involve different and specialized expertises, the question areas must be selected before the project personnel can identify any experts for consideration.

Step 3: Identifying the questions

The project personnel and experts, advisory or external, often work together to develop questions from the question areas. If the client is qualified in this subject matter, he or she may also be of assistance.

After the experts have assisted in creating some questions, they can be asked the following:

- Whether the questions are answerable (e.g., is there any datum, reference, or experience relevant to the question)
- Whether there is likely to be much diversity of opinion among external experts in answering the questions
- Whether any of the expert judgment data would be considered proprietary
- The number of questions that an expert could answer in a particular period of time

Determining in Which Steps the Advisory or External Experts Will Assist

Experts are needed to select the question areas, identify the questions (steps 2 and 3 in chapter 4) and refine the the questions (chapter 5). Either advisory or external experts can work with the project personnel to perform these tasks. However, *when* the external experts become involved in the project is a critical consideration because it affects the order

in which chapters 4, 5 and 6 should be applied. We recommend that the three following options be considered and one selected.

The external experts are not involved in question area selection, question identification, or refinement, except minimally just prior to having their judgments elicited. In this option, the advisory experts do the majority of the work, in combination with the project personnel, in selecting the question areas and identifying and refining the questions. Then, the advisory experts are used to pilot test the questions for clarity and ease of use. (Pilot testing is discussed in chapter 9.) Surveys or questionnaires are frequently developed in this manner. Later when the external experts respond to the questions, their interpretations of the meaning of the questions are recorded and constitute their refinement of them.

The advantages of this option are as follows.

- The development of the questions is controlled by the project personnel (e.g., either they make the choices or they direct the advisory experts in making them).
- The external experts do not have to be selected until after the questions are finalized (as described in chapter 5).

There are three disadvantages of this option:

- The external experts will not be as motivated to address the questions as they would have been if they had helped develop them.
- The external experts will have a more difficult time understanding the questions than if they had developed them. In particular, experts often have problems in encoding their responses into the response modes that the project personnel have picked and with which the expert has no familiarity.
- Those who review the project may believe that the experts were *led* because the experts did not develop the questions.

If this option is chosen, read and/or apply chapters 4, 5, and 6 in sequence.

The external experts are presented with the question areas and identify the questions and refine them, or they are presented with the questions and refine them. Frequently, as in the NUREG-1150 reactor risk study, the project personnel select the question areas or questions and the experts modify them (e.g., add, delete, or reword them). If more than one expert will be addressing a question, they will need to come to agree on its revision. Otherwise, their later responses cannot be compared because they will be answers to essentially different questions. For this reason, the project personnel usually monitor the experts' work in arriving at the final wording of the questions. This option of partial involvement by the external experts is favored when the project personnel want some control over the questions. One situation where the project personnel would wish to have some control over the final questions is when the client has specified which question areas he wants covered.

The advantage of this approach follows.

- The project personnel can specify the criteria that they wish met in the development of the questions and let the experts do the work.

The disadvantage of this approach follows:

- Those who review the project may believe that the experts were *led* because the experts did not develop the questions.

If this option is chosen, skip to chapter 6 on selecting the external experts.

The external experts are involved in the question area selection and in the identification and refinement of the questions. This approach works best when the experts can be gathered together over time to develop the questions. Frequent meetings require that the experts be located in the same geographical area or organization. This option is also applied when the experts are not located in the same place but can meet for a concentrated period of at least a week. For example, on the army export project, experts from different army offices met for one week to select the question areas, identify and refine the questions, and answer them for a test case. Another possibility is for the experts to meet to review and modify those question areas earlier proposed by the project personnel. Occasionally, the experts do not physically meet until later in the question refinement but work on the question areas through mail correspondence.

Early expert involvement is used if the external expert's cooperation or views of the project are critical to the success of the project. For instance, if there were very few experts in the field, the participation of each would become more important than if there was an unlimited pool of experts. Additionally, if the client or project funder would interpret any expert's reluctance to participate as a sign that the project was a failure, expert participation becomes critical.

The following are two advantages of early expert involvement:

- Early involvement has a positive effect on experts' willingness to participate in the study and later provide their judgment.
- Experts will be more supportive of the product of the project because they will view it as the fruit of their labors.

The following are the disadvantages of this approach:

- The project personnel will not have complete or direct control of the development of the questions.
- Working with a group of experts to develop the questions requires special skills (tact and ability to keep things rolling), and even then it sometimes resembles a three-ring circus.
- Having the experts meet and proceed to develop the questions requires more advance planning than the other options. (See *How to Set Up for a Delphi Situation* or *How to Set Up for an Interactive Group Situation* in chapter 10).

If the external experts will do the bulk of the question development, read chapter 6, *Selecting and Motivating the Experts*, before continuing with this chapter and chapter 5.

Checklist for Selected Questions

After the questions have been formulated, they need to be evaluated for their suitability. The following is a checklist for that purpose.

1. **Will the questions provide the data necessary to meeting the goals of the project?** Sometimes the creation of the questions takes on a life of its own and moves in a direction that has little bearing on the goals of the project. Frequently, it is helpful to outline on paper the project goals; that is, the type of expert judgment, answers, problem solving, or ancillary data that is needed for the project; and the questions that are being considered.
2. **Will the questions be within the scope of the project?** In particular, time, funding, and logistics are critical considerations.

   Time. A major consideration is whether there are the appropriate number of questions for the time allotted. Rough estimates of how much time a question will take can be obtained by considering the type of data (*answer only* or *answer plus*) it is to gather and the level of detail needed. The chart below illustrates the amount of time that it takes to elicit an expert's response to different types of questions. Note that the level of detail usually corresponds to whether answers only or answers plus will be gathered. When answers plus problem-solving processes are gathered, the amount of data being gathered multiplies and each datum becomes more complex.

Rough Time Estimates for Eliciting Expert Judgment in Different Situations

Elicitation Situation	Type of Data	Level of Detail	Approximate Time
Experts in a group	Answer only	Very low	A few minutes
	Answer plus a few sentences on their rationale		5-10 minutes
Expert alone	Answer only	Medium	30 minutes
	Answer plus problem-solving data	High	1-2 hours

Funding. The amount of personnel costs for advisory and external experts is a second consideration For example, on one large project, the costs per external expert averaged $10,000. The experts traveled to two meetings and provided in-depth answers and problem-solving data on the questions to which they were assigned.

Logistics. Will the logistics of eliciting the expert's judgments be reasonable? For instance, in a problem-solving application where a few experts are needed for long periods of time, are there experts located nearby? According to Waterman (1986:192), the availability of experts is crucial in this

situation. If the experts are not located nearby, could either the project personnel or the external experts be relocated to be together for the necessary period of time?

3. **Can the expert judgment data be obtained without extreme effort?** For example, will the data be classified or proprietary? We have found proprietary data to be more difficult to handle than classified. The procedures for handling classified data are clear, and although time consuming, do not deter the project. With proprietary data, however, experts may be unwilling to participate for fear of providing data to their competition. An example of an application containing proprietary data would be the failure rates of different manufacturer's pipes in nuclear reactors. If the data is likely to be proprietary, does the the client or the project manager have enough influence to overcome the objections to participation? Can the project personnel guarantee that access to the data will be protected and limited?

Common Difficulties--Their Signs and Solutions

Difficulty: ***Client can not provide clear information on the project's goal, the information that is to be gathered, or the question areas.*** The client may be uncertain about the project, as many are when it is in a conceptual stage, or unable to communicate a view of the project. This same sort of difficulty occurs in most consulting applications--the analyst must *extract* what it is that the client wants and needs.

Solution: First, determine at which point the client's conception of the project becomes unclear. Then, elicit from the client the information that is necessary for the project to proceed. For example, if the client has only vague ideas on what the project should be, this basic information needs to be elicited, clarified, and recorded. More frequently, the client will be able to specify the project's goals but not *how* those goals are to be accomplished; the client's reasoning being that the implementation of the goals is your job, that is why you have been hired. Again, the solution is to interview the client to obtain as much information as possible on how to proceed within the project scope as the client has viewed it. Two of the interviewing techniques outlined in chapter 7, the verbal probe and the ethnographic method, may be helpful in questioning the client. For instance, the client could be informally questioned using these two techniques and then interviewed in more depth using the ethnographic technique (chapter 7). The ethnographic technique allows the client's own words to be used in the questioning; in this way the exact meaning of his responses can be extracted.

Difficulty: ***The question developed from the question area is still too broad.*** One sign of a too-broad question is when the advisory expert or external expert has to break the question into smaller parts or request additional information before being able to answer. For example, the following question was asked about a particular nuclear plant in a reactor risk analysis study (Amos et al. 1987): "What is the frequency of ignition of the hydrogen given that there has been a station blackout

causing hydrogen accumulation?" One expert felt that the frequency of ignition would depend on whether there was high or low pressure in the vessel and therefore broke the question into those two possibilities. Additional information that might have been necessary to answer the question was an illucidation of the use of the term *ignition.*

Solution: Ask an expert to solve the question and use his decomposition of it as a starting point in narrowing the question. For instance, in the above example the question could have been decomposed into high or low pressure, and ignition could have been explained as stopping short of detonation. As another check, the expert can be asked if the questions, in their present form, are basically answerable. If the expert replies that the questions are still too vague, ask the expert to think aloud about how to make the questions more manageable. (See chapters 7 and 10 on how to use this method, the verbal protocol, of interviewing.)

Difficulty: Too many questions have been selected for the amount of time available. Selecting too many questions to be answered in the available time is a very common difficulty. An early sign of this difficulty is people's unwillingness to discuss the number of questions and the amount of available time. Perhaps because most persons are not aware of how time consuming it is to do an in-depth elicitation, they tend to estimate time from how long it would take someone to give an off-the-top-of-the-head answer. Because of wishful thinking, even persons experienced in elicitation tend to underestimate the amount of time a particular number of questions will take.

Solution: The advisory expert should be asked to provide rough estimates of how long an expert would need to respond, given the elicitation procedure being considered and the level of detail needed in the data. This amount of time can be examined in light of the numbers of experts, the questions planned, and the total amount of time available. If this rough estimate indicates that there are too many questions, there are several ways for reducing their number: fewer questions can be selected from each question area, the number of question areas can be cut, fewer experts can be sampled, the elicitation method can be made simpler and faster, or less detailed data can be gathered. In addition, it may be possible to extend the project's deadlines and to avoid any of the above measures.

The time-estimate chart shown in the previous section can also be used in approximating the amount of time that each question will take. As a general rule, elicitations of individual experts last longer than those done in groups because individual interviews are used when detailed data are needed. In addition, elicitations that gather problem-solving data tend to be more lengthy than those which just gather answers. The greater the amount or detail of problem-solving data that is gathered, the more time the elicitation will take. For example, it takes less time to obtain the experts' general rationale--a few sentences documenting their reasoning--than it does to obtain their definitions, assumptions, heuristics, references, and calculations.

5

Refining the Questions

In this chapter we describe how to refine the questions selected in the previous chapter. The aim in refining the questions is to take human cognitive limitations into account and create questions whose information can be more easily assimilated and processed by the expert. We believe that trying to minimize the occurrence of factors negatively affecting cognition will lead to better quality expert judgment. Chapter 5 includes suggestions for presenting the information necessary to understanding the question (background, definitions, and assumptions), for ordering this information, and for breaking the question into more easily understood parts. Finally, the reader is asked to consider the wording of the question in terms of clarity and bias. In addition, this chapter describes when the experts should be involved in refining the questions, and when the stage of selecting and motivating the experts (chapter 6) needs to precede this chapter.

Reasons For Structuring the Questions

The questions are refined through structuring. Structuring questions, asking them in an organized and controlled manner, is done with the aim of obtaining the best quality data. Some means of structuring the question include presenting its information in an orderly way, breaking it into more easily answered parts, representing it in a pictorial or mathematical way, phrasing it in a careful, nonleading manner, and defining the key words.

Structuring the questions provides the following benefits:

- It focuses the expert's attention on what he is to provide.
- It lessens the cognitive burden of solving the question by presenting it in a more assimilable and processed form.
- It delimits the question so that the experts are not interpreting it differently and thus answering separate questions.
- It makes the question more acceptable to the experts because it has been refined to encompass their views and use their terminology.

Which of the structuring techniques will be used and to which degree depends on the questions, particularly their complexity. By complexity is meant the amount of information required to solve the question and whether there is any means of verifying the correctness of the answer. With simpler questions, there is less information involved and

often some means of determining the *right* answer. An example of a simple question would be: "At what value would a chi-square statistic of 5.74, with 3 degrees of freedom, be significant?" (Meyer and Booker 1987b:40). A complex question would be: "What is the fraction of inventory of radionuclide group present in melt participating in pressure-driven melt expulsion that is released to containment as a result of melt expulsion?" (NUREG-1150 source-term elicitations, 4/13/88). The table below summarizes the degree to which the structuring techniques are likely to be needed, given the question's complexity.

Need for Structuring Techniques

Question Complexity	Question Information	Breaking It Into Parts	Representation	Question Phrasing
Simpler	Little (e.g., definitions/ assumptions)	Little (e.g., textual description)	Little	High
More complex	High	High	High	High

With simpler questions, less information needs to be provided to the experts. Typically, only those data, definitions, and/or assumptions that the experts are supposed to consider need to be provided. For example, on the simple question mentioned above, the observed, the expected values, and the chi-square statistic were provided as question information. In addition, the options for response were provided in a form that was standard for the experts. For example, the value of 1% was defined as meaning at or greater than 0.01; 5% meaning greater than 0.01 but less than or equal to 0.05; and so on. The magnitude of this information can be judged by the space it occupied--less than half a page. With the complex example, the experts not only needed pages of background information, but information from each other concerning the question. They shared their information during briefings before their elicitation sessions.

In addition, simpler questions are not as likely as complex questions to need breaking into parts because they are already at an answerable level. If a simple question is partitioned, it is likely to include only a few parts. Because of this simplicity, such questions are not likely to require a pictorial or mathematical representation. By contrast, complex questions are frequently diagrammed with trees and charts representing their many parts and the interrelationships between those parts.

Regardless of the question's complexity, careful phrasing is always critical to the experts' understanding of the question.

Techniques for Structuring the Questions

Presentation of the Question Information

It is extremely likely that information beyond the current statement of the question will need to be given, no matter how simple the question is. The more elaborate the question, the more information needed, and the more time consuming the planning of this presentation. Planning the presentation can be divided into two aspects: (1) determining the types of information needed, such as question background, assumptions, and definitions; (2) determining the optimal order for their presentation, and (3) roles of project personnel and experts.

Types of question information needed

Background. One of the first steps is determining the types of information that will be needed by the experts. Frequently, the experts request information on the background to the question (i.e., what events have occurred, what events are supposed to occur, and what the current status is of the thing being evaluated). For instance, in a project for determining how tank platoon leaders plan their routes, the experts needed background information on their mission. This background information included maps of the terrain, the point at which the experts were to start, the general area in which they could travel, their objective, and the probable location of enemy tanks.

Background can also be given on a physical process, if this is the focus of the question. For example, when the nuclear engineers were questioned about the performance of their code in predicting experimental results, they were given background information on the experiment. In particular, during the elicitation sessions, they were provided with text on the experiment's procedures, its equipment (lists and schematics), and boundary and initial conditions (temperatures and pressures). For the more complex questions on severe accidents occurring in nuclear reactors (NUREG-1150), the experts were sent thick packages of the latest references, prior to the elicitations.

Background information can also include representations of the question broken into parts. For example, in complex risk analyses, the possible combinations of events are depicted with tree diagrams, and the experts give estimates on the likelihood of the occurrence of the branches. These representations may have been developed previously by the project personnel and/or by the experts themselves. For example, in a simple decision analysis project on the relative safety of new automotive fuels, the experts developed possible accident scenarios (Krupka et al. 1983). The scenarios provided the framework in which the experts judged the likelihoods of particular accidents occurring for vehicles run on the different fuels.

Assumptions. Assumptions are another type of information provided to the expert. It is necessary to present the experts with the assumptions that they are to make in answering the question. If these assumptions are not specified, the experts are likely to make varying ones of their own, sometimes completely altering the question's meaning. (Thus, the original question may pass unaddressed and the experts may have answered totally different questions). The tendency for experts to make assumptions that conflict with the question's original meaning is more pronounced on complex questions. On

complex questions, it is difficult to provide the experts with all the details that they believe they need. They make their own assumptions as a means of filling in the gaps. For instance, on the question of whether a particular army technology should be exported, the experts wanted to know how many of the technologies were being requested. They said that the number being requested would affect their answers. To allow them to proceed in answering the question, they were asked to make the same assumption--that the request was for four to six of the technologies (Meyer and Johnson 1985).

Assumptions are also used when the information requested can not be provided, such as when concerning a rare physical process, but some common base must be established to allow the experts to continue.

Definitions. Definitions of terms are another type of information commonly used to refine a wide range of questions. For example, in the export study mentioned above (Meyer and Johnson 1985:7), it was necessary that the experts define *technology transfer* in the same way, so they were asked to agree upon and use a definition (e.g., *technology transfer is the means by which technologies, goods, services, data, and munitions that are deemed militarily critical by the Department of Army are transferred to a foreign country, international organization, firm or individual*).

Ordering of information

The next step is to determine the order in which the experts will need this information. One means of doing this is to consider the logical flow of the information. At each point in the planned elicitation, at each question, what information does the expert need to respond? A means for checking the flow is to have the advisory experts work through the question. If they request information that was not provided (e.g., not considered for inclusion or not available for inclusion), insert it at that point. For example, in a study of how tank officers planned their routes, the subjects frequently requested more information on the density of the forests than was offered on the maps (Meyer 1987). They felt that they needed this information to assess whether the trees would offer sufficient cover.

Finally, how humans assimilate and recall information needs to be considered in sequencing the information. People are thought to better take in new information when it fits within the context of their prior knowledge, their mental models (Waern 1987:276). Even though the subjects are experts in their field, the question information may represent a new portion or a different configuration of the data then they possess. Since the experts need to answer the question, albeit in their own ways, they must have a through understanding of it. Thus, the experts should be introduced to the various bits of information, such as the definitions to be used in common, in a manner that facilitates their mental filing and accessing of it.

In addition, in the complex and dynamic environment of solving problems, people are prone to forgetting information. For this reason, Payne (1951) has recommended that any information critical to the interpretation of the question, such as important definitions and assumptions, be included as part of the question. For example, in a study of severe accident sequences in nuclear reactors (U.S. NRC 1989), a few questions began with the assumption--for example, *given that x has occurred, what is the probability that y will?*. If these definitions and assumptions cannot be made a part of the question because they make it confusing or too lengthy, they can be given immediately before the question. For

example, in a project forecasting the weapon needs of the year 2000 (Meyer et al. 1982:7), the definition of the weapon was read immediately before the question.

Often the above considerations on how people best assimilate information leads to a hierarchical presentation of the information, such as from general to the specific or from specific to inclusive. For example, a general-to-specific ordering would provide the experts first with the general context of the question, such as its topic and the scenario, and then provide the definitions or assumptions which narrow the question. An ordered presentation of the information, whether it be from general to specific or vice versa, is done to assist the expert in assimilating the necessary information.

Roles of project personnel and experts

As mentioned in chapter 4, these are the generic roles of persons who could plan the presentation of the question information. Their roles could overlap, particularly if the advisory experts were also project personnel. In addition, other persons such as the client could have a role in presenting the question information if they had assisted earlier in selecting the questions. As mentioned in chapter 4, project personnel can include the **data gatherers** (**interviewers** or **knowledge engineers**), managers, and **analysts**.

The project personnel and advisory experts can work together on determining what information needs to be presented and in what order. One approach would be to have the project personnel draft the information and the advisory experts review and pilot test it. (Pilot testing is described in chapter 9.) The project personnel are qualified for this role because they know the project's aims and how they intend to do the elicitation. The advisory expert's know the field and can anticipate the experts' information needs and their response to the proposed question. Another approach would be to have the external experts state what background would need to be provided, agree on the definitions and assumptions to be used, and decide on the order of the presentation of this question information.

Decomposition of the Question

Another typical means of structuring the question is through **decomposition**, also referred to as **disaggregation** (Meyer 1986:88). Frequently, questions are broken into parts to ease the burden of information processing and to promote accuracy (Armstrong et al. 1975; Hayes-Roth 1980). For example, Armstrong et al. (1975) asked straight almanac questions of half of their sample. Of the other half, they asked the same almanac questions but broken into logical parts. For instance, the question "How many families were living in the U.S. in 1970?" was asked as "What was the population of the U.S. in 1970?" and "How many people were there in the average family then?" The persons answering the disaggregated questions gave significantly more accurate judgments. The information on how to decompose the question is often given to the experts as background, as mentioned in the previous section.

Complex questions are more likely to require decomposition. On simple questions, such as the almanac question mentioned above, decomposition is either not needed or done only slightly. A classic example of where question decomposition is used is in risk analysis. For instance, on the study of severe accident sequences (U.S. NRC 1989),

questions were disaggregated by the experts into a case structure. The cases were created by considering those factors that would have critical effects on the reactor phenomena being considered. For example, on the question of radionuclide release associated with pressure-driven melt expulsion in a pressurized water reactor, some factors that were thought to have bearing were the reactor coolant system (RCS) pressure, and whether the cavity was full, half full, or dry. Cases were formed of combinations of different pressures and states of the cavity (e.g., RCS pressures of 2500, 2000, and 500-1000 and a full, half full, and dry cavity or an RCS pressure of 15-200 with a full cavity).

Considerations in question decomposition

Decomposition is not a simple procedure but one that involves diverse and overlapping considerations. In addition to the question's complexity, there are several considerations that impact on how the question is decomposed. One is the purpose in performing a decomposition and the aims of the project. For instance, if the question is highly complex and the intent is to aid the expert in his information processing, a more detailed decomposition is needed. In addition, if the expert's thinking is to be documented so that someone else can track it, the decomposition will have to be taken to finer levels.

Another consideration is the amount of detail that will be needed in the data. If more detail is needed, the decomposition will need to be correspondingly finer. For example, if data on the expert's problem solving is required, this would imply more detail. As a general rule, the data gatherers must obtain data that is one level more detailed than that which is needed for analysis.

Another factor in planning the question decomposition is the external experts' involvement. How much of the decomposition will they do and at what point in the question's development? It is critical that the decomposition be acceptable to the experts. As a general rule, we recommend that the external experts be involved in the question decomposition as soon as possible. For example, the experts could begin refining the decomposition after it has been developed in-house and approved by the advisory experts. Or, the experts could decompose the questions by themselves.

A further consideration in decomposition is the relationship between the parts of the question. For instance, the relationship could be causal, temporal, or logical, as it was in the almanac example.

In addition to the above considerations, there are several concerns guiding development of a question decomposition. First, there can be problems when the experts' answers are combined for analysis if they have used different decompositions. The problem is that the experts will have answered different questions--to combine these is akin to mixing apples and oranges. Another caveat is that the decomposition be logical, that it properly model the relationship between the parts. A third problem is having gaps or redundancies in the decomposition that lead to under- and overrepresentation of the parts when these are mathematically modeled.

Roles of the project personnel and experts

As mentioned above, the disaggregation can involve the project personnel, advisory experts, and external experts. One method is to have the project personnel propose the disaggregation, the advisory experts review and refine it, and the external experts use it as

is to respond to the question. A second way is to follow the above procedure but to allow the external experts to modify the disaggregation, as was done on the reactor risk project, NUREG-1150 (U.S. NRC 1989). Still a third way is to have the experts propose their own disaggregation. Because there is always the danger that the experts will reject a disaggregation that they have not, in large part, developed, we recommend the second or third approach.

Representation of the Question

Questions are likely to need representation if they have been disaggregated in any detail. Representation is the pictorial or mathematical depiction of the question showing the factors that have bearing on the question and their relationship to one another (relative likelihood, consequence, and importance). For example, in probabilistic risk assessments of nuclear reactors, accident sequences are diagrammed and their outcomes determined. The accident sequences resemble a decision analytic model (Barclay, Brown, Kelly, Peterson, Phillips, and Selvidge 1977) in that the sequences consist of branches that display the possible outcomes that would be arrived at if particular events occurred. The accident sequences can be represented in two ways:

> . . . as event trees, which depict initiating events and combinations of system successes and failures, and fault trees, which depict ways in which the system failures represented in the event tree can occur." (U.S. NRC 1983: 2-3)

In the example below, figure 1, the simple event tree shows events leading to a safe termination of their sequence or to a specific plant-damage state. Representations can be used to accomplish the following:

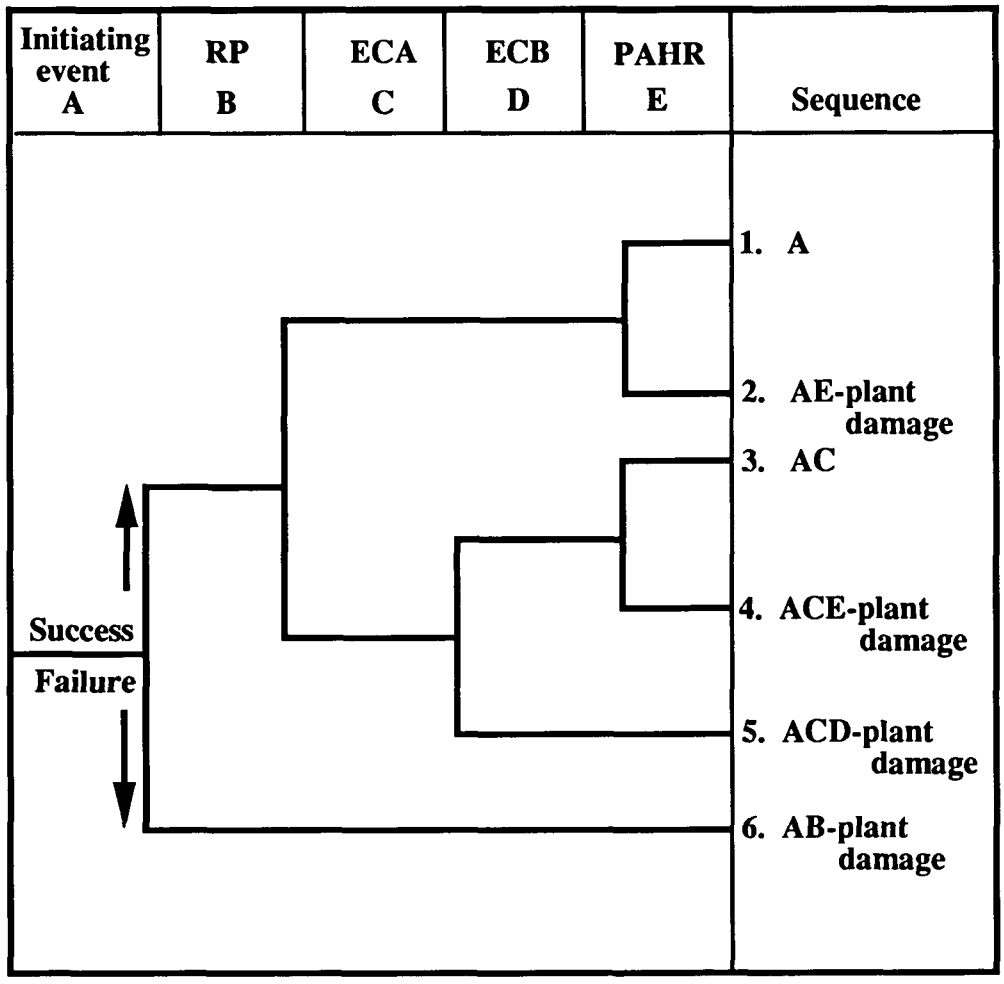
- Guide the external experts in making judgments (e.g., by providing the disaggregation that has been agreed upon)
- Document how the experts reached an answer
- Provide future guidance on how the question or similar ones are to be solved
- Provide guidance on how the expert judgment is to be processed and analyzed

Note that *representation* here is distinguished from the term *knowledge representation* used in artificial intelligence. Representation only includes the ways in which the question can be modeled; knowledge representation includes the domain of knowledge, the language for programming it, and the means for making automatic inferences. For further references on knowledge representation in artificial intelligence, see Brachman and Levesque (1985), Sowa (1984), or Skuce and Sowa (1988).

Considerations in representation

Arriving at a representation involves the same sort of considerations and concerns as decomposition (see *Decomposition of the Question* above). The main considerations in selecting a representation scheme are that it be compatible with the question decomposition, the analysis plans, and the expert's means of solving the question. If one of the existing representation schemes cannot be applied, a new idiosyncratic one can be created. A few representation schemes are described below.

The accident sequence representation, used in performing probabilistic risk assessments (PRA), is based on causal and sequential relationships. That is, one event can cause another and lead to particular outcomes. The PRA accident sequence is widely used in the nuclear risk/reliability community but has general applications to modeling physical phenomena. For example, a related representation scheme is used to diagram automobile accidents (Krupka, Peaslee, and Laquer 1985). This scheme depicts the events that could cause accidents among automobiles run on new alternative fuels. For details on how to do an accident sequence representation, see the PRA Procedures Guide *NUREG/CR-2300 (U.S. NRC 1983).*



Source: NUREG CR-2300 (U.S. NRC 1983)

Figure 1. An example of a simple event tree for Probabilistic Risk Assessment (PRA).

Another type of representation scheme is one in which the parts are temporally related. These schemes are often used for managing technical programs where for the accomplishment of the end product, each milestone must be achieved on schedule. Such representations show which activities or milestones need to be realized in parallel and

which in sequence. PERT is an early example of this type of representation. PERT was developed by governmental agencies as an aid to planning and evaluating the costs and scheduling of objective-oriented work (PERT Coordinating Group 1963).

In another type of representation scheme, the parts are alternatives which are evaluated in terms of particular attributes they are judged to possess. A common scheme of this type is Saaty's Analytic Hierarchy Process (AHP). AHP is frequently used by decision makers or experts to pick, from alternative actions or products, the one that will best address some agreed-upon criteria (Saaty 1980, 1982). For example, in the export control project, the AHP representation was used to make decisions on whether a specific Army technology should be either exported, not exported, or exported with particular conditions (Meyer and Johnson 1985).

Idiosyncratic representation schemes are those that occur when a new representation is created especially for the project. Idiosyncratic schemes differ, so the relationship between their parts cannot be characterized as being one thing, such as causal. Idiosyncratic schemes are used when one of the existing schemes cannot be applied. Thus, they may be very suited to one application but not easily generalizable to others. For example, an idiosyncratic scheme was created for a project that was to forecast the weapon needs of the United States for the year 2000 (Meyer et al. 1982). This scheme was developed because the experts needed some framework for thinking in a structured manner about what potential threats there might be to the defenses of the United States in the future--that is, what offensive weapons other countries might develop and how the United States might need to respond to these in their own weapons development.

Roles of project personnel and experts

The roles of the project personnel and experts are the same for representation as they were for disaggregation. As with disaggregation, the external experts should take a major role in creating a representation. This is particularly true if the purpose of the project is to model the experts' problem-solving processes or to create one that will serve as a guide to future decision making.

Question Phrasing

Another element of structuring is question phrasing or wording. **Question phrasing** refers to the wording of the question and of the mode in which the expert is to respond (response mode). Careful question phrasing maximizes the chances that the expert will understand the question and not be unduly influenced or biased by it . A biased wording can cause the expert to describe his thoughts or answers differently than they were and thus can be considered motivational bias as described in chapters 3 and 8. Specifically, we consider biased phrasing to be a type of social pressure because the expert's thinking is unconsciously affected by the perspective he picks up from the wording.

The biasing effect of phrasing has been shown most dramatically by Payne (1951) through his use of the split ballot technique in survey questions. The split ballot technique entails giving half of the sample one wording of the question or response option, and the other, another. For example, one wording of the question might be: *Do you believe that X event will occur by Y time?* The other wording might be: *Do you believe that X event will occur by Y time, or not?* This second option is more balanced because it mentions both

possibilities. For this reason, it would be likely to receive a higher percentage of "no" responses. Often the difference measured by the split ballot technique is 4-15% even when the rewording has been very slight (Meyer 1986:88).

In another example (Meyer and Booker 1987b), the experts were asked to identify where two curves (one generated by a computer code, the other by an experiment) diverged. An early wording of the question asked the experts to mark the places where they felt the curves diverged. This wording was leading the experts (e.g., to believe that there must be divergences) and was therefore changed to "mark the places, if any, where the curves diverge." As Payne (1951), the grandfather of surveys noted, all question phrasings are biasing; the best that one can do is aim for equal, but opposite, biases.

Another problem with unclear wording is that the experts are likely to interpret the question differently and give answers to essentially distinct questions. For example, in the study of interexpert correlation mentioned above (Meyer and Booker 1987b), the experts were found to have separate interpretations of the term *diverge*. To some diverge meant where the lines were not exactly the same; to others, it meant where the distance between the lines increased with time; and to still others, it meant where the differences exceeded the error bars of 10-20%.

Another factor that has bearing on the question's clarity is its length. Payne (1951) has found that people's comprehension of written sentences tends to drop off after 25 words. For this reason, we recommend that sentences be kept as short as possible.

Considerations in question phrasing

In phrasing the questions, consider clarity and bias. Clarity can be improved by having the project personnel and experts review and offer feedback on what the phrasing meant to them. (This procedure, pilot testing, will be described in detail in chapter 9). There is no easy procedure for combating bias--the best strategy is to be sensitive to this issue and to carefully scrutinize the phrasings.

One problem in question phrasing is creating one that will be commonly understood and acceptable by multiple experts. If a phrasing (or representation or disaggregation) has been reviewed and tested by only a few experts, it may be slanted (e.g., reflect only their experiences, views, and use of terms). If only one advisory expert is used, the chances of slanted phrasing are even greater.

We encountered this problem of slanted phrasing while drafting a linear scale for experts to use in estimating the magnitude of a computer code's divergence from the experimental results (Meyer and Booker 1987b). There was a small population of experts, and we did not wish to further decrease that population by designating more than one person to be an advisory expert. Thus, only one well-qualified expert served as advisory expert in helping develop the scale for the experts' responses. However, when we requested feedback on the scale, we learned that several of the experts did not accept a major premise of the scale. The advisory expert had equated magnitude of the divergences with the presence of code deficiencies (e.g., insufficient agreement was associated with serious code deficiencies that required immediate fixing). (See chapter 7, example 7.2 for an illustration of the linear scale). Several of the external experts did not make this direct association. Instead, they considered the possibility that the experiment had not been conducted as reported. They reasoned that poor agreement between the code's curve and the experiment's curve could be attributed to the code trying to model an experiment that

was conducted differently than reported. Thus they did not necessarily equate poor agreement with code deficiencies. As a result of the scale's wording, adjustments were being made to the scale much later than was desirable.

Roles of project personnel and experts

The project personnel can propose the first phrasing for the advisory experts' review. This first draft and later ones should be examined by the project personnel for bias. This tasking is proposed because the project personnel will be aware of the potential for bias. The advisory experts can review the phrasing to provide information on the experts' reaction to it. At the very least, the external experts need to have the opportunity to modify the question phrasing before their elicitation sessions. Otherwise, during the elicitations the external experts may state that they do not agree with the question and, therefore, can not answer it. It is generally best to allow the external experts as major and early a role in the development of the questions as possible.

When the Refinement of the Questions Should Be Preceded by the Selection of the Experts

There are three conditions when selecting and motivating the external experts should precede refining the questions:

- If the purpose of the project is to capture the expert's problem solving or to serve as a guide to future decision making.
- If there is any indication that the experts may not accept the questions.
- If outside reviewers are likely to be concerned about bias in the question selection or phrasing.

If any of these conditions exist, it is critical to have the external experts do the primary work on refining the questions and to begin this as soon as possible. Not adequately involving the external experts in question refinement occurs frequently and leads to serious problems. For example, in a reactor study, tight time schedules led to an attempt to save time by having a panel select the questions and the response scale for the experts. The experts were to have the freedom of modifying these, but when they met, did not feel that they had sufficient time to do so. As a result, several experts questioned what their answers would have been otherwise and stated that they did not wish to defend the study (Benjamin et al. 1987:appendix F). Such statements by the experts impair the credibility of a study.

If the external experts are selected first, they will assist in creating the presentation of the information, the disaggregation, the representation, and the question phrasing. Their early involvement may lessen the need for the use of the separate, and usually in-house, advisory experts. The external experts' work on refining the questions can occur in different ways. For instance, the experts may do the above tasks as part of their elicitations or they may do these as a separate step, depending on the project. For example, on the export control project (Meyer and Johnson 1985), the experts were convened once to disaggregate, represent, and phrase their questions and to give their answers. By contrast,

on the reactor risk study NUREG-1150 (U.S. NRC 1989), the experts met several times to (1) receive briefings on the project and view the disaggregations proposed by the project personnel, and (2) to give their answers to their disaggregated questions (i.e., disaggregations that they had developed from those earlier proposed).

Common Difficulties--Their Signs and Solutions

Difficulty: There was not enough input from the external experts in refining the question. Refining includes providing input on what information is needed to answer the question or assisting with the question's disaggregation, representation, or phrasing. Because refining the question is an evolving process, at its early stages the question will not have had sufficient expert input. Then too, given the means of refining the question (particularly who is doing it and at what stage), the signs of this difficulty may vary. The earliest signs may be the responses of the advisory experts who are reviewing the question. They are likely to be confused by the question, say that they do not view the question in this manner or that they can not answer the question in its current form. If advisory experts have not been used to screen the question, this difficulty may be revealed later when the external experts view the questions for the first time. They, then, may have the same reactions as the advisory experts. If this problem exists, the external experts are likely to demand more information on the question, or criticize its disaggregation, representation, or phrasing. Then, if they are to provide their answers as input into this structuring, they may refuse to do so. If, they are only to use the structuring as a first cut, they may insist on extensive modifications of it. In either case, the effects of this problem are serious. The project can lose its credibility or run over schedule when the experts challenge the questions.

Insufficient input from the experts in refining the questions commonly happens because of tight time schedules. When a project has tight time constraints, project personnel may seek to save on time in several ways. Sometimes, they will try to minimize the number of times that the experts meet or the number of times that the structuring of a question is redone. Another means for trying to conserve time is to have the project personnel do most of the structuring. Thus, the duration of this phase stays under the project management's, rather than the expert's, control.

Solution: The simplest way to avoid this serious situation is to involve the external experts as early as possible in the refining of the questions. This is particularly important if the project has any of the following conditions:

- If the purpose of the project is to capture the expert's problem solving or to serve as a guide to future decision making
- If there is any indication that the experts may not accept the questions
- If outside reviewers are likely to be concerned about bias in the question selection or phrasing

If one of the above conditions exists in the project, we recommend selecting and motivating the external experts (chapter 6) before beginning the refinement of the questions (chapter 5).

Difficulty: The question decomposition becomes too complicated or too detailed. In this situation, the disaggregation, and hence its representation, exceeds the level of detail needed for the project's goals and analyses. While it is frequently necessary to gather data one level more detailed than needed, the level of detail discussed here is excessive. The desire to disaggregate *ad infinitum* appears to be a natural tendency. Perhaps the motivation behind it could be explained as *give us enough rope and we will hang ourselves*. Newcomers to disaggregation wish to do a good, thorough job. In addition, they may be trying to take the question to a more easily answerable point, such as where experimental data can be applied. This quest is positive, in moderation, and is the reason that questions are disaggregated. However, this inclination to divide things ever more finely, if unchecked, can be counterproductive.

Solution: Usually a person needs to experience excessive disaggregation once or twice to recognize the tendency in himself and others. If someone else, such as the experts, are becoming too detailed, you can do one of two things: (1) allow them to continue so that they can come to their own realization about the practicalities of excessive disaggregation, or (2) show them through sensitivity analyses when they have more detail than they need. The first approach has the advantage of producing experts who are, after their own experience, supportive of the the more general level of disaggregation proposed by the project personnel. However, this approach has the disadvantage of being more time consuming.

Difficulty: The questions are ill defined or open to differing interpretations. This difficulty is the other side of the one mentioned above. It tends to occur when the structuring of the questions has been rushed due to tight deadlines. Often, our expressions are not as clear as we would like to believe. To us, the question, may be perfectly clear because we are reading between the lines and unaware of some of the definitions and assumptions that we are making. Anyone who has ever drafted survey questions is aware of how many ways, often other than intended, a question can be interpreted. It is to be expected that the questions will be ill defined when they are first being structured. After all, that is why refining the questions has been designated as a separate phase. However, the questions should be as specific and structured as needed before the answers are elicited.

If the advisory experts are pilot testing the question, they may provide the first sign of trouble. If this is the case, they are likely either to request more information or to try to fill in the question's gaps by making assumptions. The same thing can happen later with the external experts if the question has not been sufficiently defined in the meantime.

Solution: To avoid having ill-defined questions during the elicitations, solicit the input of the advisory experts on the question. For instructions on how to pilot test the question, see chapter 8. If the worst comes to pass and the external experts are being elicited with ill-defined questions, you have two options. The first and best approach is to

gather the experts together and have them refine the questions. Each refined question should be recorded and declared the new question to be used henceforth. The other option is to record any definitions or assumptions that the expert individually uses in attempting to answer the question. The advantage of the first approach is that the experts are answering the same question and their answers, can be legitimately combined, if one composite answer is required for analysis.

6

Selecting and Motivating the Experts

In this chapter we detail how to select and motivate the experts for the two types of applications--those meant primarily to gather the experts' answers and those meant to gather data on the experts' problem-solving processes. These two applications are so different that they determine the approach to obtaining the experts' data. For example, studies that are to gather the experts' answers usually obtain the answers in quantitative form from 4 to 50 experts. Studies that will gather detailed data on the expert's problem-solving processes focus intensively on a few experts. Thus, for the first application, experts are likely to be selected for their diversity and ability to quantify their judgments in the desired form. But, in the second instance, the experts are frequently chosen for their willingness to devote a major portion of their time to being elicited and for their ability to respond to the method; for example, can they coherently think aloud for the verbal protocol method? The experts may even be screened initially by using a sample of the elicitation method.

For Applications Whose Data Will Be the Expert's Answers

For most applications, especially in risk/reliability and decision analysis, obtaining the expert's solution is the primary objective. There may be an attempt to document the expert's reasoning behind the answer but this is done to support the solution and is not the main goal. Usually, the expert's solution is requested in a quantified form, such as a probability. Thus, the experts need to have knowledge of the subject matter as well as knowledge of the rules governing the form in which they are to respond.

Who Is Considered an Expert?

An **expert** is anyone especially knowledgeable in the field and at the level of detail (granularity) being elicited: the individual should not be considered an expert unless knowledgeable at the level of detail being elicited. For example, an expert on different types of reactors would not be knowledgeable on the probability of a specific pipe's rupture

in a Westinghouse boiling water reactor (BWR). Similarly, a specialized pipe expert might not know the comparative likelihood of loss of coolant accidents (LOCAs) in Westinghouse BWRs and pressurized water reactors (PWRs).

What Constitutes Expertise?

Two types of expertise, substantive and normative, enter into projects whose goal is obtaining answer data. **Substantive expertise** comes from the expert's experience in the field in question, such as in rupture rates of Westinghouse pipes. **Normative expertise** is knowledge related to the use of the response mode. The **response mode** is the form in which the expert is asked to give his judgment (e.g., probabilities, odds, continuous scales, ranks or ratings, and pairwise comparisons). Normative expertise is based on knowing the statistical and mathematical principles of the response mode. Several response modes, such as probability estimation, are supposed to follow particular mathematical or logical rules (e.g., all probabilities are values in [0,1]). The use of individuals with expertise in neither or only one of these areas has been a serious problem in studies of expert judgment (Hogarth 1975). Both forms of expertise enter into the giving of a judgment, so the lack of either can affect the quality of the judgment. The lack of normative expertise may be responsible for there often being little difference between the *goodness* of substantive expert's judgments and those of inexperienced lay persons (Armstrong 1981). In particular, substantive expertise does not guarantee normative expertise as discussed in chapter 2, in sections *Are Experts Bayesian?* and *Do Experts Give Better Data?*

In general, a substantive expert who is experienced in the response mode (e.g., a pipe specialist with experience in probability estimation) is a better expert than one who is not (e.g., a pipe specialist without any experience in probability estimation). Two means for coping with this major pitfall is (1) to allow the experts to give their judgments in the deterministic mode that they use in solving problems at work, or (2) to try to familiarize them in use of the response mode. (See the discussion in chapter 9 on training project personnel in how to familiarize the experts with the response mode.)

When Expertise Matters

The above-mentioned parameters of expertise (substantive, normative, and knowledgeable at the necessary level of detail) are always of importance in gathering the expert's solutions. However, under particular circumstances, another aspect of expertise becomes important--the notability of the experts. Selecting experts who are well known and respected among their peers and the broader public can lend the study greater credibility. For example, one study forecasting America's needs and resources was initially very well received because of the endorsements of its illustrious experts (Club of Rome, *Limits to Growth* 1974). Therefore, if the study has the possibility of being controversial, aim to select experts who are notable as well as qualified along the other lines of expertise (substantive, normative, and level of detail). Selecting notable experts offers a side benefit. It often motivates other experts to participate in the study in the belief that they

will be in august company. Thus, obtaining additional experts for the study becomes much easier.

Additional Considerations in Selecting Experts

Multiple and diverse experts

It is generally advisable to obtain multiple and diverse experts so that the problems will be thoroughly considered from many viewpoints. Diverse experts are those likely to view and solve the problem in different ways. For example, Seaver (1976) proposes that having diverse experts, particularly in face-to-face meetings, leads to better quality answers. Ascher (1978:202-203) who has evaluated the accuracy of different forecasting techniques in retrospect, states:

> multiple-expert-opinion forecasts, which require very little time or money, do very well in terms of accuracy because they reflect the most up-to-date consensus on core assumptions.

Diversity of participants is one way to minimize the the influence of a single individual. For example, use of a single expert will slant results toward the contents and functioning of his memory. One expert will differ from another in what he has experienced, the interpretation placed on these experiences, and the ease with which they can be recalled and brought to bear on the problem (Hogarth 1980). Expert's answers are also likely to be affected by the mental heuristics that they used to simplify and solve the problem (Hogarth 1980, Tversky and Kahneman 1974 and 1981, Meyer and Booker 1987b). The use of diverse experts allows the answers to reflect individual differences in experience, recall, and use of problem-solving heuristics.

Diverse experts are likely to be important in cases where the experts are to forecast future events or situations (e.g., predicting the market for nuclear power in the year 2000). In forecasting, the tendency is to anchor to the *status-quo* situation and not extrapolate sufficiently in considering the future (Ascher 1978). Having multiple experts with different viewpoints helps the group overcome the human tendency to anchor to one, usually conservative, reference point.

The practice of using multiple experts is being encouraged in knowledge acquisition (Boose and Gaines 1988). One advantage of eliciting and showing different expert's judgments is that the user of the knowledge-based system can pick the expert's way of thinking that he finds most useful or appropriate (Gaines and Shaw 1989).

Other studies on aggregating multiple expert estimates, such as that of Martz, Bryson, and Waller (1985), support the practice of having multiple experts. Aggregation schemes tend to show that a combined estimate has a better chance than any single expert's estimate in getting closer to the *true* value.

Number of experts

The exact number of experts that *multiple* constitutes may vary according to the elicitation method. For example, if a face-to-face meeting is involved, we recommend having from five to nine experts for each interviewer available to moderate the sessions. Fewer experts than five does not seem likely to provide enough diversity or enough

information for making inferences (chapter 18). Nine experts in a session is usually the upper limit for obtaining in-depth thinking from each expert and yet having enough control to counter potential effects arising from group dynamics, such as the follow-the-leader effect.

Selection Schemes

Most of the selection schemes are based on having the experts name other experts. In many specialized fields (e.g., seismology, high explosives, and nuclear reactor phenomenology), the experts know one another and can supply the names of other experts. The researcher starts with a few of the known experts, collects names from them, and repeats this process until more names are gathered than are likely to be needed.

The problem with using this scheme without modification is that it leaves the study open to later questions of whether people named those with similar views. Because diversity of experts is the goal, this basic scheme is often combined with additional selection criteria, such as having equal numbers of experts from academia, private industry, and the government, or from the major points of view. Some of the selection criteria may be used to define the meaning of expert (e.g., criteria beyond being a person who is recognized as being an expert by other experts). For example, only experts with particular levels of publication or experience, such as in being a plant operator, might be chosen from those named.

Then too, logistics play a role in the selection scheme. The scheme must respond to such concerns as whether the experts will be be willing to participate, have the time to participate at the necessary level, and be allowed to do so by their employer.

The selection scheme is likely to receive close scrutiny if other aspects of the study, such as its results, are questioned. The most frequent criticism is that the scheme did not select experts who were representative of the larger population and that their answers were, therefore, skewed. It is commonly believed that skewed results arise from taking the majority of experts from one place such as the same organization (e.g., especially from the same organization as the rest of the project personnel), a class of organizations (e.g., from academia, industry, or government), or one point of view. Our studies (Booker and Meyer 1988a, Meyer and Booker 1987b) have not found the expert's affiliation or education to be a significant factor in explaining similarities or differences between expert's answers. However, in the interest of trying to represent different views and to avoid criticism, we recommend selecting a balanced group of experts.

Motivating Experts to Participate

The first step of motivating the experts is to consider the proposed study from the viewpoint of the experts. Theories on interviewing predict that obtaining participants depend on maximizing those factors of the situation that experts would consider motivating, such as recognition, and minimizing those that they would find inhibiting, such as having to devote large amounts of their time (Gorden 1980).

For example, two aspects of risk assessments that could be maximized to motivate the experts are the opportunity to affect the study or contribute to the field and the

opportunity to receive recognition. If the potential experts are told that they will have input into the methods used, they will be more likely to volunteer. Generally, if individuals have control over a process, they feel better about it and will lend support to the methods used or to the conclusions reached. This optimistic attitude stems from the belief that *if I did it, it has to be good.* Also, if the experts judge that the study will be done in a manner that will bring credit to them or that their reputations will benefit from being included in this company of experts, they will be more willing to participate.

Care needs to be taken to remedy those aspects of the study that may be viewed as inhibiting the experts' participation. Generally, having to devote large amounts of time to participating in the study is a common inhibitor in risk assessments. This inhibitor can be minimized directly by reducing the time required for the study or indirectly by either increasing the attractiveness of other aspects of the study, such as offering the experts a larger role and thus a greater chance for recognition, or, if all else fails, offering to pay them for their time.

Motivating the experts through pay

Generally, we believe that paying the expert should be reserved for those situations where there are no aspects of the study that can be used to motivate participation or where participation requires major investments of the expert's time and thought. Focusing on how the intrinsic aspects of the study can be developed to encourage participation can produce more effective motivators and also improve the design of the study. Paying experts for their time should be a last resort for several reasons: it is costly; it may attract one type of participant and slant the results (Gorden 1980:118); or it may have unexpected affects on the participants' view of the study.

Payment can affect the expert's views through a means of psychological adjustment (cognitive dissonance) illustrated below. If the expert is not paid, he must convince himself that he is expending his time and effort for good reason, or he will feel duped. Studies have shown that the participant unconsciously resolves this dilemma by focusing on the merits of the study and on the benefits derived from participation (Baron and Byrne 1981:122). If, on the other hand, the expert is paid, he is not led to consider the positive aspects of the study. Then too, the expert may view the payment as a bribe for participating in a study that could not obtain experts in any other way (Baron and Byrne 1981:124). In an unconscious effort to show that his cooperation can not be bought, the expert may take an extremely critical stance toward the study. One exception to the policy of not paying the expert is in the area of travel and lodging expenses. Experts are not likely to interpret this practical coverage of expenses as payment for their time or cooperation: thus they should be paid for travel and lodging as required.

Motivating the experts through communication of intrinsic aspects of the study

How aspects of the study are communicated to the expert is likely to affect his desire to participate. For this reason, we recommend that a brief memo (about one page) be drafted as preparation for requesting participation. This memo can be used as a script for the telephone conversation or face-to-face meeting with the expert to request his participation. Generally, more individuals will respond to a request delivered in person

than by mail. For this reason, it is recommended that the experts be contacted or called first and then sent the memo.

Guidelines abstracted from communications theory (Stroud 1981, Gorden 1980) and the authors' interviewing experiences suggest that particular items of information be communicated. Typically, the potential participant will want to know this information and in the following order of importance:

1. **The reason that he is being contacted.** It is a good practice to phrase the first sentence requesting the expert's participation in a manner designed to motivate. For example, *I would like you to participate in a study of Y because of your considerable knowledge of X.* This request could be considered motivating because it is a personal appeal for assistance (e.g., *I would like you. . .*) and because it recognizes the person's expertise. It is important that this introductory sentence be motivating because many individuals decide in the first few seconds of scanning a letter whether they are interested or not and if not immediately throw away the letter. Thus, if the experts can not be called or contacted in person, the first part of the letter is a critical point in creating interest.

   In the authors' experience, scientists have responded well to these motivators:

   - **Recognition.** This recognition can come from the project personnel's selecting the experts to participate or from other experts in recommending prospective participant names. The opportunity for further recognition would come from the expert's work in the study.
   - **Altruism.** Altruism can range from helping another person (e.g., the interviewer, by agreeing to participate) to helping the human race by contributing to the advancement of science. Most scientists will be interested in taking the state of the art a little further or in examining problems with current methods.
   - **Experiencing something new and different.** Most people enjoy an occasional break in their routines, and scientists are no exception. In fact, they may have more active curiosities. (A few sources on scientists' personality traits are Mahoney 1976; Roe 1952 and 1963; Cattell 1963; and Knapp 1963.)
   - **Need for meaning.** Often scientists are interested in how their work fits into the larger picture. For example, a computer modeler of reactor phenomena might be interested in how his data is used in assessing risks and setting new standards. An individual's work becomes more meaningful if he can see how it will be used or how it will affect others.

2. **Who is conducting/sponsoring the study.** This information would generally be given before item 1 in a conversation, as opposed to a letter. For example: *Hello Dr. Jones. I'm John Smith of the Research Division at the NRC. Because of your expertise in dry subatmospheric containment in Westinghouse PWRs, I would like you to participate in a risk assessment study of the Surry plant.* The expert may have some impression of past studies done by this organization. If the expert's impression is likely to be negative, mention

how this study differs from previous ones (e.g., it attempts to rectify particular problems or to develop better methods in some area).

3. **How much time/effort this study is likely to take, over what period of time, and when it will start.** Earlier, we recommended that the reader consider the factors in the study that might inhibit or motivate potential participants. Most experts will be busy and unable to devote large portions of their time. However, if the study cannot be made less demanding of the experts' time, think about how to increase those factors that would motivate them. For example, the expert may weigh the time that the study will take against its likely contribution to the field or to his reputation. If the study's goal is to set new standards and the experts will be contributing to the creation of these, this information should be mentioned to offset the heavy time demands. If the study will include the most noted experts in the field, this factor should also be mentioned as a probable motivator.
4. **How he was selected or who referred him.** Basically, the expert will want to know how he was selected. Experts will be more interested in participating in a study for which they were specially, rather than randomly, selected. (Thus, if the experts were selected at random, it is best to gloss over this fact.) Even more motivating for the expert is to know that he was recommended by persons that he respects, some of whom may also be involved in the study. If the latter is the case, it should be mentioned to encourage the expert to assist in the study.
5. **What, in more detail, would he be doing in the study.** Before committing himself, even tentatively, to the study, the expert may wish to know his tasks or role. Try to avoid using technical jargon to refer to the methods for eliciting or modeling the experts' responses. The following is an example of a general description of the tasks abstracted from another study (Bernreuter et al. 1985): *Your role will consist of three parts: (1) helping define seismotectonic zones east of the Rocky Mountains; (2) giving your opinion on the occurrence rate and magnitude distribution of earthquakes within the zones; and (3) reviewing/refining your input and that of the other experts.*

   If the tasks seem very demanding, the expert may request the provision of background materials or training. If the details on providing these to the experts have not yet been worked out, state that background materials and training will be provided, as needed.

   The expert may also want to know if he will be required to give answers on tasks or on areas where he is not knowledgeable. Scientists often raise this issue when they are still new to the study and concerned about their ignorance. Tell the expert that he will not be asked to provide his judgment until he has received training in the response mode and has become familiar with the study. [Note: If after the expert has received the training and briefings, he is still reluctant to provide his judgment, he should not be forced for two reasons: (1) he is probably not an expert in this area if he does not feel qualified to give his judgments; and (2) his reaction to being forced is likely to be negative and to detrimentally affect his view of the entire study. This reaction is illustrated by

reviewer's statements about a study: "The participants were forced to provide unsubstantiated guesses as input." (Benjamin et al. 1987:F-5,6)].

In general, the expert can be told that he will not be forced to give judgments where he is not expert because that would detrimentally affect the quality of the data. Emphasize that the goal of the project is to collect judgments that are based on careful consideration and experience.

6. **Will the judgments be anonymous, and if so, how will confidentiality be maintained.** If the study will be sensitive in nature (e.g., very hot politically), potential experts will probably want to know how confidentiality will be handled before learning about the study in more detail. Thus, for a sensitive study, the information on confidentiality should be given before the details of the study listed in item 5. In the case of a sensitive or controversial study, consider making the experts' estimates and comments anonymous. However, it is best to be guided by the experts' wishes on this issue. If there is some question about anonymity at the time of contacting the experts, the experts can be asked to help establish how anonymity will be handled. It may be that the experts wish to have their estimates or thoughts identified if they perceive the study as being important and if they have had a significant role in shaping it. However, expert judgment studies have traditionally followed the social or behavioral science norm of confidentiality, of not identifying the expert's data.

   Some levels of confidentiality are to list the organizations or offices that have contributed experts, to list the names of the experts plus their affiliations, or to identify the data provided by each expert in addition to listing them and their affiliations.

   Generally, the wishes of those experts who request the highest level of confidentiality should be applied. For example, the majority of experts may want their judgments kept anonymous but want to be listed as having participated in the study. One expert may oppose being listed as an expert or even having his organization named as having provided a representative. The one expert should be given the level of anonymity he requests, even to the point of extending this level to include all the experts, if necessary.

   For the occasional highly sensitive study, the potential participants may wish to hear exactly how confidentiality will be maintained (e.g., how the data files will be stored and who will have access to them). They will use this description to judge whether they would be likely to lose the protection of confidentiality if they chose to participate.

7. **The anticipated product of the study and their access to it.** Generally, one product of the study will be a report and the experts will wish to know if they will be sent copies. For many people, knowing that there will be something tangible to show for their efforts is a major source of motivation.

8. **Whether participation will be required or voluntary.** For most studies, participation will be voluntary, and this fact should be stated. However, usually this statement need not be made until after the benefits of participation have been fully elaborated.

For Applications Whose Data Will Be Problem-Solving Processes

It has become more common to obtain problem-solving data, perhaps as a result of the influence of artificial intelligence. However, this focus occurs for a variety of reasons, such as when the application is to do one of the following:

- Determine how problems are currently being solved and perhaps set new standards
- Use the data for evaluating novice's methods of solving problems and for the training of novices
- Gather data for the building of an expert- or knowledge-based system

While the expert's answer is usually gathered in these studies, it is not considered the main data but rather a part of the problem-solving information.

What Is Needed in an Expert

Applications whose goal is to gather problem-solving data require more of the expert than those whose goal is primarily to gather answers. The expert needs to be extremely skillful in solving problems (Welbank 1983:8), articulate in describing his or her problem-solving processes (Waterman 1986:192), and willing to commit to this difficult, time-consuming task.

In particular, articulate experts are rare. In the process of becoming expert, many of the expert's basic thought processes have become automatic or unconscious and thus inaccessible for articulation. It is thought that humans progress from learning and consciously manipulating rules, such as those of grammar, to more abstract thinking and less conscious use of rules or procedures (Dougherty 1986, Denning 1986). Yet, experts must somehow regain awareness of their thoughts to assist in explaining, representing, checking, and refining that process.

In general, the expert must be available for providing this information. Usually, availability requires that the interviewer or knowledge engineer and the expert be in the same city. How accessible the expert is can be a separate concern. Frequently, even with the expert being in the city, he becomes less accessible as the project drags on and his interest decreases.

If the goal of the project is to gather problem-solving data for building an expert- or knowledge-based system, the qualities of the expert become even more critical. McGraw and Harbison-Briggs (1989:99) list the following personal characteristics and attitudes as desirable:

> domain experience, sense of humor, good listener, sense of committment, patience, ability to communicate ideas and concepts, introspective of own knowledge, willingness to prepare for the session, honesty with self and others, and persistence.

Method-Driven Selection

Often, the interviewer or knowledge engineer does not have a choice of experts--the expert is simply appointed by the organization that is funding the work. If this is the case, then the methods of elicitation must be tailored to the expert so that he is able to respond to them. If however, there is a plethora of experts, they can be selected according to the methods planned. Because of the in-depth nature of this application, the focus is usually either on one expert or one elicitation method at a time. Thus the expert can be chosen for his willingness and ability to be elicited by a particular method. (See chapter 7 for a description of the elicitation methods.)

A trial of each elicitation method can be conducted on each expert to determine which combinations of expert/method work best. (To run a trial, select a sample problem and follow the instructions in chapter 10 on how to administer the verbal protocol, the verbal probe, or the ethnographic technique.) Of the elicitation methods, the verbal probe and ethnographic technique can be used on the greatest number of people. The verbal protocol is more restrictive in that some experts can not use it. We have found that about one in thirty experts becomes extremely frustrated in using this method because it interferes with their thinking.

Motivating the Expert

As mentioned in the section on motivating experts to participate in answer-gathering applications, the goal in problem-solving applications is to maximize those aspects of the study that humans find motivating and minimize those which have the opposite effect (see *Motivating Experts to Participate* above). The information is then communicated to the potential participant in the same manner as detailed in *Motivating the Experts Through Communication of Intrinsic Aspects of the Study.* The differing aspects of motivating participation in a problem-solving application is described so that the communication to the expert can be adjusted accordingly.

The main factor discouraging expert's participation in the problem-solving application is the great amount of time that it requires. Sometimes, the burden on an expert can be lessened by using several experts. However, if the number of experts is limited or if learning one expert's thinking in depth is important (as is common at the beginning of such projects), this may not be possible. A second inhibitor can be the experts' fears that the model/system could replace them or show their thinking to be faulty. For example, if management volunteered their participation, they may be suspicious of why management wishes to examine their thinking. These fears can be alleviated by explaining the purpose of the project and eventual capabilities of its product. To further reassure them, state that there are no *right* answers, explain that this is not a test, and convey a nonjudgmental interest in learning their thinking. A third inhibitor is the experts' concern that they will not be able to tell how they solve problems because they do not know. Their concerns can be answered by explaining that this lack of awareness is part of being an expert and that methods have been developed to extract the information. (For inhibitors specific to experts involved in building expert- or knowledge-based systems, see McGraw and Harbison-Briggs, 1989:117-125.)

Just as the problem-solving application presents additional inhibitors, it offers stronger inducements for participation than its answer-gathering counterpart. In addition to the motivators mentioned earlier for the answer-gathering application, the problem-solving application offers a greater opportunity for motivation through altruism and the experiencing of something new and different.

Some of the altruistic rewards offered by problem-solving applications are the opportunity to help different groups of people. For example, there is an opportunity to aid students if the application is to provide the expert's thinking as instruction as in Elston et al. (1986). Others in the expert's field can benefit if the application performs a function such as identifying likely locations for petroleum deposits (PROSPECTOR) or offering decision-making guidance in export control (Meyer and Johnson 1985). The work can also serve the larger public if the application is to provide a service, such as medical diagnostics (MYCIN).

Problem-solving applications offer participants the opportunity to experience something new and different--insight into how they think. While this opportunity is appealing to most people, it is especially so to scientists. In fact, scientists will often request references on how practitioners in their field think and act. (Mahoney 1976, Roe 1952 and 1963, Cattell 1963, and Knapp 1963 are good sources for this information).

Lastly, if the product of the application will be marketed, recognition and financial gain can be incentives for participation.

A major problem is keeping the expert motivated through time. Generally, the incentives that were used to interest the experts in the project can be used to maintain their interest. As Waterman (1986:194) notes:

> Making the expert feel like an integral part of the system-building process will motivate him or her, as will showing the expert how the system will ultimately produce a useful tool.

He adds that long periods between interviews should be avoided because they diminish the expert's interest. Welbank (1983:10) recommends the following means for recharging the expert's enthusiasm: selecting an interesting problem, such as one which has received recent publicity; holding panel discussions in the belief that many experts enjoy criticizing one another; and producing a prototype of the expert's thinking for him to review.

Common Difficulties--Their Signs and Solutions

Difficulty: The experts do not wish to participate. In most applications, somewhere between one-third and three-quarters of those experts called will initially agree to participate. If fewer than one-third agree to participate, be alerted to a potential problem in motivating the experts.

Solution: While it is likely that the communication with the expert is the culprit, we recommend gathering more specific information before reworking the presentation of the rewards of participation. One of the experts earlier contacted can be questioned to learn what may be inhibiting participation. If one of these experts is an acquaintance or known to be outspoken, he would be a natural choice. Ask this

expert to provide, in confidence, his real reasons for not participating. If asking an expert for this information seems awkward, consider asking the advisory experts. The advisory experts, because of working on the project, are likely to have a different view of it than the external experts. For this reason, the advisory experts are not our first choice. The advisory expert will need to be briefed on what was said to the other experts and on the experts' responses. Ask the advisory expert to place himself in the expert's place and to suggest reasons for the low participation rate.

If the above suggestions are inappropriate or unsuccessful, itemize the motivators and inhibitors of the study again. It is likely that the motivators are not strong enough to overcome those things inhibiting the expert's participation. Inhibitors may be discovered that were not taken into account. In our experience the one factor that most causes experts to balk at participation is the belief that their effort and judgment will not be used. We have been told, in some such cases, that the experts thought that decisions had already been made at high levels and that their judgments would not be used as input. If this perception, true or false, is hindering expert participation, it needs to be addressed. The project's client and the other project personnel can be helpful in suggesting solutions.

Difficulty: Everyone, including nonexperts, wishes to participate. The participation of everyone may seem to be an embarrassment of riches to those encountering the difficulty mentioned above. However, it is a difficulty even though it can be easily resolved.

Solution: Everyone can participate, if participation is broken into classes and run according to a few rules. In general, the volunteer participants should be grouped according to their expertise. For example, in the army export control project, most of the experts were knowledgeable in one specific question area but wished to provide their inputs to all the areas. All the experts were allowed to present their views and answers in a structured manner. However, only those experts who had been previously designated as experts in the area gave their answers as votes, which were then documented for analysis. The other expert's answers were simply recorded as commentary. Both groups were satisfied. The nonvoting experts had had the opportunity to express their views and ensure that the voting experts were not overlooking some important information. The voting experts knew that their expertise had been acknowledged and that their answers would be used to establish policy.

Frequently, the participants can be asked to sort out their participating rights and statuses. In the above example, it was the participants who suggested the voting scheme.

Difficulty: The real experts are too busy to participate at the needed level. As a general rule, this difficulty is considered much more serious if it occurs in problem-solving applications rather than in answer applications, especially if the application is to emulate the expert. The signs of this problem are similar to the above-mentioned one of the experts not wishing to participate.

Solution: If this difficulty occurs when the experts are first being called, the problem is in the communication. Refer to the suggestions given above. Note that the best experts *are* often more busy than those less expert or that they may need greater or different motivators to be persuaded to participate. Consider asking the expert what would convince him to participate.

Difficulty: The system for selecting experts is criticized. We have observed that selection schemes are frequently criticized. The most common complaints are that the selection is biased or that it does not include the *real* experts.

Solution: If the credibility of the selection is questioned after the expert judgment has been elicited, there is little that can be done. The only thing that can be done is to document the selection scheme as a means of explaining and defending it. We recommend recording the selection criteria, the reasoning behind the use of the criteria, and the number of those who were invited to participate versus those who actually participated.

If the selection scheme is criticized while it is still in the design stage, there are two approaches. One approach is to rethink the selection scheme. We suggest targeting the top experts in the field and trying to motivate as many of them as possible to participate. If the top experts accept, the selection will not be open to the common criticism that no real experts participated. If the top experts cannot participate full time, accept their partial help. If none of the top experts will participate, at least their participation was sought. In addition, we recommend designing the selection scheme around features that are expected to make the judgments of the experts' differ. For example, the experts are often expected to differ according to where they have worked or gone to school. Experts can be selected to represent these different features.

Another approach to handling criticism of the selection scheme is to put the criticizers in charge of designing a selection scheme. Having to design a selection scheme will make them more appreciative of the difficulties involved. Double check their selection scheme to ensure that it is not open to one of the common criticisms. Then, use them as manpower in calling the experts. The experience of trying to implement their selection scheme will make them aware of how things often go awry. If they do not participate in making the calls, they may later wonder why their scheme was not implemented exactly. If they assist in the calls, they will understand that some experts could not participate and that substitutions had to be made.

Difficulty: There is a conflict between those wanting to identify the expert's data and those wanting to preserve anonymity. The conflict stems from the two views of identifying the expert's judgments. One view is that the expert data will be more credible if it is labeled by the expert's name. Proponents of this view believe that the experts will exercise more care in giving their judgments if these judgments are attributed to the experts and that better quality data will result. The second viewpoint favors anonymity in the belief that the experts will not give their true answers if others can trace the answers back to their

sources. Proponents of this second view argue that the confidences of interviewees have traditionally been protected.

A sign that this difficulty is occurring is disagreement among project personnel, clients, or experts on anonymity.

Solution: The means for resolving this conflict of views is to have the experts decide how they wish to have their judgments identified. The experts are the persons to make this choice because they are the providers of the data and they may withhold their judgments if they are uncomfortable. Explain to the experts that their decision on anonymity will be followed but that they will have to reach a consensus. This puts the burden on them for swaying the members of their own party that have differing ideas. If the experts fail to reach a consensus, impose the highest level of protection requested, even if only one expert wishes his judgments to be anonymous.

7

Selecting the Components of Elicitation

This chapter is designed to guide the reader through planning what is needed to obtain and later analyze expert judgment data for a particular application. It offers the checklist shown below to assist the reader in determining which of the five basic components, the building blocks, of elicitation will be needed. Then, in subsequent sections, it aids the reader in selecting the most appropriate methods from within the selected components. For example, the reader might use the checklist below to decide that the expert's problem-solving processes need to be elicited. The reader may then select the verbal protocol and the verbal probe as the best combination of methods for accomplishing this task.

The basic methods are presented because people often wish to use an existing method rather than try to create a new one and possibly reinvent the wheel. Also many people prefer to use an existing and accepted method in the belief that it will enhance the credibility of their work.

In the next chapter (chapter 8) information is provided on how the methods selected in this chapter can be tailored to the reader's application.

Determining Which of the Five Components Are Needed--Checklist

Check	When a Component Is Needed
_____	1. **An elicitation situation is needed if** expert judgment is to be gathered. Regardless of specific project requirements, some staging is necessary for arranging how the experts and data gatherers will meet and how the expert judgment will be obtained. ***Definition of Component:*** *An elicitation situation is the setting in which the expert's judgment is elicited. Expert's judgments can be elicited in private, in a group setting, or when the experts are alone but receiving information on the other expert's judgments. More than one elicitation situation can be used in a project.*

For example, the experts could meet as a group to discuss and revise the questions. Later, they could be interviewed separately for their final judgments.

_____ **2. A response mode/dispersion measure is needed if** the expert's answers must be in a specific form and if the experts will be asked to make this conversion mentally as opposed to having someone later translate their judgments into the desired form. Often, project personnel wish to use a particular model and therefore want the expert's answers to be in a particular form. This means that either the expert must conform to the desired mode or the analyst must transform the expert's judgments into that form. In general, we favor using a response mode rather than converting the expert's judgment into the desired form later. We believe that the former practice is less likely to lead to misinterpretation or misrepresentation of the expert's data, providing that the expert can accurately encode his thoughts into the requested response mode (either naturally or with training).

Definition of Component: *A response mode is the form in which the expert is asked to encode his judgment. Some modes that are handled in this book are estimates of physical quantities, probability estimates, odds-ratio, probability distributions, continuous scales, ratings or rankings, pairwise comparisons, and Bayesian updating. Additionally, the expert is often asked to provide a measure of dispersion on his judgment (e.g., 0.9 ± 0.1). A dispersion measure is the amount of variation or spread in the data. Dispersion measures can also indicate the amount of uncertainty in the data. The dispersion measures covered below are ranges, percentiles, and variances or standard deviations.*

_____ **3. Elicitation of problem-solving processes will be needed, if**

(a) the goal of the project is knowledge acquisition, such as in building a knowledge-based or expert system;

(b) there is likely to be interest in how the experts arrived at their answers;

(c) the answers are to be aggregated. If the experts defined the questions differently, combining their answers could be like mixing apples and oranges. Use of data on their problem solving could prevent this mistake.

Definition of Component: *Elicitation of problem-solving processes involves obtaining data on how the subject solved the problem. Data on problem-solving processes can be gathered to any level of depth.*

_____ **4. Aggregation of expert's answers is needed if** multiple experts (and, therefore, usually disagreeing experts) will be used and a single representation of their answers is needed. For example, one might want to combine several expert's problem-solving procedures to produce one procedure for use in an expert system. Similarly, in a

risk analysis application, one might aggregate several expert's estimates to enter one estimate into the model.

Definition of Component: *Aggregation is a means of obtaining a single datum from multiple and differing expert data. For example, experts can be required to agree among themselves as to what answer they give or they can be allowed to give different answers that are then, later, mathematically combined (chapter 16).*

_____ **5. Documentation is needed if** having a permanent or semi-permanent record of the expert's answer and problem-solving processes is desired. Documentation can include information on which expert gave each estimate and their reasons for giving these answers. Documentation is often used to provide traceability on the expert's decision. Traceability becomes important if the judgments are likely to be reviewed or to require updating.

Definition of Component: *Documentation is a record of the expert's judgment and/or of how that judgment was reached.*

Selecting From Within Each Component

Selecting From Elicitation Situations

There are three major methods or situations for eliciting the expert's judgment: with the interviewer in a private face-to-face interview with the expert; with the interviewer in an interactive meeting of the experts; and with the expert in physical isolation from the interviewer or other experts but communicating his data by mail (electronic or postal) or telephone. These situations can be tailored and combined to fit the application. For example, the interactive group could be structured to be like a technical conference where each experts is scheduled to present his views prior to the group's discussion. The use of the interactive group method could be combined with that of the individual interview to elicit each expert's judgments apart from that of the other experts.

Interactive group

The interactive group is where the experts meet in a face-to-face situation with one another and a session moderator or interviewer. The expert's interactions with one another can be structured to any degree. An unstructured group resembles a traditional meeting; a highly structured group is carefully choreographed to prevent spontaneous interaction (to limit the negative effects of interaction, such as group think).

Advantages:	*Generates more accurate data, particularly for predictions, and a greater quantity of ideas than the other two situations. (These two results are attributed to the synergism created by expert's sharing their thoughts.)*
Disadvantages:	*Possesses the potential for group-think bias. Poses logistical problem in scheduling and handling multiple interacting experts, particularly if there are more than four or five.*

Summary of studies: *According to Seaver (1976) who did comparative studies of these three elicitation situations, the interactive group method produces a greater quantity of ideas and higher member satisfaction with the product than the Delphi He also noted that the majority of subjects in a group of diverse membership improved their accuracy following a discussion. Fogel (1967:375) commends the interactive group for solving problems that require "originality and insight" and not routine tasks. He found that predictions made by groups were more often correct than those made by individuals. In general, studies comparing a structured interactive group with one or more of the other methods favor the former (Seaver 1976, Armstrong 1981, Gutafson et al. 1973, Gough 1975, and Van de Ven and Delberq 1974).*

Delphi

Delphi is where the experts do not directly interact with one another or the moderator. The experts, in isolation from one another, give their opinion data. These are collected by the moderator, made anonymous, and distributed to the experts to allow them to revise their previous judgments. The experts can be allowed to revise their estimates until consensus, if it is desired, is achieved. This method was developed by RAND to limit the biasing effects of interaction.

Advantages: *Designed to avoid biases arising from group dynamics. (However, some question whether it accomplishes its design purpose.)*

Disadvantages: *Limited in the amount of data that can be gathered (e.g., not suited to gathering data on how the experts solved the problem, except for their sources of reference. Less synergism than in the interactive group. Usually, the most time consuming of the three situations because of the turnaround time through the mail. (If the Delphi were done by electronic mail, it would be less time consuming.)*

Summary of Studies: *For all of its use, the Delphi method has not had extensive empirical investigation, and its reviews range from the positive to the negative. Several researchers consider the structured interactive group to be better than the Delphi in terms of avoiding the bias the Delphi was designed to avoid (Seaver 1976 and Armstrong 1981). It would be natural to hope that the experts have converged on the "right" answer when they have reached consensus in the Delphi. However, Dalkey (1969) found that the number of rounds in Delhi corresponded to increasing agreement but not to increasing accuracy.*

Individual interview

Individual interview is where the expert is interviewed alone, usually in a face-to-face situation with the interviewer. This situation can be structured to any degree. An unstructured interview occurs when the interviewer has not outlined data-gathering goals or questions in advance.

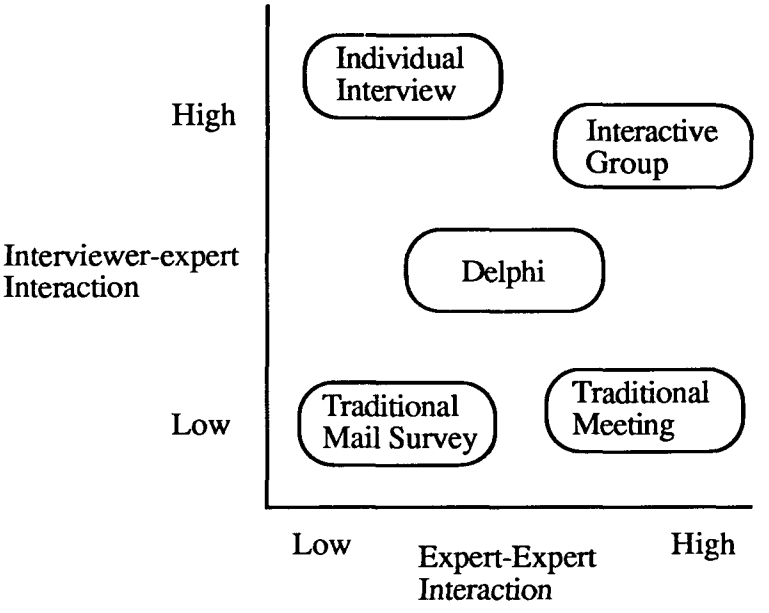
Advantages: *Best method for obtaining detailed data. Main method used for obtaining data on the expert's problem-solving processes. Avoids potential bias from group dynamics and data can be combined later, usually by mathematical aggregation.*

Disadvantages: *Time consuming. No synergistic effects from interexpert discussion.*

Summary of Studies: *Seaver (1976) found that mathematically combined estimates from individual interviews outperformed single estimates. When the answers from the individual interviews are mathematically combined, the individual interview is termed the staticized or nominal group method.*

Results from the staticized group situation have been judged poorer than those from interacting groups (Stael von Holstein 1971). However, it should be noted that the above-mentioned studies were not using the individual interviews to elicit deep problem-solving data, the task for which it is most suited.

EXAMPLE 7.1: ***Relative Interactiveness of Elicitation Situations***



(This figure was excerpted in part from Armstrong 1981:104)

■

Selecting from Response Modes

The response modes given below are organized according to the forms that are commonly needed to answer the types of questions asked of experts. For example, a question on the number of homeless currently in the United States would require an estimate of a physical quantity. In contrast, a question on the likelihood of some occurrence would need probability estimates, odds, or distributions. Questions based on the comparison of two or more things would require pairwise comparisons, continuous scales, or ranks or ratings.

In addition, some questions require single estimates, such as one probability value, while others need multiple estimates, as in an expert's probability distribution. Of the response modes listed below, the estimate of a physical quantity, the probability estimate, the odds ratio, the ranks or ratings, and the continuous scale are utilized to obtain single estimates. Probability distributions, Bayesian updating, and pairwise comparisons are most frequently employed to obtain a set of estimates from the expert, although the continuous scale, and ranks or ratings can also be used for this purpose.

To help the reader find the response modes appropriate to his questions, the modes are listed separately as well as under the more inclusive mode to which they belong. For

example, odds ratios are listed separately and also as one of the ways of eliciting a probability response.

Estimate of physical quantity

The expert gives an estimate of a physical quantity, such as of temperature, time, pressure, volume, or flow rate, in response to a technical question. For instance, the expert could be asked to provide the engineering specifications for a component, the parameters needed to achieve a particular milestone in magnetic fusion, or the amount of coal reserves currently in the United States. Sometimes estimates of physical quantities are part of the phrasing of the question and thus can be used in combination with other response modes. For example, in a magnetic fusion project (Meyer et al. 1982:424), the parameters needed for confinement were elicited from an advisory expert, refined by an external expert, and then used in questioning several experts. These experts estimated the time needed (a physical quantity) and the probability (another response mode) for achieving these physical parameters. Similarly, in the reactor risk project NUREG-1150 (Ortiz et al. 1988), the values of physical variables were sometimes elicited at given cumulative probabilities (e.g., *what is the inner core temperature such that the probability is 0.90 for that temperature or higher?*).

Advantages: *A convenient and flexible form for answering questions on poorly understood or difficult to measure physical processes. The expert has little or no difficulty in understanding estimates of physical quantities as response modes because he uses them in his work. Thus, the expert does not need to be trained in the use of this response mode.*

Summary of Studies: *To our knowledge, there are no studies that evaluate estimating physical quantities with respect to some other response mode. However, there are two types of studies that address estimates of physical quantities: (1) studies (e.g., Ascher 1978) that examine in retrospect the accuracy of experts' predictions, such as the size of United States' market for petroleum in some year; and (2) studies that examine properties of human judgment by asking questions whose answers can be determined. There are many more studies of the second variety and most of these have used almanac questions (e.g., Armstrong et al. 1975, Martz et al. 1985), where the subjects estimate physical quantities, such as a city's population of persons over age 50. The results of these studies have shown that the accuracy of expert's estimates of physical quantities can be affected by the assumptions that the expert makes (Ascher 1978) and by decomposition of the question--decomposition being associated with greater accuracy. (These results are thought to apply to the other response modes also.)*

Probability estimate

A probability estimate is a single value given by the expert (e.g., 0.45) in response to the question. Usually probability estimates are used to predict the likelihood of some event occurring. Probability estimates can be asked for in different ways: *What is the probability that an event will occur?* (fixed value); *what are the number of occurrences in n total trials using a log scale?* (log odds); and *what are the number of occurrences or events in n total trials?* (See *Odds Ratio* below.) Multiple experts' estimates to the same question are sometimes linked to form a probability distribution, which is then used as the answer. (See *Probability Distribution* below.)

Advantages: *Commonly used in decision analysis. In fact, one advantage in using a probability-based response mode is the existence of established elicitation and analysis techniques. (For this reason we recommend the use of decision analysts and decision analysis techniques if this response mode is chosen.) In general, probability estimation is a very convenient form for modeling and analysis.*

Disadvantages: *Most experts are not good estimators of probability values and may be reluctant to use this response mode. (The use of probability wheels and training in probability estimation can mitigate these problems.) Done correctly, probability estimation is a very time-consuming process. It can fatigue the expert.*

Summary of Studies: *It has generally been shown that humans are biased in their estimation of probabilities (Tversky and Kahneman 1974), that they do not properly interpret probabilistic phenomena--like randomness, statistical independence and sampling variability (Hogarth 1980), and that most do not feel capable of using this response mode (Spetzler and Stael von Holstein 1975, Welbank 1983). Hogarth (1980:149), who has done extensive studies on cognitive and motivational biases, acknowledges that "probability theory itself is difficult to learn and apply" but argues that it is the best choice for expressing uncertainties. He cites the numerical precision of probabilities and their logic for structuring relationships between events as reasons for using probabilities. Welbank (1983:28) notes that the precision of probabilities is not always needed or justified but states that probabilities may be useful where the knowledge is vague and where there is need for weighting of the answers.*

Odds ratio

An odds ratio is a response that follows the form of x chances out of n total trials. For example, an expert could state that there is 1 chance in 1000 of a particular event occurring. The odds ratio is most frequently used to estimate the frequency of rare physical events. (Odds ratio is also mentioned above under *Probability Estimate* and next under *Probability Distribution.*)

Advantages: *A convenient form for estimating the likelihood or frequency of rare events. We believe that it is easier for most people to think of rare events in terms of odds (e.g., 1 in 1000) rather than in probabilities (e.g., 0.001).*

Disadvantages: *If the expert is given the total number of trials (n) from which he is to estimate the occurrence of some event, setting too small a total can affect the expert's judgment and/or frustrate him.*

Summary of Studies: *In general, humans tend to overestimate the likelihood of rare events. According to Seaver et al. (1983:2-7), odds are one of the best procedures for estimating relatively unlikely events, especially if the odds are on a logarithmic-spaced scale. In general, rare event estimates are thought to be less biased when they are elicited as odds, such as 1 in 1000, than when they are elicited as decimals, such as 0.001 (Boose and Shaw 1989: 71).*

Probability distribution

Probability distribution is used here in a very broad sense to mean a set of possible values for the estimate and the associated likelihood or probability for each value's occurrence. The set should include the absolute maximum and minimum values possible.

The set is ordered into a distribution of values. Functional forms of probability distributions are commonly used to represent the probability per unit interval of values. These are **probability distribution functions**, f(x), and they are equations in terms of the estimate, called a **random variable**, x. (See chapter 11 for a more detailed discussion of both of these terms.) The Gaussian curve or bell-shaped curve are common names for the normal probability distribution function.

The questions for eliciting the probabilities associated with each value in the set can be asked in different ways: *What is the probability that an event will occur?* (direct value); *given a probability, what is the value or lower of the variable in question?* (cumulative probability); *what are the chances in n trials of an event occurring on this log scale?* (log-odds); and *what are the chances in n trials of an event occurring?* (odds ratio). For example, the estimate for the probability of a pipe rupture in a specified sequence of events would be given by the following:

Pipe Break Estimate	Probability of That Estimate or Less
0.001	0.01
0.005	0.05
0.010	0.10
0.05	0.20
0.10	0.50
0.15	0.70
0.20	0.90
0.25	0.95
0.27	0.99

Advantages: *Commonly used in decision analysis. In fact, one advantage in using a probability-based response mode is the existence of established elicitation and analysis techniques. (For this reason, we recommend the use of decision analysts and decision analysis techniques if this response mode is chosen.) In general, this is a very convenient form for modeling and analysis.*

Disadvantages: *Most experts are not good estimators of probability values and may be reluctant to use this response mode. (The use of probability wheels and training in probability estimation can mitigate these problems.) In addition, the concepts of probability distribution may not be fully understood by the experts and training may be required. Done correctly, probability estimation is a very time-consuming process. It can fatigue the expert.*

Summary of Studies: *It has generally been shown that humans are biased in their estimation of probabilities (Tversky and Kahneman 1974); that they do not properly interpret probabilistic phenomena, like randomness, statistical independence and sampling variability (Hogarth 1980); and that most do not feel capable of using this response mode (Spetzler and Stael von Holstein 1975, Welbank 1983). Hogarth (1980:149), who has done extensive studies on cognitive and motivational biases, acknowledges that "probability theory itself is difficult to learn and apply" but argues that it is the best choice for expressing*

uncertainties. He cites the numerical precision of probabilities and their logic for structuring relationships between events as reasons for using probabilities. Welbank (1983:28) questions whether the precision of probabilities is needed, at least in expert systems, but states that probabilities may be useful where the knowledge is vague and where there is need for some weighting of the answers.

Continuous scales

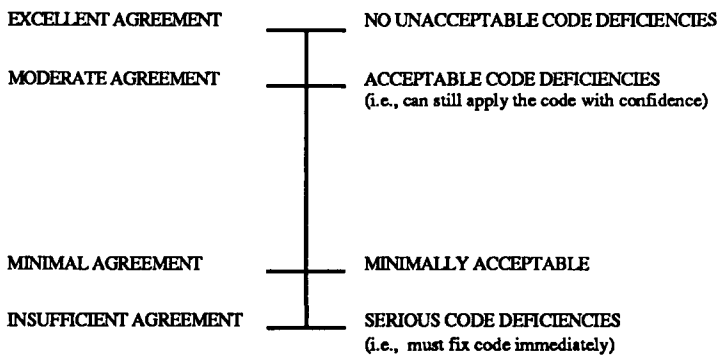
Continuous scales have continuous number lines either with linear or log spacing of values. The end points of the scale should represent extreme values and be labeled with text or numbers. Thus, the scale could be labeled with integers, textual categories, probabilities, odds, categories, ratings, or measurements of physical quantities such as temperature. The expert can mark his answer at or between any of the delineations on the scale. We recommend that the labels on the scale be clearly defined, especially if they are not measurements of some physical quantity. Frequently, categories or rating require additional clarification, such as given in the linear scale in the example 7.2. This scale was used by the experts to compare the data gathered from an experiment to the results generated from a computer code that simulated the experiment. The experts used the scale to rate (1) the agreement between these two sources of results, and (2) the performance of the code in capturing the experimentally generated reactor phenomena (Meyer and Booker 1987b).

Advantages: *Requires little instruction in how to use. Easily converted to numerical, continuous variables for analysis. Most people seem to be reliable estimators when using these scales.*

Disadvantages: *Developing a continuous linear scale to fit a particular application requires time. Care must be taken to guard against biased wording of either the labels or of the definitions of these labels. Training may be needed if log scales are used.*

Summary of Studies: *Seaver and Stillwell (1983) compared pairwise comparisons, rankings or ratings, and continuous linear scales. The continuous linear scale (labeled by probabilities) was ranked by them on the basis of their experience as being an empirically tested means of estimating probabilities and as requiring little preparation for analysis.*

EXAMPLE 7.2: ***A Continuous Linear Scale***



■

Pairwise comparisons

Pairwise comparisons is a process of having experts rate a set of objects, events, or criteria by comparing only two at a time. Comparisons can be made in terms of importance, likelihood of occurrence, or possession of some characteristic (e.g., cost). If there is a set of n things, $n(n\text{-}1)/2$ comparisons are required to do all possible pairwise comparisons. Comparing element A to B is considered the reciprocal of comparing B to A. Thus, if A is considered more likely to occur than B, B is less likely to occur than A. Using Saaty's (1980) type of pairwise comparison, the expert might be asked *Which is more important, A or B?* and then *how much more important?* To answer the latter, the expert would use a scale of values, such as the one listed below that Saaty designed for pair comparisons (see chapter 11, figure 11.8 for the complete scale). A response of "3" would indicate that he considered A to be weakly more important than B.

Number	Description
3	A slight favoring of the first item over the second
5	A strong favoring of the first item over the second
7	A demonstrated dominance of the first over the second
9	An absolute affirmation of the first over the second

Advantages: *Most people are reliable estimators using pairwise comparisons, in part because they only have to consider two things at a time. Thus, they do not exceed the capabilities of their information processing, as described by Miller (1956). Some methods, such as Saaty's (1980) Analytic Hierarchical Process, offer means for verifying the mathematical consistency of the expert's estimates. After a brief introduction, experts find pairwise comparisons an easy method to use. Another advantage of pairwise comparisons is that they can provide a numerically based analysis for qualitative data.*

Disadvantages: *Time consuming to elicit all possible combinations. Pairwise comparisons provide only relative data relations. A baseline scale or value is needed to translate relative comparisons into an absolute relation (Comer et al. 1984).*

Summary of Studies: *There is a body of research showing that people make better relative, indirect judgments, such as with pairwise comparisons, than direct estimates. (See Stillwell, Seaver, and Schwartz 1982 for a review.) In a study by Seaver and Stillwell (1983:2-12), pairwise comparisons was ranked by the authors, on the basis of their experience, as being acceptable to experts, as producing a high quality of judgment, and as having a strong theoretical base. Another study examined the usefulness of paired comparisons and continuous linear scales (labeled by probabilities and matching odds ratio) for obtaining estimates of human reliability in reactor risk assessments (Comer et al. 1984). The study's conclusion was that the analyses did not dictate the selection one response mode over another, but practical considerations could lead to a preference. The paired comparison method used required more experts and more of the expert's time than the linear scale.*

Ranks or ratings

Ranks or ratings involve assigning numbers or descriptions to the objects, events, or values in question. They are listed together because the judgments that they require are thought to be based on the same underlying psychological model (Comer et al. 1984).

Ranks can be integer numbers in ascending or descending order or ordinal descriptions, such as good, neutral, and poor. For example, to select the questions that would be addressed later in the elicitation sessions, the experts could rank them in importance (e.g., a "1" for the most important to address and a "5" for the least important).

Ratings are usually numbers or choices from a given scale or set of choices, such as a scale from 1 to 10 or a multiple choice set. We recommend using both numbers and words to further describe the ranks or ratings if there is a chance that their labels will not mean the same thing to each expert. For example, the qualitative term *small chance* has been found to mean from 1% to 40% depending on the expert's interpretation (Keeney and von Winterfeldt 1989). To foster consistency in experts' interpretation of the ranks or ratings, try to use both numerical and qualitative descriptors. For example, weapon's planners gave one of the numbers below (example 7.3) to rate how potential weapons related to a U.S. defense need (Meyer et al. 1982). The textual descriptors on the right were provided as part of the scale to prevent inconsistent use of the numerical values.

EXAMPLE 7.3: ***A Rating Scale***

3 ----------Completely related, approximately 80% related
2 ----------Significantly related
1 ----------Slightly related, approximately less than 20%
0 ----------Not at all related

■

The possibility of having experts make different interpretations of the rating was the reason that Sherman Kent developed the rating scale (example 7.4) below for government use.

Advantages:	*Experts find ranks and ratings easy to use, with little instruction. Ranks and ratings are good for applications with qualitative information or with only a limited set of possible answers. Many analysis techniques are available for rank data (Conover 1971).*
Disadvantages:	*People have difficulty keeping more than 7 (± 2) things in their minds at once (Miller 1956); therefore, comparing and ranking many items is difficult and inaccurate. Comparing more than 7 things requires either the use of more experts or the use of more time with fewer experts (Seaver and Stillwell 1983).*

EXAMPLE 7.4: ***Sherman Kent Rating Scale***

Order Of Likelihood	Synonyms	Chances In 10	Percent
Nearly Certain	Virtually (almost) certain We are convinced Highly probable Highly likely	9 8	99 80
Probable	Likely We believe We estimate Chances are good It is probable	7 6	60
Even Chance	Chances slightly better than even Chances about even Chances slightly less than even	5 4	40
Improbable	Probably not Unlikely We believe not	3 2	20
Nearly Impossible	Almost impossible Only a slight chance Highly doubtful	1	10

■

Summary of Studies: *Seaver and Stillwell (1983:2-12) compared pairwise comparisons, rankings or ratings, and continuous linear scales. On the basis of their experience, they evaluated ranks or ratings as being relatively easy to collect and acceptable to the experts. Also they considered the use of ranks or ratings to have sound theoretical justification.*

Bayesian updating

Bayesian updating is a process of revising estimates by combining different sources of information. Bayesian updating can be done by combining measured data with expert judgment data, by combining data previously supplied by one expert with that of the same expert at a different time, or by combining expert judgment from different experts. The expert judgment can be elicited using any of the previously mentioned response modes. This technique is also a means of aggregation because it is used to combine data sources. (Some background on Bayesian methods is given in chapter 11, and some applications are

found in chapters 16 on aggregating estimates and in chapter 17 on handling uncertainties.) For example, suppose a component was being tested for failure and 0 failures were found in 10 tests. Imagine also that an expert provided an estimate of 1 failure in 100. Using a binomial process for the data and assuming a beta prior distribution for the expert's data, Bayes updating would combine the test data and the expert's estimate as

$$\frac{0+1}{10+100} = 0.009 \quad .$$

Advantages: *Provides a convenient way to combine various information sources and accounts for the conditional nature of the data.*

Disadvantages: *Requires assumptions about the distributions of each source of data. May also require additional estimates by the experts for the parameters of the assumed distributions.*

Selecting from Dispersion Measures

The expert gives a dispersion measure when he is asked to provide some measurement of the amount of variation or uncertainty in the data, such as the error bars on experimental measurements. In addition, the expert's answer itself can be the datum on which a dispersion measure is requested. For instance, the expert could be asked to provide the absolute maximum and minimum possible value on his estimate of a physical quantity or a probability. The dispersion measures covered below are ranges, percentiles, and variances or standard deviations.

Ranges

Range is the difference between two values that represent a likely interval where the estimated value lies. Usually the expert is asked to provide an absolute maximum and absolute minimum possible value. The expert may be asked to provide two values that represent his version of what is likely. Error bars and uncertainty ranges are examples of these. For example, an expert estimates the probability of an event as 0.001. He is then asked to estimate a minimum and maximum on that event. He gives the minimum as 0.0001 and the maximum as 0.005.

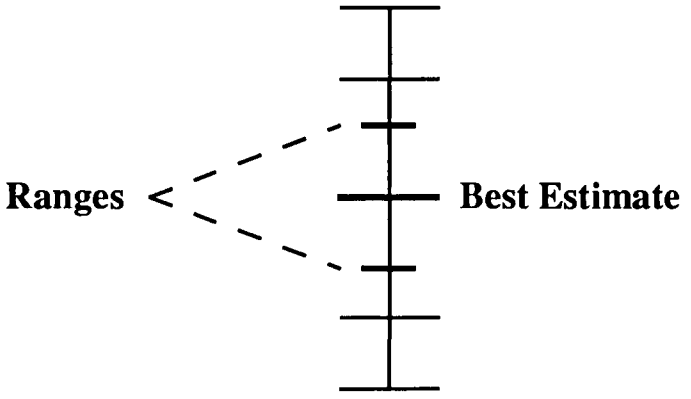
Advantages: *Ranges are easily elicited. Ranges are an acceptable measure of dispersion for analytical purposes.*

Disadvantages: *Humans are not very good estimators of absolute maxima and minima. They tend to underestimate maxima and overestimate minima. Ranges may not be sufficiently defined to analyze. For example, if the expert is asked to give a range and supplies two numbers, the analyst will not know what interpretation to place on these values. Similarly, if the expert is requested to mark specified ranges, such as ± 0.10, the analyst cannot assume that the expert's values represent ± 0.10 because people are known to generally underestimate such intervals.*

Volunteered ranges

Volunteered ranges are the same as ranges except that the experts are not asked to provide any ranges of dispersions during the elicitation. For example, in a project evaluating the performance of a computer code (Meyer and Booker 1987b), the experts used a linear scale, marked their best estimate with a tick mark, and then voluntarily made two marks on either side:

EXAMPLE 7.5: ***Volunteered Ranges on A Best Estimate***



■

Advantages: *Can be used to represent additional values of those elicited. Thus, the volunteered ranges can be used to fill out the sample of expert estimates. The values indicate the expert's uncertainty in his original estimates. Obtaining volunteered ranges requires no special efforts, except perhaps to use a response mode like continuous scales that encourages the experts to mark their ranges.*

Disadvantages: *It is difficult to interpret the meaning of the ranges and therefore difficult to analyze.*

Percentiles

Percentiles are where the expert estimates are specified values of a distribution such that the distribution is cut into the predesignated pieces. The expert may be asked to provide an estimate such that 5% of the values are smaller than that estimate, or such that 95% of the values are larger than that estimate. Confidence interval estimation is the same as estimating two different percentiles. For example, the expert estimates the failure probability for an event as 0.001. He is asked to give an estimate of the probability of failure such that there is a 5% chance or less or that the probability will be less than the estimate. He estimates this 5th percentile to be 0.0001. He is then asked to give an estimate of the probability of failure such that there is a 95% chance or more that the probability will be greater than the estimate. He estimates this 95th percentile to be 0.01. This interval of (0.0001, 0.01) is the expert's estimate for a 90% coverage interval.

Advantages: *Convenient for many analyses and modeling techniques.*
Disadvantages: *Humans are not very good estimators of specified percentiles. Training may be needed for the expert to understand what a percentile is.*

Variances, standard deviations

Variances, or **standard deviations**, are statistics estimated by the expert. The variance measures the average squared deviations of all possible values of the estimates from the arithmetic mean. The standard deviation is the square root of the variance. For example, the expert could give a best estimate for the probability of an event as 0.001. He might estimate that the variance about this value is 0.000001.

Advantages: *Convenient for many analyses and modeling techniques. Standard deviations are easier to estimate than variances because they are for the same order of magnitude as the estimates themselves; whereas, the variances are squared quantities.*
Disadvantages: *Humans are not good estimators of variances or standard deviations. Experts must be trained in the concepts of variance or standard deviation to use this response mode.*

Selecting from Methods for Eliciting Problem-Solving Processes

Three very simple means of eliciting the expert's problem-solving data are the verbal protocol, the verbal probe, and the ethnographic technique. While these are not the only methods for obtaining problem-solving data, they are three of the easiest and least prone to introducing bias. For additional information on methods, see McGraw and Harbison-Briggs (1989), LaFrance (1988), and Spradley (1979)

Verbal protocol

Verbal protocol involves instructing the expert to think aloud as he progresses through the problem (Ericsson and Simon 1980 and 1984). For example, the expert is given a written copy of the problem as follows:

> What feed program would you start this colt on? The colt is 6 months, 550 lbs., has an average metabolism, and will receive light exercise through ponying. Please solve this problem as you do others that you receive in this field. Please try to think aloud as you work your way through the problem. Your thinking aloud is as important to me as the answer that you reach.

The expert's verbal protocol resembles someone talking to himself or herself. This technique is from psychology.

Advantages: *Avoids introducing motivational bias (altering the expert's reports of his thinking) because the interviewer does not question the expert and thus is less likely to "lead" the expert (Meyer et al. 1989).*
Disadvantages: *Must be used only on one expert at a time. Not suited to group situations or those where the expert and interviewer are*

communicating by telephone or mail. Very time consuming. It usually takes the expert at least twice the time to verbalize his thoughts as it does to simply have them. The expert may not be able to verbalize all of his thoughts, as is a problem with all elicitation methods. These thoughts may be unconscious or they may be difficult to articulate because they must be translated from one form into another, such as from a mental picture into words.

Summary of Studies: *A main concern with this elicitation method has been that the process of verbalizing may negatively influence (bias) the expert's problem-solving. Ericsson and Simon (1980) , the authorities on this method, argue that the introduction of bias depends on the presence of the data in short-term memory and the type of information that the expert is asked to verbalize (McGraw and Harbison-Briggs 1989:335). If the information is in the expert's short-term memory, thinking about it should not alter it. Similarly, if the expert is to simply report his considerations and steps, less bias is expected to occur than if the expert were to conceptualize about his reasons for these steps.*

Verbal probe

Verbal probe is questioning done at a particular time and in a specific manner. The type of verbal probe discussed here is used immediately after the expert has reached a solution. The probe focuses on only one problem, the one that the expert has just solved, and is indirectly phrased so as to minimize influencing the expert's thinking. For example, immediately after the expert has solved the problem and given the answer, the verbal probe is used to learn why that answer was given.

Interviewer--Why did you give that answer--that feed program?

Expert--Well, it provides the right amount of protein, calcium, and phosphorus for a horse to grow at this age.

Advantages: *Quick means of obtaining general data on the expert's reasoning in solving the problem. Can be used on experts individually or in a group (Meyer et al. 1989).*

Disadvantages: *The verbal probe is best used where the expert can respond verbally in a face-to-face situation. Written responses to the probe are generally inadequate. The verbal probe is slightly more likely to induce motivational bias than the verbal protocol because the probe's questioning can cause the expert to invent descriptions of this thinking (e.g., plausible reasoning).*

Ethnographic technique

Ethnographic technique involves transposing the expert's words into questions. For example, the ethnographic method could be used to probe on one of the expert's responses to obtain an operational definition that could then be entered into the knowledge base. The expert has just said that the colt's feed program may need to be adjusted if the colt is not keeping his weight on.

Interviewer--Not keeping his weight on?

Expert--Yes, not gaining as he should at this time.

Interviewer--At this time?

Expert--At his age, 6 months, he should be gaining between 1.5 and 2.0 lbs. per month.

Advantages: *Can be used to obtain the greatest amount of detail on the expert's problem-solving processes. The ethnographic technique is a relatively nonbiasing form of questioning (Meyer et al. 1989) because it is based on the expert's own words. In addition, this technique can act as a check on the interviewer's or knowledge engineer's tendency to assume that she knows the meaning of the expert's terms (misinterpretation bias).*

Disadvantages: *Generally time consuming, except when used to elicit a few definitions. The ethnographic technique is not suited, in its usual time-consuming form, to group elicitations. Also this technique should not be administered while the expert is still solving the problem because it can distract him.*

Selecting the Type of Aggregation

Frequently there will be a need in the project to obtain a single answer from multiple and differing expert answers. There are two basic ways of obtaining a single response: (1) have the experts work until they reach a single consensus response, or (2) have the experts' answers mathematically aggregated into one answer.

Behavioral aggregation

Behavioral aggregation relies on the experts reaching a consensus (Seaver 1976). The aggregation occurs during, rather than after, the elicitation session. In interactive group situations, the experts are informed by the group moderator that they must reach a consensus and usually use persuasion and compromise to do so. In the Delphi method, successive iterations are used to reach one datum. (How to set up for behavioral aggregation is discussed in chapter 8.)

Advantages: *Produces an aggregated result during the session. Behavioral aggregation protects anonymity because no individual can be linked to the consensus response. Encourages the experts to support the product of their labors and view it as a group effort (e.g., perhaps from the thinking of "if we do not hang together, we will all surely hang separately").*

Disadvantages: *Needs advance planning because it should be used in conjunction with particular elicitation situations and measures for countering bias. In particular, it can foster a group-think situation where no one truly thinks but simply unconsciously acquiesces. Behavioral aggregation can be very time consuming if group think is not facilitating unconscious agreement. Employing behavioral aggregation can suppress expressions of difference and thus the chances of discovering the "right" answer. This means of aggregation obscures the differences between the experts' answers and the reasons for the differences, both of which can be critical to the understanding, analysis, and use of this data.*

Mathematical aggregation

Mathematical aggregation is the use of mathematical means to combine multiple expert's data into a single estimate or single distribution of estimates. Some mathematical methods weight the experts' data equally, such as the mean; others weight the experts' data differently in attempts to give higher weights to the "more expert" or more valued data (These methods are not included in chapter 8 but in chapter 16 because they can be considered after the elicitation.)

Advantages: *Does not have to be planned as early or as closely in conjunction with the elicitation methods as the behavioral aggregation. (However, the choice of the response mode may limit which aggregation methods can be applied.) Also, different mathematical schemes can be applied in succession to the individual's data, whereas with the behavioral aggregation the process can usually only be done once.*

Disadvantages: *Like any type of aggregation, it obscures the differences between the experts' answers and the reasons for the differences. It is easy to do mathematical aggregation incorrectly, such as by combining the estimates of experts who have made such different assumptions in answering the question that they have essentially solved different questions. Then too, mathematical aggregation can lead to the creation of a single answer that all of the experts would reject. For example, if the experts had given different distributions for describing a physical process, the aggregation of these distributions might describe a phenomena that could not occur physically. (For further discussion of the difficulties of mathematical aggregation , see chapter 16.)*

Summary of Studies: *The problem of how to aggregate expert's estimates has received much attention recently (Morris 1986, Clemen 1986, Schervish 1986, French 1986, Winkler 1986). Most schemes require that the experts' estimates be independent, even though experts' judgments are not known to be (Booker and Meyer 1988a:135). The few exceptions are those discussed by Winkler (1981) and Lindley and Singpurwalla (1984) but the latter two schemes assume that the structure of the correlation in the expert's estimates is somehow known or estimable (Booker and Meyer 1988a:136). Comparisons of different aggregation schemes indicated that the equal weighting of experts' estimates performs the best in covering the right answer (Martz et al. 1985).*

Selecting From Methods for Documentation

There are two major options for recording expert judgments: (1) recording the expert's answers, and (2) recording both the expert's answers and any information on the expert's problem-solving processes, including the expert's background. The second option can be done to any level of detail. We describe three levels, or documentation schemes, for the second option: summary documentation, detailed verbatim documentation, and detailed structured documentation.

Answer-only documentation

Answer-only documentation is a written record of only the answers or solutions.

Advantages: *Is the quickest and easiest method of documentation.*

Answer-only documentation

Answer-only documentation is a written record of only the answers or solutions.

Advantages: *Is the quickest and easiest method of documentation.*

Disadvantages: *The details of how answers were reached cannot be reconstructed later, if this becomes necessary. These details, then, cannot be critically reviewed, a seeming blessing but at closer inspection, a true disadvantage (i.e., what cannot be reviewed cannot be improved).*

Answer plus problem-solving documentation

Answer plus problem-solving documentation is written records of the experts' answers and how they arrived at these answers They can vary in their degree of detail. Three levels of detail are listed in *Advantages* below.

Advantages: *Allows for the defense of the judgments or the processes of elicitation. Provides the data for revising or updating the judgments. Provides the data for conducting detailed analyses on which factors (e.g., of the expert's problem solving or background) may correlate to the answer.*

Disadvantages: *On the other hand, the data is documented and can be criticized by reviewers. We believe that it is better to document how the experts arrived at their answers and receive possible criticism of specifics than it is to not document this information and to be criticized for inscrutability. There seems to be a growing trend among reviewers to lambast studies for missing or confusing documentation (e.g., the early drafts of the NUREG-1150 reactor risk study).*

SUMMARY DOCUMENTATION

Summary documentation is when experts or project personnel provide a few sentences or paragraphs on their thinking (e.g., references used, major assumptions, and reasoning). Sometimes this type of documentation is used to annotate the expert's answers or to compare them to each other or to some other baseline.

One version of summary documentation gives a few sentences of explanation for each answer. Experts from Army offices rated the importance of particular factors to the export decisions. For example, the importance of the factor "Triggers US Conventional Arms Transfer Restrictions" was given the high rating of "3" because "these restrictions are set by Congress and are outside of the Department of Army's control" (Meyer and Johnson 1985:50).

Advantages: *Less labor intensive or time consuming than the other means of documenting problem solving.*

Disadvantages: *Generally does not provide enough detail for tracing an expert's thinking.*

DETAILED VERBATIM DOCUMENTATION

Detailed verbatim documentation involves transcribing, usually from an audio or video recording, the expert's elicitation session. This method is more common to knowledge acquisition applications in artificial intelligence.

Advantages: *Requires minimal advance planning of what should be recorded and how.*

Disadvantages: *Transcribing tape recordings is very time consuming and labor intensive. Frequently, verbatim documentation provides too much and too undifferentiated data to be useful in tracing the expert's thoughts. We have also noticed that project personnel often prefer to recontact the expert to clarify some question rather than search through tapes or transcribed records of that expert's session.*

DETAILED STRUCTURED DOCUMENTATION

Detailed structured documentation usually involves providing the person tasked with documentation with a format of what is to be recorded. The format lists those aspects deemed to be the most important (e.g., answers and uncertainty levels, assumptions, and rules-of-thumb) and to the level of detail desired. For example, in the large reactor risk study NUREG-1150 (U.S. NRC 1989), the experts and project personnel recorded the following:

1. The issue name (e.g., the Temperature Induced PWR Hot Leg Failure)
2. The sources of information (e.g., analysis of results of running computer codes RELAPs/SCDAP)
3. Subissues into which the issue was divided (e.g., likelihood of hot leg circulation cell and ballooning occurring, ...)
4. Assumptions relating to the subissues (e.g., that hot leg nozzle and pipe and surge pipe are maintained as designed, ...)
5. Answers (e.g., plot of probability of failure, given temperature)

Advantages: *Provides the most traceable and analyzable problem-solving data.*

Disadvantages: *Requires the most planning and coordination. Very labor or time intensive.*

Common Difficulties--Their Signs and Solutions

Difficulty: The literature on the different methods is either scarce or conflicting. When the project personnel are selecting the components of their elicitation, they would like the literature to provide them with guidance. Specifically, they would like to be able to look up which method would be best for their particular application or situation. Unfortunately, there is little comparative literature on the methods, and the few sources that exist may provide conflicting advice.

There are several reasons for this state of the literature on elicitation components. One, it is very difficult to evaluate which methods provide the best results, if the result, the experts' answer to the question, is not something which can be externally verified, such as through measurement. Certainly this is a problem with evaluating the individual interview, Delphi, and interactive group methods, the response modes/dispersion measures and aggregation schemes. For this reason, elicitation components are more often compared on practical considerations such as how

acceptable they are to the experts or how much time they take to administer (e.g., Seaver and Stillwell 1983). Sometimes the results of one comparative study conflict with another. One reason for this difference may be that authors define methods differently. For example, the term *Delphi* has been used to refer to a variety of elicitation situations--some where the experts are kept physically separate, their data made anonymous and redistributed to the experts and others where the experts are together but restricted as to when they may interact. Then too, the authors may have different views about what is valuable in a method. For instance, response modes/dispersion measures can be compared in terms of how easy they are for the expert to use or how tractable they are computationally. While pairwise comparisons are easier for many experts to use than probabilities, they are not as easy to analyze. Depending on the importance that different authors attribute to these almost opposite considerations, different methods could be recommended for the same type of application.

Solution: To select the best elicitation components for a given situation, we recommend the following two steps. The first step is to review the relevant literature on methods. The references mentioned in the above text, particularly in the *Summary of Studies* paragraphs found throughout the section *Selecting from Within Each Component,* should provide a starting point. In addition, there are a few new attempts in the field of knowledge acquisition to compare and evaluate different elicitation techniques (see Dhaliwal and Benbasat 1989; Shaw and Woodward 1989).

The second step is to list the possible elicitation components and their strengths and weaknesses for the situation being considered. This list will serve as a decision aid in selecting the components of the elicitation for the application. The information on the method's strengths and weakness can be gleaned from the literature and also from the views of the project personnel. Try to use the definitions of the methods rather than the author's names for them as the basis of comparison. Points where the different sources of information truly differ, that is support the choice of different methods, should be noted and as much explanatory information should be provided as possible.

Later this list can be used to justify the selection of methods, such as in the final report for the project. We suggest that the report give the theoretical and logistical reasons for selecting the particular method. In addition, if the standard methods were tailored in any way, the documentation should explain why.

8

Designing and Tailoring the Elicitation

In the previous chapter the reader was guided through the selection of the building blocks or components of elicitation for gathering expert data. This chapter is designed to assist the reader in tailoring these components to a particular application. Five considerations are presented in the following sections to guide the tailoring of the elicitation: (1) *Logistics and Costs of Convening the Experts or Interviewing Them Separately*, (2) *Structuring the Elicitation Process*, (3) *Handling Bias--A Starting Point*, (4) *Documentation During and After the Elicitation Sessions*, and (5) *Presentation of the Question--A Quick Check.* There are so many possible combinations, given the components of the elicitation and these considerations, that arriving at the appropriate elicitation design requires detailed planning on paper. In general, people often neglect this planning phase, only to regret it later. Indeed, the design is not final even after following all the guides in this chapter. In chapter 9, *Practicing the Elicitation and Training of In-House Personnel,* a description is given of how to find any remaining glitches by testing the elicitation design.

Considerations in Designing the Elicitation

Frequently the elicitation design is driven by a combination of the considerations mentioned above and project-specific constraints, such as schedule. However, if this is not the case, the reader can selectively read those sections below pertaining to his or her areas of concern.

Logistics and Costs of Convening the Experts or Interviewing Them Separately

Even though the basic elicitation situation for extracting the expert's data has been selected (chapter 7), it can be tailored by adding or substituting other methods. Other elicitation situations and modes of communication (face to face, telephone, and mail) are frequently combined to create the quickest and least expensive means of elicitation. In particular, if the interactive group has been chosen as the basic elicitation situation, it could

be combined with less expensive situations, such as the Delphi and the individual interview. The expense of the interactive group situation lies in having the experts and project personnel meet together. It is usually cheaper for the interviewer to communicate by mail or telephone with the experts, as in the Delphi, or to meet with each expert separately, as in the individual interview. In addition, it is logistically difficult to schedule more than a few, usually busy, persons to meet for a week or for a series of shorter periods. Given the high expense and difficulty of gathering experts together, these meetings are often reserved for when they are absolutely necessary Other situations, such as the Delphi and individual interview, and modes of communication are substituted for the group meetings.

Frequently, other methods of eliciting besides the interactive group method are used for the first, second, and fourth stages of the larger elicitation process. The stages are as follows:

Stage 1. The selection of the question areas
Stage 2. The refining of the questions
Stage 3. The elicitation of the expert data
Stage 4. The documentation of the data

For example, in the NUREG-1150 reactor risk study (Ortiz et al. 1988), individual interviews had been selected for the eliciting of the expert data, stage 3, but other means were used for the rest of the stages. A preliminary set of questions were selected by the project staff and then sent, with background information, to the experts (stage 1). The experts were given several months to review these on their own and modify them, if they so chose. Then, the experts met together to be briefed on the use of the response modes and on techniques for refining the questions through decomposition. This refinement of the question (stage 2) continued as follows: the experts separately prepared their preliminary disaggregations; they met a second time to present their preliminary results so that they could benefit from the exchange of information; and they met as an interactive group for a third time to do a final refining of the questions and to discuss their problem-solving approaches. Their answers and aspects of their problem-solving process were elicited through individual interviews (stage 3). They were contacted singly after their elicitations to review the documentation of their data (stage 4).

The one exception to using different methods for the other stages occurs when the individual interview has been selected for eliciting in-depth, problem-solving data. Obtaining in-depth data in stage 3 must be preceded by working toward that level in the other stages. Thus, group situations (the interactive group) or ones in which the expert is physically isolated from the interviewer (the Delphi) cannot be used in the other stages because they are not suited to obtaining detailed data. For this reason, if the individual interview is being used for the 3rd stage, the other two situations cannot be used in the other stages.

Three modes of communication

Different modes of communication can also be employed in creating the optimal means of elicitation. The three modes of communication that can be used are face to face, by telephone, and by mail. Each has its advantages and disadvantages.

Face to face. The face-to-face mode is particularly adapted to obtaining detailed data. In fact, if eliciting deep problem-solving data is the goal, this is the only suitable mode of communication. Usually this type of information is elicited from one expert at a time through intensive interviews. For example, one project focusing on crisis managers required two 2-hour interviews with each of the experts. With projects whose goal is extracting the expert's problem-solving processes for building an expert system, the interviews are likely to be even longer. The latter can be a "prolonged series of intense, systematic interviews, usually extending over the period of many months" (Waterman 1986). The knowledge engineer usually travels to the expert's place of work and investigates the expert's problem solving *in situ.*

Telephone. One advantage of telephone communications is that it is generally less expensive than either face-to-face interviews or gathering the experts together. Another advantage of telephone communication is that it has a shorter turnaround time than the traditional postal method. The data can be obtained while the interviewer is still on the telephone, rather than later by mail.

On the other hand, the telephone is not good for relaying detailed or long pieces of information. For example, the telephone could be used to learn the expert's major reason for giving a particular answer or the main reference that he used. The telephone would not be suited to probing for the assumptions that the expert made in arriving at an answer. However, the telephone could be used for obtaining answers, such as to questions sent in the mail.

Traditionally, the mail survey has been used in combination with the Delphi, but the telephone could also be used, if only limited bits of information were being communicated. For example, the experts could give their responses (answers, a sentence or two on their reasoning, and/or the names of the references that they used) over the phone. The coordinator could make this information anonymous and relay it to the other experts. If there were many experts and thus magnitudes of data to be relayed, the above-mentioned information could be sent by mail. After the experts had received the mailed information they could be interviewed by telephone for their revised responses.

Mail. Traditionally the mail survey has been conducted by post but electronic mail is beginning to be used for this purpose. The traditional mail survey is good for eliciting simple data from a large sample just as the individual interview is suited to obtaining more detailed data from a small sample. Like telephone elicitations, mail surveys are much cheaper than face-to-face interviews or meetings of the experts.

The mail survey has the greatest problem of all of the modes of communication in its response rate, probably because it is easier to ignore a mailed request than to refuse one made in person or over the phone. In addition, recording one's thoughts and mailing them requires more effort than verbally reporting them. Again, the added effort of writing one's response leads many to abandon the attempt. For this reason, a mail survey is not recommended if it is critical to receive a response from a high proportion (i.e., over half) of the targeted population. The response rate can be boosted by calling the sample to request their response, but these attempts usually generate only half again as many as that of the earlier rate. (Note that electronic mail surveys may have higher responses rates than those sent by post. In addition, electronic mail surveys may have faster turnaround times than the traditional surveys.)

Neither the mail nor the telephone are suited for the transmission of complex instructions or detailed problem-solving data. For this reason, complicated response modes that require training should not be used with either of these modes of communication. Similarly, interviews for eliciting in-depth problem-solving data should not be conducted by mail or telephone. In particular, the verbal protocol, or thinking aloud method, should never be used in mail or telephone communications with an expert.

In the final method decision, the reasons for gathering the experts together will need to be balanced against the possible costs. These reasons and costs are listed below to aid you in making this decision. (For additional information on the relative costs and speed of these modes of communication, see Armstrong (1981:104-108,122).

Reasons for gathering the experts together

The reasons for gathering the experts together are as follows:

1. **If complex response modes and dispersion measures, such as probability distributions or percentiles, have been selected for use with more than a few experts.** The experts will require training in these complex modes and measures, and if there are more than a few experts, it is convenient to give them their training at the same time. In addition, a less complex response mode, Saaty's pairwise comparison, should also be introduced in an interactive setting. Experts sometimes experience initial confusion in using this mode (e.g., *Does a three indicate that A is weakly more important than B, or the reverse?*). For this reason, they need the extra clarification that being in a face-to-face situation offers.
2. **If the synergistic effect of group discussion is necessary to the elicitation process.** For example, if the experts in this field have not previously or recently gathered, if the field is evolving, or if the questions require knowledge of the field's state of the art, then meeting together as a group would be advisable. In addition, if it is critical that the group of experts identify with the outcome of their problem solving (i.e., see it positively as the product of their labors), meeting as a group is recommended. Member satisfaction with the product is higher in interactive groups than in Delphi ones, according to Seaver (1976).
3. **If the experts will not be able to give their uninterrupted attention to the project because of other demands on their time.** Having the experts meet in a place of your designation allows you to place controls on other's demands on the expert's time. For instance, controls can be placed on when and how the expert's telephone messages are delivered to the meeting room. Occasionally, we have had the experts themselves request that the meetings be held away from where they could be easily reached by their offices so that they could focus on the project (Meyer and Johnson 1985). We believe that gathering the experts together is especially helpful if their management has not given your project top priority.

Expenses for gathering the experts together

Possible expenses for gathering the experts together are as follows:

1. **Payment for the experts' travel and lodging,** if they do not reside in the same geographical area.
2. **Payment for the meeting room and any refreshments** (e.g., coffee to help keep the experts awake).
3. **Payment for having the sessions videotaped.** This record is not only useful for documenting the data but for allowing the experts who have missed a session to catch up. For example, in the NUREG-1150 reactor risk study, the experts gave presentations to the group on issues related to the technical question. These presentations and their subsequent discussions were taped. These tapes can then sent to the experts who were unable to make the presentation meetings so that they will receive the same information as the other experts.
4. **Payment for miscellaneous administrative costs,** such as for typing, copying, and mailing any background material that the experts need to see before coming together as a group. Copying may need to be done throughout the meetings. For example, after the experts have met, often they have materials that they wish to share with the other experts. In addition, toward the end of the process, the copying machine can be used to provide copies of the project personnels' documentation of the experts' data for the experts' review.
5. **Payment for the expert's time,** if this was part of the project budget.

Structuring the Elicitation Process

Structuring means imposing controls on the elicitation process. Structuring with respect to presenting the expert with a clear and assimilable statement of the question was discussed earlier in chapter 5. When the concept of structuring is applied to the larger elicitation, it can include using a predesigned set of questions to guide the elicitation, allowing only particular kinds of communication between the experts and requiring that the experts answer using one of the response modes.

Why structuring is done

Structuring the elicitation is done for a purpose, such as aiding the interviewer in interviewing, making the elicitation easier for the expert, or limiting the introduction of bias. For example, the structuring may be imposed by the use of a specially designed interview instrument--one that prompts the interviewer to ask particular questions at specific times (e.g., as described in Meyer 1987). The structuring may be put on the experts' group interactions to prevent some experts from acquiescing, without thinking, to other experts, as occurs in group think bias. For example, the Delphi was designed to counter biases arising from group interactions and thus is structured in this manner.

In general, we have observed that structuring the elicitation limits the intrusion of extraneous factors, such as bias. It seems to keep the field of observation clearer and thus eases the task of gathering and analyzing the expert data.

Degrees of structuring

Structuring can be done to varying degrees to different aspects of the elicitation process. As a general rule, use of one of the following components imposes structure: the response mode and dispersion measure, a method for eliciting problem-solving processes, the use of behavioral aggregation, or a documentation scheme. Using one of these components means that there is a plan in place and a specific procedure that will be used. For example, if the pairwise comparison response mode is used, the experts' judgments will be elicited by the interviewer or moderator asking particular questions and by the experts responding with answers on the appropriate scale. The response mode or dispersion measure, methods for eliciting problem solving, and behavioral aggregation are typically used when the expert data is being extracted (stage 3). The documentation component is often used in the last stage, 4, to record the data or the data-gathering methods.

Examples of structuring applied to each of the stages of elicitation are given below. The more that these options are applied, the more highly structured the elicitation situation becomes.

Structuring options applied to the stages of elicitation

Stage 1. Selecting the questions. For most elicitation situations, any one of the three structuring options could be applied to selection of the questions.

- Input could be obtained from the advisory experts on what would be good questions to pose to the external experts. This option is used frequently when the project personnel are unfamiliar with the field or if the project funder has not requested specific questions.
- The project personnel or the external experts can rank and select the questions according to some criteria. For example, on the reactor risk study NUREG-1150, (U.S. NRC 1989), the questions were initially selected by the project personnel according to which held the greatest potential to produce uncertainty in risk. The experts then reviewed the proposed set of questions and added, deleted, or modified these with respect to their own criteria.
- Alternatively, the external experts could be polled to learn, in advance, which problems they consider themselves qualified to address or which they would like to work on.

Stage 2. Refining of the questions.

- In interactive group situations, the experts and/or project personnel can work together in disaggregating the question. The experts or project personnel can present to the other experts and to the group moderator their interpretation of the question, means of disaggregating it, or any other relevant information.
- In any face-to-face situation, the experts can be required to review the definitions, the assumptions, or any other information that defines the question and then to refine the question for the last time.
- In individual interviews, the external or advisory expert could work on refining the questions that will be asked of him or other experts. For example, the expert can assist the interviewer in determining how to word the question and how to define particular variables.

- In a Delphi situation or a mail survey, the external experts can be sent information on the question and asked to provide information on how to set it up. For example, in a seismic study (Bernreuter et al. 1985) the experts were asked to break the Eastern United States into various earthquake zones in preparation for defining the question. It is also possible to have qualified project personnel, such as advisory experts, do most of the work in refining the question and then have the external experts review and modify it. For example, in the NUREG-1150 reactor risk study, the project personnel created sample disaggregations of the question, and the experts had the options of using these disaggregations as the starting points for their own.

Stage 3. Eliciting the expert data:

- During elicitations in interactive groups, the external experts can record their own data on a documentation format. The experts can also be asked to verbalize their judgments or thinking to the group and/or to the group moderator. For instance, the experts could state their names and their answers in the desired response mode. Only some of the experts need give data on particular questions; namely, those who were earlier judged (by themselves or the project staff) to be the most qualified. On the export control project, all the experts voiced their opinions in the early discussions, but only those experts who were assigned to particular questions were allowed to *vote* on the answers (Meyer and Johnson 1985). If a group of experts are verbally giving their responses, one expert can present his response while the rest are asked to remain silent. The natural and official leaders in the group can be asked to give their responses last or privately, if group think is a concern. The verbal probe or ethnographic technique can be used briefly to question the experts on aspects of their problem solving.
- In the Delphi method, the experts can be sent questions and asked for their data (e.g., their answers, a few lines of explanation for each answer, and a footnote of the references used). This data can be made anonymous and redistributed to the experts to allow them to revise their earlier answers. This process can be repeated as long as necessary, until consensus (behavioral aggregation) is achieved. If the experts are to give their answers in a particular response mode, they can be sent a format, such as a copy of a continuous linear scale, and instructions on how to use it.
- In individual interviews, the questioning can be guided by a list of topics that the interviewer has prepared. Alternatively, the interviewer could be guided by the format on which the data is to be recorded. Or, the interviewer can simply let the use of one of the pre-existing methods for eliciting problem-solving data, (e.g., verbal protocol, verbal probe, or ethnographic technique) guide her gathering of the data.

Stage 4. Documenting the data and/or the process by which it was obtained:

- In interactive group situations, the experts can use one of the documentation schemes to guide them in filling out their own records of their answers or

problem-solving data. Similarly, the group moderator can use one of these to guide her in gathering and recording all the required data.

- In a Delphi, the documentation format, along with instructions, can be sent to the experts.
- In individual interviews, the expert or interviewer can use the documentation scheme to record the desired information. A computer program that prompts the expert for his inputs and/or reasoning would serve this same purpose as a written documentation scheme (Meyer 1987).

General rules in structuring the elicitation

As a general rule, the more structuring that is imposed on the elicitation, the greater the time needed to plan and conduct it. However, these greater demands are balanced by the greater effectiveness of the structured techniques. For example, unstructured individual interviews are described as being less effective than their structured counterparts, at least in knowledge acquisition (Hoffman 1987, McGraw and Harbison-Briggs 1989:73).

A higher degree of structuring also seems to correspond to the gathering of more detailed data. It makes sense that a structured approach allows the interviewer to focus the questioning to a finer level. In an unstructured interview, there is less to prevent the expert from jumping from one topic or level to another, often to the interviewer's dismay.

We would also emphasize that conducting a structured elicitation often demands more tact than administering an unstructured one. This is because the experts do not always follow the structuring. They may be confused or ignorant about what they are supposed to do, or they may simply decide that following instructions takes too much effort. Tact is needed in getting the experts to follow the plan because the interviewer is dependent upon their good will for obtaining data. For example, imagine that the interactive group method has been structured to minimize the occurrence of group think bias, otherwise known as the follow-the-leader effect. The experts have been asked to present their views to the group before entering into a discussion of all the presentations Additionally, the natural leader of this group has been requested to go last in stating his or her views. However, that which was requested is *not* what happens. The leader interrupts and criticizes others' presentations. Diplomacy is needed to make this session run again according to the plan. One way of handling this problem would be to take the leader aside and tell him or her about the group think bias and how it negatively affects people's thinking. In addition, the experts could be briefed again, as a group, on the elicitation procedure and the reasons for structuring it in this manner.

Handling Bias--A Starting Point

Designing elicitation with regards to bias, as we advocate, is a new approach. The information presented here was largely developed from our own experiences. For this reason, it is intended to be a starting point for those who wish, as we did, to become more aware of bias and how to handle it. There is a small but growing body of literature on human biases that can be applied to handling bias in expert judgment--Payne 1951; Hogarth 1980; Tversky and Kahneman 1974 and 1981; Kahneman and Tversky 1982; Cleaves 1986; Meyer et al. 1989; and Meyer and Booker 1989).

There are two reasons for paying attention to bias: First, the occurrence of some biases has been shown to degrade the quality of the results (Hogarth 1980, Kahneman and Tversky 1982); and concerns about its presence, such as among reviewers of a project, affect the project's credibility.

There are two views or definitions of bias, as mentioned in chapter 3.

The first view of bias, sometimes termed **motivational bias**, proposes that *bias occurs when the expert's reports of his thoughts or answers are altered by the elicitation process*. (For an explanation of why this altering occurs, see chapter 3, *Causes of Bias--Motivational bias*.) For example, if the expert gave a different answer from what he believed because of the interviewer's comments, this would be considered bias. Using this first view, if the expert's estimates or problem-solving data do not represent the expert's thinking, these are not quality data. In addition, reviewers or clients are liable to question the validity of this data if they suspect that the experts have been *led* by the interviewer. In our experience, reviewers have been most sensitive to biases occurring because the interviewee was led either by the interviewer or by other members of the group to give an answer other than the one in which the interviewee believed. Also, many people are familiar with bias arising from a conflict of interest, such as when the expert's wishes or interests influence his judgment. (This bias is called *wishful thinking* in this book). These biases need not actually occur to impair the credibility of the work; they need only to be suspected of having occurred.

The second view of bias, sometimes termed **cognitive bias**, *defines bias as occurring when the expert's estimates do not follow normative, statistical, or logical rules.* Cognitive bias is frequently determined by checking the expert's data against the mathematical and statistical standards that apply to the response mode that he has been asked to use. To illustrate, if an expert gives probability estimates on all outcomes to a problem (previously defined as being mutually exclusive), and these probabilities do not sum to zero, these data would be considered biased because they do not follow normative statistical rules. While the majority of people are not yet as aware of cognitive bias as they are of motivational bias, this situation is not likely to continue. Such authors as Hogarth, Kahneman, and Tversky are leading the field in showing the cognitive limitations to which the human mind is prone. The old view that the brain acts like a computer (i.e., in being mathematically correct) is rapidly being debunked.

The approach proposed in the section below is to anticipate which biases are likely to occur given the planned elicitation and then to tailor the elicitation methods accordingly. This step, *Anticipate the biases,* is followed by others in the later chapters: *Make the Experts Aware of the Potential for Introducing Bias and Familiarize Them with the Elicitation Procedures--Step 2* (chapter 10); *Monitor the Elicitation for the Occurrence of Bias--Step 3* (chapter 10); and *Adjust, in Real Time to Counter the Occurrence of Particular Biases*--Step 4 (chapter 10). *Analyze the Data for the Occurrence of Particular Biases--Step 5,* is discussed generally in chapter 14. In chapter 3 in *Determining Which Steps to Apply,* we describe how the reader can determine which of the above-mentioned steps to use.

How to anticipate bias is described below.

Anticipate the biases to which the planned elicitation is prone and redesign the elicitation, as needed--Step 1

Applying this step presupposes that the reader has selected a motivational or cognitive view of bias, as described in *The Selection of the View of Bas*---chapter 3, subheading *Determining Which Steps to Apply*. While we consider the cognitive and motivational views to be equally valid definitions of bias, we believe that one way of construing bias may be more useful than another for a particular project. We suggest that the reader select and use only one view of bias at a time to avoid being contradictory. For example, use of the cognitive definition would propose that a mathematically incorrect judgment be modified. This act would cause a misrepresentation of the expert's data, a bias, according to the motivational definition of bias.

Selected biases and situations in which they occur

To anticipate some of the biases to which an elicitation method is prone, determine the components of elicitation and/or modes of communication that you intend to use and look them up in the table *Index of Selected Biases* in chapter 3--*Steps in a Program for Handling Bias*. The index will list some of the biases to which particular situations are prone. To obtain more information on the motivational and cognitive biases, locate the item in the section following the table *Definitions of Selected Biases* (chapter 3). It should be noted that the biases listed in the *Index* and *Definitions* are not the only ones that can occur in gathering expert data. However, they are the biases that we have frequently encountered and have developed means for handling. These sections are meant only to give the reader a start in dealing with bias.

The other option for accessing information in this section is to go directly to the segment *Definitions of Selected Biases* and skim it to learn if the planned elicitation is likely to be susceptible to one of the selected biases. The elicitation situations and components that are prone to these biases have been underlined in this segment to allow them to be found more easily. This segment also provides information on *why* the elicitation is prone to a particular bias.

The information on why the elicitation is prone to bias can provide the reader with ideas for redesigning the elicitation, if the reader wishes to do so now. For instance, under wishful thinking bias the following information is provided in chapter 3: "that its effects will be most pronounced when the expert does not have to explain this reasoning." Thus, if wishful thinking bias was a concern, this information could be instrumental in modifying the existing elicitation to require that the experts provide explanations of why they gave their answers.

Documentation During and After the Elicitation Sessions

How the expert data will be recorded is one of the considerations guiding the tailoring of the elicitation. The other considerations mentioned so far have been the logistics and cost of having the experts meet together; the structuring the elicitation; and the treatment of bias. Before we can present the options for who performs the documentation and when, we need to describe what can be documented.

What documentation can include

1. **The statement of the question, in its final version.** This statement would include (1) any background information that clarifies the question, such as scenarios, and (2) definitions or assumptions that the experts are to use.

   Stating the question is not as easy as it sounds because often the statement of the question is changing until the actual beginning of the elicitation, or in some cases until the expert gives an answer. For example, in the NUREG-1150 reactor risk study, the experts were developing their own disaggregations of the questions, essentially their statements of the question, until the moment that they gave their judgments.

2. **The identity of the experts.** The names of the experts may need to be recorded in order to answer questions about the expert's responses or to update their responses after the elicitation. For example, if the experts are to review the project personnel's documentation of their responses, records must be kept of each expert's response, name, and date of elicitation. In addition, records of the expert's identities and responses will be needed for report writing. However, even when all of this information is recorded, it need not be divulged in a report. There are three ways of presenting the expert's identities depending on the level of confidentiality that has been selected. (In chapter 6, see item 6 *Will the Judgments Be Anonymous...* under *Motivating the Experts Through Communication of the Intrinsic Aspects of the Study.)* These three ways are to either list (1) the organizations or offices that have contributed experts; (2) the experts names and affiliations; or (3) the expert's names and affiliations in connection with their responses. The last of these is the most demanding in terms of the records that will have to be kept.

   The identity of the experts can also refer to aspects of their professional background, such as how long they worked in their area of expertise and where they went to school. We recommend recording and testing this information to learn if it correlates to the expert's answers or other factors. In our experience, people intuitively expect the experts' answers to correlate to some aspect of their background, such as where they went to school. Although we have found no such evidence of correlation (Booker and Meyer 1988a, Meyer and Booker 1987b), we suggest documenting this information for testing.

3. **The methods by which the expert data was obtained.** The elicitation methods need to be documented if any of the project personnel intend to use them again, if the project is likely to be reviewed by outsiders, or if a written report is a required product of the study. In general, if the methods are to be documented in a report, we suggest that two types of information be supplied: (1) a summary of what methods were used and how they were applied; and (2) an explanation of why each method or combination of these were selected. Support for the use of the methods can include references from the literature and other considerations such the data-gathering objectives of the project and the need to reduce costs, time, or the occurrence of particular biases. This support can also include descriptions of how the methods were pilot tested or rehearsed and revised. In addition, the supporting information can include statements on what the documentation of the expert judgment represents. Readers of the

report may not know that the documentation represents the expert's state of knowledge at the time of the elicitation. For this reason, it may be necessary to state that the expert might, with time and new knowledge, give a different answer to this same question.

4. **The expert's responses.** The expert's responses can be documented using several schemes. The documentation schemes, described in chapter 7, are basically of two types: where only the expert's answers are recorded and where both the expert's answers and thoughts in arriving at these answers are recorded. There are approximately three ways of documenting the second type. The first is a summary documentation where the expert's answer and a few sentences or paragraphs on his thinking is provided. The next is detailed verbatim documentation where everything that the expert says or does is written. Often this scheme involves mechanically recording the elicitation session and then transcribing it. The last and most involved means of documentation is the detailed structured scheme. In this method, there is usually a format to guide the person doing the documentation. The format includes blanks labeled for the recording of the aspects of the elicitation that are considered most important, such as the sources of information (e.g., code simulations, references, experiments, personal communications), assumptions, algorithms, and equations that the expert used in solving the problem.

The documentation can be used in a report in its original form or it can be rewritten to be more general; that is, the report can include less or more general data than was gathered but not the reverse. For example, if the expert's names were not originally recorded with their responses, this data will not be available for inclusion in a report.

Logistics of who will record and when

The documentation of the expert judgment can be written by the project personnel, by the external experts, or by a combination of the two. For example, the judgments can be written by the data gatherer and reviewed for accuracy and completeness by the expert. In addition, there is sometimes an option for when particular types of information are documented--before, during, or after the elicitation session.

Step 1: The first consideration in determining how to do the documentation is the decision of who is qualified and motivated to do it.

This information is listed below according to what is being documented.

1. **The final statement of the question.** The project personnel, such as the data gatherers (interviewers and knowledge engineers) are usually the most qualified and motivated persons to document this information. The one exception would be if the external experts had been totally responsible for the refining of the question. Then, the experts would be the most qualified, although, not necessarily the most motivated, to record this information. (The experts are generally not as concerned with documenting the elicitation as they are with solving the problem.) For this reason, even when the experts are the most qualified to document the statement of the question, the project personnel

often do it. The project staff may request the latest statement of the question from the experts and then record it.

2. **The expert's identity.** Writing the expert's names by their responses does not demand as much qualification and motivation as recording the types of information mentioned, described in items (1), (3), and (4). Noting the expert's identity can be done by whoever is recording the expert judgment. If the experts are recording their own judgments on a special format, they can simply enter their names on the predesignated blank. If the data gatherers are using formats (writing the expert's responses in an organized manner on the board or using interview guides), they can label the responses by the expert's name. If the data gatherers are mechanically recording individual interview sessions, the tape can be labeled in the same fashion. In general, the labeling should be double checked because the documenter can forget to record a name or record the wrong one. Occasionally the data gatherer records characteristics of the expert, in addition to the expert's names and responses. These characteristics are gathered when later analysis is planned, such as for learning if the characteristics correlate to the expert's data. Some commonly recorded attributes are the expert's education, years of experience, job title or position, area of specialization, and psychological test results.
3. **The elicitation method.** We consider the project personnel, specifically those who conduct the elicitation, to be the most qualified and willing of persons to record information on the elicitation methods. Information on the elicitation methods is the only one of the four types of information under *What Documentation Can Include* that can be written both before and after the elicitation. The elicitation methodology can be initially written up after its last revision (chapter 9). Then after the elicitations, the description of the methods can be updated because they may have been conducted slightly differently than planned.
4. **Expert's responses.** Who is qualified to record the expert responses depends on how detailed or technical this data is. Almost anyone is sufficiently qualified to write down a short answer (e.g., a probability estimate, rating, or ranking). Frequently, the project personnel record this information if it is given verbally in an interactive setting (e.g., group or individual interviews). If the elicitation is conducted over the telephone, the interviewer records the answer. On simple mail surveys, the expert records the answer, using the desired response mode.

   If, however, the expert data includes detailed problem-solving data, the expert is the most qualified to document it. Unfortunately, experts are not usually as well motivated as the project personnel to document this information. Generally, they are more willing to deliver the data verbally than they are to take the time to record it in written form. Thus, documenting the problem-solving data needs to be made as easy as possible to increase its chances of being done. Documentation formats (also called *interview instruments* or *forms*) are often used for this purpose. (See sample forms included at the end of this chapter.) The data gatherer or the expert can be guided by these formats in documenting the information. Please note that if the experts fill out the formats, the

completed formats should be checked by the project personnel before the experts leave. Experts frequently fail to document their thinking as they were instructed to do. They may have misunderstood the instructions or simply tired of writing their data. Similarly, if the project personnel did the documentation, the expert should check the notes for accuracy and completeness. For example in the NUREG-1150 reactor risk study, an extra precaution was taken--the experts signed the final documentation to show that they had approved it. Then too, if the elicitation is tape recorded, someone should check that the expert's voice is audible and understandable, especially if this will be the only record of the session.

If the data gatherer is unfamiliar with the field and if complex questions are being asked of the experts, the data gatherer may no longer be the most qualified to document. Arrangements beyond those mentioned above may be needed. For example, in the NUREG-1150 reactor risk study (U.S. NRC 1989), the elicitations were performed by decision analysts who were experienced in the elicitation but not in the technical areas. For this reason, project personnel who were knowledgeable in the question areas attended the sessions and, like the decision analysts, provided written records. In addition, all sessions were tape recorded and the expert was encouraged to document his reasoning for his judgments immediately after the elicitation (Wheeler, Hora, Cramond, and Unwin 1989, 3-7).

Step 2: The second consideration is whether the persons tentatively selected for the documentation role will be able to do it.

The selected person may be too busy eliciting the expert data or solving the question to do the documentation. A list of tasks that people can do simultaneously is given below to help you make this assessment.

Experts

- Experts can solve the problem and verbally deliver or write down the answer
  IF it is a short answer
  IF it is given in a manner (response mode) that the expert is used to
- Most experts can solve the problem and verbally describe their problem solving. A few have difficulty doing these two tasks simultaneously, as noted in chapter 7 in the description of the verbal protocol.
- Most experts cannot solve the problem and write down their thought processes in solving it
  IF their thought processes are to be recorded in any detail

Interviewers or knowledge engineers:

Working with a single expert

- The interviewer is able to request the expert's answer, record it on a documentation format, and run the tape recorder.

- In addition, to the above tasks, the interviewer can usually also document the expert's problem solving
  IF the interviewer is familiar with the subject area

Working with multiple experts face-to-face

- The interviewer can request the expert's answer and record it on a documentation format
  IF obtaining the expert's answer does not require in-depth elicitations
  IF the experts will be orderly in giving their answers, as in a structured elicitation situation
  IF the interviewer is not also running mechanical recording devices
- In addition to the above tasks, the interviewer can usually document the experts' problem solving
  IF only one or two sentences are needed (e.g., on the references that the experts used or the group's rationale for their answer)
  IF the interviewer is familiar with the subject area

Working in any elicitation situation

- Interviewers cannot write all of an expert's problem solving as the expert is verbalizing it
  IF this is being elicited in great detail (because people usually verbalize their thoughts much faster than the average person can write. Peoples' rate of speech is one of the reasons that tape recorders and stenographers are used as backup to notetaking).

Note: More is involved in running a tape recorder than there might seem. In all sessions, someone needs to turn the tape recorder on, check that the tape head is turning, and change the tape cassette as needed. In group situations, the omnidirectional microphone needs to be turned on and have its batteries checked. In individual interviews, the expert's lapel microphone needs to be attached and detached so that the expert does not walk away with it, tape recorder dangling.

Step 3: If the documenter is asked to do more than what was listed as feasible in step 2, consider the three following aspects of documentation. It may be that manipulating one of more of these will lighten the documenter's load:

1. **Who does the documentation or how many others assist in it?** For example, if the data gatherer is to do the documentation, could the expert help? Could additional project personnel (e.g., a stenographer, secretary, or editor familiar with the technical area) assist the data gatherer, such as by taking notes or running the tape recorder during the session?
2. **The timing of when the documentation is done.** It may be that some of the documentation can be done in advance of, or after, the elicitations. Frequently, the statement of the question, the elicitation methods, and the

expert's identities can be documented in preliminary form before the elicitation. Then, during the elicitation, brief notes can be taken on any changes that were made. For example, the statement of the question often evolves or changes during the process or elicitation. Similarly, the list of experts assigned to the questions may change if some of the experts do not come and substitutes are used.

- ***The leeway in timing the documentation of the expert's problem solving.*** The problem-solving data cannot be recorded in total before the elicitation because it often changes significantly during the session. However, you may be able to document some of the basics of the expert's problem solving if it has been written prior to the elicitation. For example, in the NUREG-1150 reactor risk study (U.S. NRC 1989) the experts disaggregated their questions before the sessions in which their answers were elicited. However, even in this project, the experts changed their disaggregations and thoughts during the elicitation. Documenting the problem-solving data totally after the elicitation does not work either. Waiting to document this data results in portions of it being forgotten and lost. Tape recording the problem-solving data, although helpful, is not a panacea because much of what people say is unclear without their expressions or gestures. (We recommend that tape or video recording be used only as a backup to note taking because the mechanical recordings often malfunction and they are very time consuming to play back, e.g., to clarify particular points or to transcribe.). Thus, some documentation of problem solving must be done during the elicitation if this data is to be recovered and reported later.

- ***The amount of time needed to verify the expert's data.*** Generally, more time is needed to verify expert data if it was documented after the elicitation rather than during. For example, if the experts' problem solving was documented after they had left, they will have to be contacted to verify what has been written. Therefore, verifying documentation long after the session is more time consuming than verifying it during or shortly thereafter by having each expert read and initial it while he is still present.

3. **The level of documentation.** If none of the suggestions mentioned above resolve the documentation problem, simplifying the documentation demands may be the last resort. Would a less detailed or structured documentation scheme still provide the necessary data? The general documentation schemes listed in chapter 7 are listed in order of the simplest, or least demanding, to the most demanding.

Sample documentation formats

Formats serve as guides for the project person or the expert who is taking down the information. For instance, formats have been developed for questioning the expert on the subject area (Spradley 1979); for obtaining background on the expert (McGraw and Harbison-Briggs 1989, Meyer 1987, Meyer and Booker 1987b); and for obtaining short explanations of the expert's problem solving (Meyer 1987, Meyer, Booker, Cullingford

and Peaslee 1982). Formats that have been designed specifically for knowledge acquisition projects are given in McGraw and Harbison-Briggs (1989).

Two generic formats for recording the expert's answers and their problem solving are illustrated at the end of this chapter. The first focuses on documenting experts' ratings and only gathers a few notes on why the expert gave these ratings. The second focuses more on the expert's problem solving. It documents the experts' disaggregation of the problem, their answers, and their reasons for giving both of these.

Presentation of the Question--A Quick Check

Although this topic was covered in chapter 5, we readdress it at this time to ensure that the presentation of the question fits the revised elicitation. In chapter 5, the question was refined by considering the information that the experts needed to answer it (e.g., the background, assumptions, and definitions), the order in which the information needed to be presented (e.g., general to specific or specific to inclusive), the decomposition of the question, and the phrasing of it. Now these same aspects of presenting the question need to be reconsidered in terms of the revised elicitation.

In general, you need to note the changes that you have made to the elicitation as a result of following the suggestions in this chapter and ask whether the presentation of the question still fits. A few considerations follow:

1. **If the components of elicitation have been structured, is the presentation of the questions structured to a comparable degree?** For example, if the elicitation components were structured to provide a clearer field for observation, an unstructured presentation of the question could compromise this aim. An example of an unstructured presentation would be to ask a question without having determined which wording was clearest and what information would be needed to answer it.
2. **If the basic elicitation situation was modified to include a combination of the other situations, will the planned presentation need changing?** For example, imagine that you decided to break an interactive group situation into individual interviews to avoid the possibility of group think during the elicitation. If you had planned to present the question (i.e., its background, definitions, and assumptions) only once to the group, you would have to rethink this presentation for the individual. The question could be announced to the group before individual interviews began but if there was much of a delay before interviewing the last expert, the question would have to be presented again to refresh the expert's memory. In a case like this, the question could be presented on paper and then verbally reviewed by the interviewer at the onset of each elicitation.
3. **If the telephone or mail was to be used as an alternate mode of communication, will the means by which the question was to be presented need changing?** For example, if the plan was to save on costs and time by eliciting the expert's solutions over the telephone, can a simple and brief presentation of the question be given over the telephone? If not, perhaps the question can be sent by mail and then the experts called for their answers.

4. **Does the documentation scheme support the presentation of the problem and vice versa?** For example, if a detailed structured documentation scheme was selected, the presentation of the question should exhibit comparable care.

Any remaining conflicts between the presentation of the question and the elicitation methods will become known when the pilot testing is done. The pilot testing described in chapter 9 will be the last stage in refining the elicitation process before the elicitation is conducted.

Common Difficulties--Their Signs and Solutions

Difficulty: ***Realization that sufficient time has not been allotted to planning the elicitation.*** This is one of the most common problems that we have observed. Often people view the planning stage as an unnecessary delay to starting the elicitation. Signs of insufficient time for planning usually come when the project personnel are looking at their project schedule. The schedule may show that the elicitation, not its planning, is to begin *now*.

Solution: One remedy is to determine if the planning can be dropped without severe consequences later. Planning is less necessary, if only a few experts are being interviewed, especially face to face. For this situation, you do not have to coordinate a group of interacting experts nor polish the elicitation materials that you would send to another group of experts (e.g., a Delphi group). You have greater license to plan as you go because there are fewer things requiring advance preparation. Planning is also less necessary, if unstructured techniques are used. As mentioned earlier in the chapter, the more highly structured techniques require more planning. We believe that the structured techniques offer, in compensation for their greater planning time, greater effectiveness in elicitation and analysis. For all other situations, planning is likely to be critical. If sufficient time for planning has not been allowed, we urge you to consider modifying the schedule.

Difficulty: ***Ignoring the possibility of bias.*** Frequently, the possible occurrence of bias is disregarded because of the project staff's ignorance. They may have been totally unaware of its existence, or they may have chosen to ignore it. Both of these are understandable responses. The topic of bias in expert judgment has not commonly been addressed and the physical scientist, in particular, is not accustomed to dealing with it.

However, ignorance is definitely not bliss. We have seen an increase in external reviewers' criticism of expert judgment studies. Generally, these reviewers have been sensitive to the bias that results from the interviewer's *leading* of the experts. For example, a committee that reviewed NUREG-1150's early use of expert judgment (Kouts et al. 1987) criticized this effort for having the experts work from questions proposed by the project personnel. They advocated allowing the experts to independently generate and refine their own questions. Another frequent concern with

external reviewers is the possibility of group think, perhaps because of the influence of Janis' book (1972). The concern is that some experts may have mouthed or tacitly accepted the opinions of others without thinking or expressing their own thoughts. Conflict of interest, called wishful thinking in this book, is another source of bias that seems to catch reviewers' attention. When this bias is present, the expert's judgments tend to reflect what the expert would like to happen, or in an extreme case, the position that the expert has been paid to support.

Solutions: One approach would be to use this chapter's section on anticipating bias to design the elicitation before proceeding. Another narrower approach is to selectively focus on handling the three biases mentioned above (social pressure from the interviewer, group think, and wishful thinking). Still another approach is to proceed with the idea of gathering enough data to test for the presence of selected biases after the elicitation. The analysis techniques for testing bias are described in chapter 14.

For example, on a magnetic fusion project (Meyer et al. 1982), two tests were run to examine the problem of wishful thinking. The purpose of the study was to obtain estimates from those working on the fusion project of whether they would meet scheduled milestones. After the elicitation, the sample of experts were divided into those with more to gain from their answer--the managers, and those with less--the scientists working on the hardware. Their data was analyzed. These two groups could not be found to give significantly different answers to the same questions. The experts answers were also compared to the wished-for outcome. That is, the expert's predictions of when they would meet various scheduled milestones were compared to the planned schedule and found to differ significantly.

The last approach, not considering bias until during the analysis, has one advantage over doing absolutely nothing about bias. If reviewers of the work raise questions about a particular source of bias, the results of the analysis can be presented. If the results show no evidence of the bias and the reviewers find no fault with the methods of analysis, the reviewers can turn their attention to other matters.

EXAMPLE 8.1: *Sample Format for Documenting Mainly Answers*

Expert's name:_____

Date:_____

Time:_____ to _____

Interviewer:_____

Definition of question:_____

_____

	Variable 1	*Variable 2*	*Variable 3*	*Variable 4*
Variable w				
Variable x				
Variable y				
Variable z				

Expert's comments on reasons for giving estimates:

Variable w and *1*:_____

Variable x and *2*:_____

Variable y and *3*:_____

Variable z and *4*:_____

■

EXAMPLE 8.2: *Sample Format for Documenting Answers and Problem Solving*

Expert's name:_____

Date:_____

Time:_____ to _____

Interviewer:_____

Question name:_____

Information defining the question: _____

_____

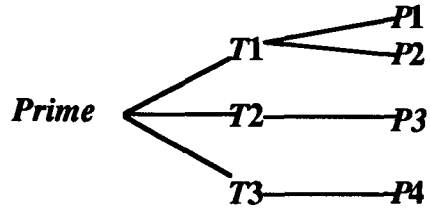
_____

_____

Sketch and label each question, subquestion, and branch, as shown in the example.
Provide a few sentences on the expert's reasoning for each point, as shown in the example.

Footnote any references that the expert uses.

Example



Reasoning:

Reasoning:

T1:_____

T2:_____

T3:_____

P1:_____

P2:_____

P3:_____

P4:_____

■

9

Practicing the Elicitation and Training the In-House Personnel

The purpose of this chapter is to provide the last check on all aspects of the elicitation design before conducting the elicitation, which is described in chapter 10. The logistics, ease of use, interface, and timing of the parts of the elicitation process are examined to find any remaining glitches and resolve them. The problems are identified by practicing and pilot testing the different parts of the elicitation.

Practicing the Elicitation

The following aspects of the elicitation are frequently practiced: the briefings given to the experts; the elicitation procedures; and the documentation, aggregation, and entry of the data into the model. Practice is defined loosely here to mean rehearsing some act to become more proficient in it. Practicing serves another purpose--the training of in-house personnel. One subset of practice is pilot testing. **Pilot testing** involves taking a very small sample of the expert population, presenting them with the aspect of the elicitation that is to be tested, obtaining these experts' feedback, and revising the elicitation accordingly.

What Needs to Be Practiced?

The following items should be rehearsed to resolve any difficulties and to learn how long they will take.

1. **Introduction of the experts to the elicitation process.** Introducing experts to the elicitation process includes familiarizing them with the general steps and schedule that will be followed, the questions or problems that they will address, the response mode that they are to use, and the biases to which they may be prone. If the experts were not heavily involved in developing the elicitation, they will need clear and concise briefings on what they are to do. Conducting such briefings do not occur naturally but requires practice. If the

expert is confused about any of the above-mentioned topics, his confusion could make him more prone to bias, such as inconsistency. It could also make him reluctant to participate. Practicing these items will make the communication of them clearer and will give the project personnel greater confidence.

Introducing the experts to the elicitation process is such a critical step in establishing the project's credibility and the experts' understanding of their tasks that we offer the following suggestions. (The most appropriate time to introduce the experts to the elicitation process is *after* they have been briefed on the general elicitation procedures.)

- We suggest that the experts be given sample questions to work so that they can practice using the response mode. If there are any techniques for properly using the response mode, they can be introduced and practiced here. For example, if the response mode is probability distributions, Hogarth (1980:149) offers eight keys to assigning probabilities.
- We recommend that the experts be briefed on the biases that were identified in chapter 8 as being likely to occur. This briefing should include an explanation of why the selected biases occur and of how the expert can reduce his tendency to commit them. (The section *Definitions of Selected Biases* in chapter 3 provides examples of this type of information.)

  The bias briefing should include exercises that are designed to evoke the selected biases. The interviewer can read the correct answers and allow the experts to score their own exercises. These exercises can convince the expert that he, like everyone else, is prone to these biases. If the briefing is given without exercises, we have noticed that the experts are not as effective in countering their tendencies toward bias, perhaps because they were never convinced that they, too, would be biased.

2. **The elicitation procedures.** Even when the elicitations have been carefully planned, much can be learned by testing them on sample experts. For example, you may observe that the elicitations last longer than expected and that the sample experts become fatigued during them.
3. **The documentation procedures.** Pilot testing the documentation will provide information on the logistics and the format, such as whether or how these need to be revised. For example, the pilot tests may show that the interviewer cannot conduct the elicitation and simultaneously fill in the documentation format.
4. **The mathematical aggregation of the expert's answers, if this will be used.** Although mathematically combining the expert's responses may seem straightforward, there are many methods available and some may not be appropriate to the data elicited (chapter 16). Therefore, it is advisable to practice the chosen method. At the very least one can learn how long this procedure is likely to take, and if necessary, automate it. The data from the limited pilot tests are used to practice the aggregation. In addition, if the experts have broken the problem into its logical parts and the project personnel will reaggregate the experts' responses, as in the NUREG-1150 reactor risk study (U.S. NRC 1989) this step should also be practiced.

5. **The entry of the data into a model, if this is planned for its analysis or end use.** The data gathered from the limited pilot tests, as described below, is used for rehearsing the entry of the data into a model. Frequently, this rehearsal reveals a major disconnect between the form in which the expert's data was obtained and that which is needed for the model.

What Needs to Be Pilot Tested?

The procedures that need to be pilot tested are those that the expert participates in, such as in giving his judgment. The expert's understanding is critical to his participation. Therefore, anything that the expert must understand, is a good candidate for pilot testing. Usually, however, a smaller set of all those things that the expert must understand are pilot tested because pilot testing is time and expert intensive. As a general rule, pilot testing is done on the expert's understanding of instructions whether these are given orally or in writing, such as on how to fill out a survey. The following parts of the elicitation are suggested for pilot testing.

1. **The expert's understanding of the problem or question.** Sometimes the statement of the question or problem has already been developed and is being presented to the experts for their solutions. In other projects, such as the NUREG-1150 reactor risk study, the experts are to evolve their own statement of the problem from a beginning version. In the first case, the pilot test provides information on the expert's interpretation of the problem. If the project personnel did most of the question development, their interpretation is likely to differ noticeably from that of the experts. In the second case, the pilot test checks that the expert understands that he is to take the basic problem area and refine it.
2. **The expert's use of the response mode.** The pilot test allows you to check that the expert understands how to use the mode. The expert's response can be checked against both the instructions and logical or mathematical standards. For example, probabilities for all mutually exclusive events should sum to one.
3. **The expert's understanding of the elicitation procedures that he will be expected to follow.** For example, if verbal protocol was to be used, it would be important to check that the expert knew that he was to *think aloud* and that he could do it.
4. **The expert's use of any documentation format, such as on a mail survey, if he, rather than someone on the project, is to fill it out.** The pilot test provides feedback on the expert's understanding of the directions on how to fill out the format. We have found that experts typically do not provide as thorough a documentation of their judgment as they are requested to do.

How to Pilot Test

Sample selection and sizes

In pilot testing, it is desirable to have experts who represent the range of experts who will later be elicited. For instance, if the experts have been drawn from positions in the government, private industry, and academia, the pilot sample should contain members of each of these positions. Consider the factors that were used in selecting the expert populations and use these in choosing the sample for pilot testing. In selecting an expert pilot sample, consider that this selection decreases the pool of experts whose judgments can be elicited later. Generally, it is not advisable to use the pilot test sample or those who have otherwise assisted in the method's development in the final elicitation. Similarly, we would caution you to avoid using the advisory experts, those who have helped develop the questions, in the pilot sample. The advisory experts will not be able to approach the test materials from a fresh perspective if they helped develop them.

The size of the test sample will depend on the size of the expert population for the elicitations. Test samples for expert elicitations rarely, if ever, have the large sample sizes (10% of the total) associated with traditional mail surveys. Typically, expert samples for pilot tests are five persons or less because the largest expert population does not usually exceed 50. Because of these small sample sizes, the strategy for pilot testing in expert judgment studies has been to test intensively. In other words, try to obtain as much benefit or feedback as possible from these few experts.

There are two types of pilot tests mentioned below. We recommend conducting the intensive pilot test first, if you intend to do this test. The **intensive pilot test** was developed (Meyer, 1986:92) to trace the expert's understanding. It consists of structured, in-depth interviews and observations. The other type of pilot test is called **limited** to distinguish it from the traditional (high sample size) and the intensive pilot tests mentioned above. Limited pilot tests are best done after intensive pilot tests have been performed and the elicitation revised. The limited pilot test allows the in-house personnel to practice the elicitation procedures, time their duration, and check how these procedures fit together.

Sequence for pilot testing

The sequence for performing these pilot tests follows:

1. **Apply part one and part two of the intensive pilot test (as described on the next page) to the elicitation.**
2. **Use the pilot test results to revise the tested parts.**
3. **Practice the entire elicitation process from beginning to end with limited pilot tests done on those parts where the experts will be involved.** For example, the in-house personnel would rehearse their introduction to the elicitation in front of the experts for the limited pilot tests. The in-house persons would practice the elicitation procedures on these experts. The expert data from the practice elicitations would be used to check the other steps of the elicitation. Specifically, the data would be documented, aggregated, and used as input just as it would be when the elicitation is done for real. Each of the elicitation procedures is timed during the practice.

How to conduct intensive pilot tests

This type of pilot test is the only one to our knowledge that allows people's thinking to be tracked through information presented in written form. The intensive pilot test provides two kinds of information: (1) how, in general, the expert progresses through the information, his general impressions, and when and why he decides to respond to particular questions; and (2) how the expert specifically interprets each direction, statement of the question, or response mode option.

First, the materials to be pilot tested are selected using the list provided above under *What Needs to Be Pilot Tested?* A typical selection would be the set of problems; any written directions on assumptions or definitions that the expert was to use; directions on what information the expert was to document on the form and in what manner, such as on a continuous linear scale.

Intensive pilot testing is done with one expert at a time in a face-to-face situation. The interviewer sits to the opposite side of the expert's handedness to allow easy viewing of his writing; that is, the interviewer sits to the expert's left, if the expert is right handed.

INTENSIVE PILOT TEST--PART 1. For the first part of the intensive pilot test, the expert is instructed to fill out the written materials as he would naturally if no observer were present. The expert is also asked to *think aloud* (verbal protocol) as he reads through the written materials. The interviewer has a copy of the same materials given to the expert for recording the data.

While the expert pages through the written materials, the order in which he looks at the materials, his pauses, gestures, facial expressions, and words are recorded by the interviewer on the interviewer's copy. For example, the expert may skim through the introduction, cover letter, or directions and then flip through the rest of the materials before returning to read each of these more thoroughly. The expert may make such comments as: "I have problems with this page and will probably let it sit on my desk for several days."

In addition, if the intensive pilot tests will not be followed by limited pilot tests, the interviewer can record the expert's starting and ending time on the first part of the intensive test. The limited pilot tests described in the next section provide better time estimates because the expert is not having to think aloud throughout the interview. However, some data on time is better than none, so this data should be gathered now, if they will not be gathered by other means. Note, if you are trying to obtain time data, you will have to save your questions of the expert until the second part of the intensive test. Otherwise, your questions will inflate the time estimates of how long the expert takes to respond to the written materials.

The intensive pilot test provides better time estimates than might be expected. It is reasonable to expect that the time estimates obtained from the first part of the intensive pilot test would be very high because the expert continuously verbalized his thoughts (and will not do so for the actual elicitation). However, we have found that this measurement provides an adequate estimate of the upper limit of time needed. This is because the experts selected for the pilot test may not represent the range of experts, some of whom could take a significantly longer time to finish. For instance, we have had pilot tests last two hours and the actual elicitations range from 1 hour to 2 hours and 15 minutes.

Frequently, the expert is allowed a brief break between the first and second part of the intensive pilot test.

INTENSIVE PILOT TEST--PART 2. During the second part of the intensive pilot test, the expert is asked to paraphrase in his own words, the meaning of each direction, question, or response option. This information allows the interviewer to track the expert's interpretation in detail. It has always amazed us that something could be interpreted in so many different ways, ways that we had not thought of because we *knew* what we meant The interviewer can also question the expert about any reactions, such as a look of puzzlement, noted at one of the questions during the first part of the pilot test. This questioning can jog the expert's memory so that he can give a more thorough description of his impression.

All information is recorded on the interviewer's copy of the written materials, and the expert's hard copy is collected. The expert is thanked and asked not to discuss the details of the pilot test with anyone else because it could influence their responses The data from the intensive pilot test is examined and used in revising the elicitation procedures or the logistics of who performs the different tasks. If limited pilot tests are planned, they are the next step.

How to conduct limited pilot tests

In limited pilot tests, the elicitation procedures are conducted as closely as possible to the way that they will be done for the actual elicitation. There would be little point in examining the expert's responses to a form or a procedure that did not resemble the one to be administered. For this reason the intensive tests are done first to allow the methods to be revised. Thus, in the limited pilot test, the experts are given the briefings or the introductions that have been planned for the larger population of experts. Similarly, the sample experts receive whatever forms and instructions on providing their responses that will be used during the actual elicitation sessions.

The interviewer elicits the data in the planned manner and obtains the experts responses. In addition, data is gathered on how long each elicitation lasts so that this can be added to other time estimates to produce a total. If the elicitations are being done in person, the interviewer may also record the amount of time it takes an expert to respond to a particular problem.

The following is an illustration of how the timing data is used. The practice briefing of the group of sample experts lasted 2 hours and one elicitation lasted 2 1/2 hours (counting the expert's review of the documentation of his responses). Assuming that there is only one elicitation team to perform 10 individual interviews, about 27 hours would be needed just to perform the elicitation.

As with the other pilot test, when the expert finishes, he is thanked for his cooperation. The data from the pilot test is used in practicing the other stages of elicitation, mathematical aggregation and modeling, as was mentioned in the *Sequence for pilot testing* above.

In addition, the limited pilot test may be used in training in-house personnel, as will be discussed below.

Training In-House Personnel

Project personnel are frequently lay persons in the area of expert judgment and require training in elicitation methods. Even those experienced in expert judgment may need training if they are unfamiliar with the specific methods selected. Training provides the personnel with instructions (directions) and the opportunity to practice until they become proficient.

Training is done in the same areas that are practiced.

1. Introduction of the experts to the elicitation process.
2. The elicitation procedures.
3. The documentation procedures
4. The mathematical aggregation of the expert's answers if this will be used.
5. The entry of the data into the model if modeling it in some way is the end use.

The training can be done to differing levels of proficiency depending on what the project demands and the personnel desire. For instance, if the experts were likely to be reluctant to give their judgments, we would recommend that the data gatherers be trained to a higher level of proficiency to obtain the judgments. In addition, there are particular situations that make training necessary.

When Training Is Needed

Training the in-house participants is especially necessary under the following conditions.

1. **When different people have planned the elicitation than will be conducting it.** In many projects, the elicitation methods have been designed by experienced persons or selected by a committee. However, the task of implementing the methods is often given to those who are less experienced and, perhaps, those who did not participate in the selection of methods.
2. **When the persons who will be performing the elicitation and documentation are uncomfortable with this assignment.** This reaction may stem from the situation mentioned in item 1. Others, such as their managers, selected the methods, and now they are being asked to implement the methods even though they have not had any input into the selection process. Then too, the persons who are to perform the elicitations may be initially reluctant to do so because of their inexperience. Training can make them feel more confident of their ability to perform these tasks.
3. **When more than one person or team will be gathering and documenting the expert data.** With more persons, there is the chance that each will perform the tasks differently and that data will be inconsistent. Training the data gatherers promotes consistency in these tasks. In addition, the process of providing instruction often allows the instructors to identify any looseness in the procedures and to provide stronger guidelines.
4. **When more than one in-house person will be involved in an expert's elicitation.** More than one person may be required to perform the elicitation if one cannot do all the tasks simultaneously. Then too, the elicitation

can require very different skills, such as in interviewing and in the technical subject area. Frequently, several people are used because there is not time to train one person in both skills (e.g., train the technical person to interview). Thus, one person may elicit one type of information from the expert and a second person another. For example, in the NUREG-1150 reactor risk study, a decision analysis method of elicitation was selected, and three decision analysts were brought in to do the elicitations (U.S. NRC 1989). However, the information being elicited on reactor phenomenology was so technical that an in-house person needed to participate in the elicitations to answer the expert's technical questions about the problem, to record, and to check the expert's technical data for gaps. In another project examining tank tactics (Meyer 1987), one person obtained the expert's reasoning as he responded to a computer-simulated battleground, and the other ran the computer. The computer person who was familiar with the computer program, ensured that the computer ran properly and answered any of the expert's questions on it. If a two-person interviewing situation exists, the in-house persons must rehearse their roles together. Otherwise, they can appear like a slapstick comedy routine--bumping into one another, interrupting and confusing each other and the expert.

How to Train

There are different means of training in-house personnel to execute the five items mentioned earlier in *What Needs to be Practiced.* The paraphrased items are listed below and followed by training suggestions.

1. **The briefing of the experts on the elicitation.** The in-house personnel can simply rehearse this talk in front of an audience, preferably one unfamiliar with the planned elicitation.
2 and 3. **The elicitation and any documentation done during it.** It is more difficult to train personnel in these two areas because the areas are more complex than those mentioned in 1, 4, and 5. Elicitation and documentation involve interactions with the expert and perhaps simultaneously with another data gatherer. The instructions that can be given do not encompass all that could happen in the elicitation. Therefore, these tasks must be rehearsed with someone playing each role as it will be done in the actual procedure.

Three training options and their advantages and disadvantages.

The first option is to instruct the trainees in the procedures and then have them conduct the limited pilot tests as practice. The advantage of the first training option is that it does not require a large pilot test sample nor much time.

The second option is to have the experienced data gatherers perform the first limited pilot tests while the trainees observe quietly and then have the trainees conduct the last tests. The advantage of the second option is that the trainees are able to learn by observing before doing. The disadvantage of this option is that it requires a large enough sample for more than a few tests.

The third option is to have the trainees observe the limited pilot tests or videos of them and to practice their skills by taking turns role playing the interviewer and the expert. The advantages of the third option are (1) that there is no special requirement for a particular number of pilot tests, and (2) that there are benefits derived from role playing the expert. The trainees who play the experts gain insight into the elicitation from the expert's perspective and become better interviewers. When they play the experts, they learn first hand how the interviewees wish to be treated. For example, the trainees as interviewees may view some of the procedures as condescending and dictatorial.

4 and 5. The mathematical aggregation and entry of the the data can be done simply. The trainees can be given the data from the limited pilot tests, instructed on what to do with it, allowed to perform and time these operations, and asked to report any problems.

Common Difficulties--Their Signs and Solutions

Difficulty: Pilot tests show that the sample experts have significant difficulty with the response mode. If there are problems in using the response mode, the intensive pilot tests will have indicated this. The sample experts may have paused on the section describing the response modes; they may have seemed confused; or they may have remarked that they did not understand. In addition, the answers may give further evidence of the expert's difficulties. For example, it may be that the response mode is to follow particular logical or statistical standards. If the sample expert's responses consistently violate these, there is a problem.

Solutions: One of the first approaches is to rethink the briefing on the response mode that was given to the experts. Perhaps, this briefing or the written instructions on the response mode were not clear. The revised briefing should be given to a new sample of experts, and the use of the response mode should be intensively pilot tested again.

If this sample of experts has the same amount of difficulties, you may wish to reconsider using this response mode. In selecting another response mode, consider what would fit with the other phases of the elicitation, such as the aggregation and modeling of the responses.

Difficulty: Pilot tests indicate that the elicitation is likely to need more time than is available. It is a common occurrence to find that your elicitation requires more time than you have allotted. This discovery is evidence that all of us are prone to wishful thinking--to overestimating how much we can do in a given amount of time and to underestimating uncertainty--forgetting those things that can take more time. Signs from the limited pilot test indicate that the interviews by themselves or perhaps in combination with the other stages, such as aggregation, sum to more time than was scheduled.

Solutions: A first option is to review the information from the limited pilot tests on how long the different stages take and decide which of these stages could be done differently

to cut time; redo the stage; pilot test the elicitation again; and record its new duration. A second option is to change the schedule. In our experience, people have often chosen the second option because it does not require the careful consideration that the first option does.

Difficulty: ***In-house personnel resist the training.*** Personnel can resist elicitation training for several reasons: they do not view elicitation as part of their job; they do not feel qualified to perform elicitation; they fear being blamed if the elicitation fails; and they resent having to do something that they did not help plan. If those receiving the training appear uncooperative or object to the training, you need to find out why.

Solutions: Any solution rests on learning the person's reasons for resistance and addressing those reasons. If only a few persons seem to be inspiring this reaction in the rest, addressing the concerns of the minority may resolve the problem.

Difficulty: ***The rehearsal shows that the documentation scheme does not meet the other needs of the project.*** The documentation can fail to mesh with the rest of the elicitation process if it is at a different level of detail (granularity) or form than needed for the later reporting, aggregating, inputing, or analyzing of the data. If the documentation scheme was not pilot tested, this difficulty is often not discovered until after the elicitations have been conducted.

Signs of the above situation will show up during the pilot testing as processes that do not fit or flow. For example, you may have tried to aggregate the expert's solutions mathematically while taking into account their differing assumptions so as not to *mix apples and oranges.* Imagine, however, that this data is missing or that it is in a fuzzy form and cannot be used for comparing the experts' solutions. Another example could be that the documented data cannot be input into the model. It is in the wrong form. The expert's solutions were given on a linear scale of 1 to 5 and the computer model requires probability distributions with confidence ranges. Similarly, you may have tried to analyze the effect of some variable on the expert's problem solving only to find that you did not have the necessary data. Perhaps, you used a summary documentation scheme and could not find any record of this variable.

Solutions: If this problem is caught before or as a result of pilot testing, the solution is simple. Select and develop another documentation scheme (chapters 7 and 8) that will address these needs. Perform limited pilot tests again and use the data gathered in them in testing the aggregation and the entry of the data into the analysis model.

If this problem is not detected until after the elicitations have been completed, the options are limited. Basically, the aggregation, report, or analysis cannot be done as planned. If the documentation was in the wrong form, perhaps, it can be translated into the desired one. However, in translating the data from one form to another, there is always the risk of misrepresenting it, of making an assumption that is not valid, such as concerning its distribution.

Difficulty: ***The documentation of the expert's problem solving has been done differently or to different levels.*** We have frequently encountered this difficulty and have observed many other situations in which it has occurred. Generally,

inconsistency in the documentation occurs when (1) multiple project people or experts have done the documentation, or (2) when the documentation includes questions of differing complexity.

For example, in the NUREG-1150 reactor risk study, some of the final documentation was done by the decision analysts who were performing the elicitation and some by the project staff who were familiar with the technical area of the expert's problem solving (U.S. NRC 1989). Given the differences in these documenter's backgrounds and the information that each was to record, they could not, in all likelihood, have produced the same documentation. In another instance, our documentation was inconsistent even though it was done by the same person using approximately the same documentation format. The expert's means of solving problems were gathered in detail--first on classical statistical questions and then on judging the adequacy of a computer code in modeling reactor phenomena (Meyer and Booker 1987b). While the statistical questions could not be called simple, they were simpler than the evaluations of the code. The experts seemed to think and verbalize very differently about solving the simpler versus the more complex questions. Thus, it was difficult to document the same level of detail on these two questions.

Solutions: If this problem has been detected during the practice elicitations, the rehearsals have served one of their purposes. There are several actions that can be employed to promote consistency in the documentation. First, the guidelines on the documentation can be tightened. Often, one of the reasons that people have done their documentations differently is that their instructions on the format have been open to differing interpretations. Second, additional training on documentation could be offered. If project personnel will be doing the documentation, they can rehearse it by doing limited pilot tests and taking turns at playing the experts. They could turn in their practice documentation for review and feedback on which aspects of their elicitation still needed to be changed. If the experts will be doing the documentation, the instructions and format given to them should be carefully pilot tested for clarity.

As a final check on documentation, a qualified person can be appointed to review each record as it becomes available after the real elicitations. The project person who will be aggregating, modeling, or programing the expert data or writing the report would be a natural choice for this task. The designated individual could be asked to check each record, preferably before the experts left, if the experts had been convened for the elicitation sessions. The appointed person could quickly check that all the needed information was there and that it was legible and comprehensible. This type of check is particularly necessary if the experts will be documenting their own solutions or problem solving.

If the documentation problem was not discovered until after the elicitations, which is when it is typically found, the options for resolving it are more limited.

The first option is to try to reconstruct the elicitations to fill in the gaps. This entails using all the records, both written and mechanically recorded, to find the missing information. If the information cannot be found, recontacting the experts may be necessary, perhaps even to have them rework some parts of the problem. This option is so tedious and embarrassing that it is rarely done.

The second option, is to adjust what and how the data is reported so there will be no gaps. Usually this means reporting or modeling the data at a more general level, a

coarser granularity. For example, if some of the notes on the expert's problem solving are more detailed than others, the less detailed will have to be used as the common denominator.

10

Conducting the Elicitation

This chapter provides details on how to schedule, set up, and conduct the elicitation. The different elicitation situations and the three techniques for eliciting problem-solving data are covered. In addition, guidance is given on how to monitor and adjust for bias during the elicitations.

In Part III, information is given on how to analyze the data collected from the elicitations.

Scheduling the Elicitations

There are several fine points in scheduling and confirming meetings that will make this phase go more smoothly. In general, these are simple courtesies that set the stage for good relations with the experts. The next two sections--*Scheduling the Meetings* and *Confirming the Scheduled Meetings*--may be skipped if there will not be meetings with the experts, such as if the mail survey, Delphi, or telephone interview are to be used.

Scheduling the Meetings

To schedule the meetings, such as for individual interviews or interactive group situations, call the expert and follow the steps given below:

- **Introduce yourself and your affiliation for this project.** For example, *Hello, Dr Jones, this is Mary Smith calling for the Division of Risk Analysis, Nuclear Regulatory Commission* (the name of the organization that is funding the study).
- **Ask if it convenient for the expert to talk for a short, specific amount of time now.** An example of how to ask this is, *Is it convenient for you to speak to me for about five minutes at this time?* We have all had the frustrating experience of answering the telephone when we are busy, having the caller speak nonstop, and not being able to interrupt to explain that we cannot talk at this time. Similarly, it can be irritating to the expert to be called when he is leaving for a meeting or holding one in his office, and his irritation may

emerge later (e.g., in his reluctance to be scheduled for an interview). For this reason, we recommend checking that the call is at a convenient time. The following illustrates such a check, *Have I called at a convenient time, or am I interrupting something?* If the expert answers in the negative, quickly ask when it would be a good time to call back.

- **State your intention--to schedule a meeting for interviewing the expert.** Briefly explain to the expert that the purpose of this call is to schedule an interview with him. In general, we have found that it is best to say *interviewing* rather than *obtaining answers*. The latter seems to lead many experts to think that they will be asked point blank for their answers. This imagined scenario seems to make them uncomfortable or leads them to doubt the validity of the caller's intentions (e.g., anyone who thinks that these answers can be given so easily must not be very knowledgeable).
- **Tell the expert about how long the session will last.** This time estimate will allow the expert to decide when he can block out an appointment of the necessary length. If pilot tests were performed (as described in chapter 9), estimates of the length of an elicitation will be available. We frequently say something like *the session is likely to last from one to two hours, depending on you.* (Later, if the expert complains about the amount of time that his session took, emphasize again that the duration depends on the expert, and then thank the expert for his *thoroughness*.)
- **Emphasize selecting a date and time that is at the expert's convenience.** We have found it effective to emphasize our wish to schedule the meeting at the expert's convenience. This courtesy initially promotes the expert's good will. If the meeting will include other experts, mention the times that the other experts have suggested and ask this expert which ones are the most convenient. (Continue this process until one preferred time and one alternate time are found that are acceptable to all the experts.) If the meetings will involve only one expert, ask the expert to pick a convenient time. For example, *Could you pick a convenient time for a two-hour meeting within the next two weeks?* Request that the appointment be in the next few days, few weeks, or months, depending on the deadline for completing the interviews.
- **After the expert has selected a time and date, verify this information by repeating it.** Repeating this information gives both the caller and the expert time to record the appointment and to catch any misunderstandings. Also repeat the location of the meeting, especially if it will be held somewhere other than the expert's office. If this is the case, either give or request directions to the meeting place.
- **State that the appointment will be confirmed.** For instance: *I will try to call on Thursday, the day before, to check that the meeting time is still convenient.*
- **Thank the expert.**

Confirming the Scheduled Meetings

- **Introduce the caller and state the intention of confirming the meeting.** For example: *Hello, this is Mary Smith calling about our meeting from 9:00 to 11:00 a.m., tomorrow, the 31st, in your office. Is this meeting time still convenient?* Confirm the meeting the day before or on a Friday, if the meeting is on a Monday. (If the experts will be traveling to the meeting, confirm the meeting a few days before their trip.) This simple reminder has saved us many frustrations and wasted trips.
- **If the meeting was to be with only one expert who then needs to cancel, reschedule the meeting. If the meeting involved other experts, offer to call the expert back after talking to the other participants.** The following things need to be considered in rescheduling a meeting: whether other experts or only this one will be unable to attend, whether the meeting could be video taped, whether there will be costs associated with cancelling (e.g., travel, lodging, or meeting room), and whether it is possible at this late date to contact the other experts to cancel the meeting. Offer to call the first expert back after talking to the other participants. If only one expert cannot attend, he could be shown a video tape of the session and his interview could be conducted later. If several experts cannot attend, consider using the alternate (backup) date.

Setting Up and Conducting the Elicitations

Tips on Setting Up for the Elicitations

Setting up for the elicitation means bringing the necessary papers and or supplies and physically arranging the meeting room. Many of these preparations are not critical to the success of the elicitations but make them easier or more pleasant. The preparations are listed below and can be used as a checklist or memory aid, if so desired.

How to set up for an individual interview

For an individual interview, we recommend that you assemble the following:

1. **The expert's name** because forgetting it in the midst of the elicitation can be embarrassing. It is also a good idea to take the address and telephone number of the meeting place if there is a chance of becoming lost or arriving late. Maps of the area are useful.
2. **A short letter on the project** that includes who is sponsoring it, who is conducting it , and what its product will be. (A longer version of this letter is described in *Motivating the Experts Through Communication of Intrinsic Aspects of the Study*, chapter 6). This information on the proper letterhead helps establish the interviewer's and the study's credentials and refresh the expert's memory on the project. In addition, we recommend stating this information to the expert, rather than just handing it to him for reading.

Conveying this information verbally will get the interview started and ensure that the expert has heard the necessary information at the beginning of the interview.

3. **The documentation format, list of question topics, or any form** that will be guiding the questioning. These can be labeled with the interview's date and the expert's name or identification number prior to the interview.
4. **Copies of the questions and/or background materials** (references, charts, tables, etc.) that the expert will be using. The expert will need one copy and the interviewer will need another in order to follow what the expert is viewing or commenting on. Again, the expert's identity can be recorded in advance on these papers.
5. **Extra pencils or papers** for note taking.
6. **The mechanical recording device, cassettes, extra batteries, and/or an extension cord.** The tape cassettes can be labeled in advance with the expert's name and the time and date of the elicitation. This labeling is particularly important if an expert will be interviewed more than once on the same question because his thinking is likely to change with time.

How to set up for a Delphi situation

Setting up for a Delphi is different than the other two situations because its communications will be by mail and/or by telephone. As mentioned in chapter 7, one of the greatest problems with the mail survey is that it has a low response rate. It is therefore advisable to put as much effort as possible into communications with the expert to increase the chances that he will respond.

If the experts will be receiving and returning the questions by mail, the following need to be prepared.

1. **The cover letter for the set of questions.** The cover letter contains an abbreviated version of the letter first sent to motivate the experts to participate. (See chapter 6, *Motivating the Experts Through Communication of the Intrinsic Aspects of the Study*) The cover letter should be carefully composed because it is one of the most important tools for encouraging the experts to answer.
2. **Copies of the set of questions.** These are assumed to include instructions on how to fill out the questions, explanations of the response mode, and directions on how the expert judgments should be returned. If the expert's data is to be returned by postal mail, we recommend self-addressed envelopes with stamps, if possible. Self-addressed envelopes have been found to increase response rates on mailed surveys.
3. **The names, addresses (electronic or regular mail), and telephone numbers of the experts.** If the list of addresses is likely to be outdated, current mailing addresses should be verified by calling the experts.
4. **Advance publicity to increase the response rate.** If the experts work in the same organization, a brief article can be inserted into the organizational news bulletin or a memo can be sent from the expert's management. If the experts are not all in the same place, call them before sending the set of questions. In general, we recommend calling the experts in addition to publicizing the study through the use of articles and memos. Frequently, so

much time will have elapsed between the time when the experts were selected and when they receive the set of questions that they will have forgotten about the study.

If the expert will be giving his answers to the interviewer over the telephone, the same things as above will need to be done, with only slight modifications. Usually with a telephone interview of this sort, the expert is mailed the information and then called to obtain his judgments. The interviewer goes through each question over the telephone with the expert. Please note that the telephone communication between the expert and interviewer should not last more than about fifteen minutes per call. As with scheduling interviews, the interviewer should ask the expert if it is a convenient time to talk before proceeding. The interviewer can use the written cover letter as a guide in introducing the expert to the elicitation. We do not propose that the interviewer actually read the cover letter over the telephone because this practice causes most people to sound repetitive and flat.

How to set up for an interactive group situation

Many of the support materials listed below are things that a visitor or meeting coordinator could assist with, if one were available. In any case, we recommend having the following papers and supplies ready for the experts:

1. **Materials that the group moderator/interviewer will need,** such as lists of participants, the introduction to the elicitation, copies of the expert's background materials and statements of the problem, the moderator/interviewer's question topics, and documentation formats.
2. **The mechanical recording devices,** cassettes, and extra batteries or extension cords needed. If the sessions are to be video recorded, the equipment (and the experts' seating) should be set up in advance and tested. It is a good idea to test everything in place to determine if the machine can receive sound or picture from each location. The cassettes can be labeled by the meeting date before the meeting.
3. **Name tags for the experts,** especially if they are not all known to the meeting moderator or each other. With large groups, we recommend that the names be printed across each side of 8 1/2- by 11-inch papers that have been folded lengthwise and placed before the expert for easy viewing. If the experts are supposed to sit in particular seats, their name tags can be set out in advance to show them where to sit. Otherwise, it is very effective to present each expert a packet of the materials listed below and labeled by his name tag.
4. **The program schedule.** A general schedule includes the goals and deadlines. More detailed schedules include the names of speakers, topics of discussion, and their times. On projects where there will be presentations, we recommend listing the speaker, title, times of the talks, and the person in charge of each session. This procedure has been effective in keeping elicitation meetings on schedule. Please note that time overruns are sometimes necessary, such as when the experts are clarifying a question. The person listed as being in charge of the session can be asked to decide if the discussion should continue or if it is being unproductive and should end.

5. **The expert's copy of background materials, statements of the questions, and documentation formats,** if the experts will be recording their own data. Often the experts will wish to provide current references to be distributed at the meetings. It is much easier to receive and copy these in advance of the meeting than during it.
6. **Refreshments.** If the experts will be meeting for more than a few hours or for more than one day, the use of caffeinic beverages and snacks can prolong their productivity. Sweet, fatty foods, such as doughnuts, should be avoided because of their sedating effects.
7. **A list of restaurants and things to do** (sights, tours, and activities), if the meetings will last more than a day and are being held outside of the experts' work places. We have found that the extra effort of treating the experts like VIPs is more than rewarded by their good will and favorable impression of the project. As a professional visitor coordinator told us, "People may not remember the technical content of the meetings but they will remember how they were treated."
8. **Extra note paper and pens** for the experts' use.

Tips on Conducting the Elicitations

For convenience we have divided the elicitation sessions into several parts: introducing the experts to the elicitation process, gathering and recording the data, and monitoring the session for bias. General suggestions are offered below on how to do each of these. However, we recommend that you use what was learned from practicing or pilot testing as your primary guide in conducting the elicitation.

Introducing the Experts to the Elicitation Process

This section represents step 2 in the program for handling bias, as discussed in chapter 3 *Steps in a program for handling bias.* However, the procedures mentioned below are so generally beneficial to the expert's performance that we recommend following them even if the bias program will not be used. If the program for handling bias will not be used, one procedure, that of briefing the expert on the biases to which he may be prone, can be omitted without detriment.

Make the Experts Aware of the Potential for Introducing Bias and Familiarize them with the Elicitation Procedures--Step 2

HOW TO SET UP FOR AN INDIVIDUAL INTERVIEW

1. **Introduce yourself,** if you have not already met the expert. Give the expert the cover letter (mentioned in *Setting up for the Individual Interview*) to establish the interviewer's credentials. If the expert will be questioned on classified matters, show the expert further identification, and name a person known to him who will vouch for you. After the expert has glanced at the cover letter, quickly deliver this same information verbally. (Giving the information verbally ensures that all the experts will receive the same

information, something that cannot be said of their reading it.) For example, we usually mention the sponsor of the study, how the expert was selected, what the expert will be doing, how long it is likely to take, how his data will be protected, and the anticipated product of the project and the expert's access to it. At the end of this description, ask the expert if he has any questions. Start the mechanical recording devices, if these are to be used.

2. **Start with questions on the expert's professional background,** if these will be asked. These simple questions will allow the expert to get into the flow of being interviewed and will reassure him that he is capable of answering later questions. Examples of these questions: *How many years have you worked in your present position? What educational degrees do you hold? In what fields are your degrees?.*
3. **Give the expert some sample questions** to familiarize him with the use of the response mode, if that mode is likely to be a difficult one for him.
4. **Brief the expert on any biases** that were identified as being likely to occur (chapter 8). Give the expert ideas on how he can strive to counter the tendency towards these biases. (The section *Definitions of Selected Biases* in chapter 3 provides examples of this type of information.)
5. **Give the expert the set of questions and verbally go over any instructions.** Ask the expert if he has any questions.
6. **Tell the expert that he can begin.** Record the expert's beginning time if a record of duration of interviews is being kept.

HOW TO SET UP FOR AN INTERACTIVE GROUP SITUATION

1. **Distribute the materials described above in** *How to Set Up for an Interactive Group Situation.* Turn on the recording devices if these are to be used.
2. **Introduce the meeting moderator/interviewer, the project staff, and the experts.**
3. **Review the purpose of the project, its schedule, and, in general the elicitation procedures for the benefit of the experts.** Some descriptions of the elicitation procedures are as follows. *You will meet together for this week to develop detailed statements or representations of the problems. On the last day, Friday, you will vote on what you think the answers should be.* A more detailed overview of elicitation procedures is as follows. *You will meet here three times: First, to become familiar with the project and the elicitation procedures; second, to present up-to-date technical information and refine the rough-drafted questions; and third, to give your expert judgment in private meetings with an interviewer.*
4. **Give the experts sample questions to work** so that they can practice using the response mode. If there are any techniques to properly using the response mode, they can be introduced and practiced here. For example, if the response mode is probability distributions, Hogarth (1980:149) offers eight keys to assigning these.
5. **Brief the experts on those biases that were identified in chapter 8 as being likely to occur.** This briefing should include an explanation of

why the selected biases occur and of how the expert can reduce his tendency to introduce them. (The section *Definitions of Selected Biases* in chapter 3 provides examples of this type of information.) In addition, the briefing on bias should include exercises that are designed to evoke the selected biases. After the experts have completed the exercises, the moderator/interviewer can read the answers and allow the experts to correct their own. These exercises can convince the expert that he, like everyone else, is prone to these biases. If the briefing is given without exercises, we have noticed that the experts are not as effective in countering their tendencies toward bias, perhaps because they were never convinced that they, too, would be vulnerable.

6. **Ask if there are any questions.** Afterwards, state that the introduction is concluded and the sessions will now begin.

How to Set Up for a Delphi Situation

If the entire Delphi will be conducted by mail, the expert will not be introduced to the elicitation by the moderator/interviewer in person. Instead the expert will receive the cover letter and set of questions described above.

If part of the Delphi will be conducted by telephone, call the the expert to assist him in understanding the set of questions or just to obtain his answers. Use the items listed for individual interviews above as a basis for introducing the expert to the elicitation process.

Gathering and Recording the Expert Data

Using the individual interview, group interactive, and delphi situations. As a general rule, let this phase be guided by the results from the practice runs or pilot tests (as described in chapter 9). If the elicitations were not pilot tested, we recommend reading *Common Difficulties--Their Signs and Solutions* at end of this chapter before proceeding. Reading the section on common difficulties may prevent you from encountering some of them.

Using the three techniques for eliciting problem-solving data. The three techniques for eliciting problem-solving data--verbal protocol, verbal probe, and the ethnographic technique--are frequently used in combination with individual interviews. Occasionally two of them, the verbal probe and ethnographic technique, are used with the interactive group situation to gather a few sentences on how the experts solved the problem. Details are provided below on how to administer these techniques.

Verbal protocol. To review, verbal protocol involves instructing the expert to think aloud as he progresses through the problem (Ericsson and Simon 1980). For example, the expert is given a written copy of the problem:

> What feed program would you start this colt on? The colt is 6 months, 550 lbs., has an average metabolism, and will receive light exercise through ponying. Please solve this problem as you do others that you receive in this field. Please try to think aloud as you work your way through the problem. Your thinking aloud is as important to me as the answer that you reach.

The expert's verbal protocol resembles someone talking to himself. This technique is from psychology. The following are suggestions for setting up to conduct the verbal protocol.

- **Place the interviewer's chair slightly behind the expert and to the opposite side of his handedness** (i.e., for a right-handed expert, the interviewer is on the expert's left). There are several advantages to this positioning. First, the expert will not be as able to see what or when the interviewer is taking notes. If the expert becomes aware of what the interviewer is interested in, this awareness may influence, or bias, the response. Second, this position allows the interviewer to see, unobtrusively, what the expert is looking at, marking, or writing.
- **Design the questions, at least initially, so that they are like those in which the expert has expertise.** Otherwise, there is little point in using the expert. The elicitation sessions should be set up so they resemble, as much as possible, the environment in which the expert usually solves such problems. Otherwise, factors may be introduced into a particular elicitation session that change the expert's usual way of thinking and make him inconsistent. The sessions can be conducted in the expert's customary work place, if interruptions can be controlled. For instance, telephone calls can be handled by call forwarding.
- **Obtain copies of whatever visual aids (e.g., tables, graphs, and equations) or references the expert will be using.** This practice allows the interviewer to follow what the expert is viewing even when it can not be seen over his shoulder because of the distance and the print size. It also provides a hard copy for recording what the expert is looking at and marking. Hard copies should also be obtained if the expert is solving problems on a computer. In the later case, the hard copies are copies of the computer screens.
- **Take notes rather than rely solely on the expert or recording devices.** Experts are unreliable in taking notes and should not try to provide detailed written accounts because this activity is likely to be done at the expense of their thinking. Thus, not only is little data obtained but it is likely to be unreliable as well. Recording devices, by themselves, do not provide complete records. For example, they do not show the marks that experts make on their papers as they think. In addition, recording devices malfunction, so a backup copy, your's, is likely to be needed.
- **Emphasize that the expert is to work through the question rather than talk hypothetically about how it could be solved.** Experts are often unaware of, or mistaken about, how they actually solve a problem. They, therefore, provide more reliable (less biased) information if they are verbalizing while they are solving the problem. The message on actually working the problem need only be delivered the first couple of times that you work with the expert.
- **Stress the importance of thinking aloud when instructing the expert to begin solving the question.** The first time an expert solves a question may remind him of a test situation in school. As a result, his tendency may be to rush through the question to give the solution. The interviewer must contrive to convince the expert to think aloud and must be careful not to emphasize the importance of verbalizing at the expense of problem solving.

The latter alters the way the expert would normally work the problem and is therefore undesirable.

Example 10.1: ***Illustration of the Verbal Protocol***

The interviewer in this example wishes to learn, in general, how the expert plans a feed program, what information on the horse and on the types of feed are needed, and what concepts (e.g., the horse's metabolism) are considered. The expert is presented with a written copy of the problem:

> What feed program would you start this colt on? The colt is 6 months old, 550 lbs., has an average metabolism, and will receive light exercise through ponying. Assume that the colt has normal good health.

Interviewer--Solve this problem as you do others you receive in this field. Please try to think aloud as you work your way through the problem. Your thinking aloud is as important to me as the answer that you reach.

Expert--*(The expert slowly reads the question aloud.)* The first thing that I'll need to find out is what this colt would weigh full grown. *(The expert scans some xeroxes that were previously copied from reference books.)* Let's see, at six months and 550 lbs., he will be about 1100 to 1300 lbs. full grown.

Next, I need to find the balance between hay and grain that I'd feed a horse of this age. *(The expert looks at another table.).* Four lbs. of hay and 8 lbs. of grain. However, I like to feed more hay, so I will aim for 6 lbs. of hay and 7 lbs. of grain.

Other constraints that I have are the amount of protein in pounds needed for a colt that will mature to a 1100- to 1300-lb. horse. *(The expert refers to another table.)* I'll want to balance between 1.74 and 1.89 lbs. of protein per day.

(The expert examines another chart.) I also need to balance protein in the overall diet to about... 14.5%, calcium to... 6%, and phosphorus to... 0.45%.

Foods that I like to feed in the diet are oats, corn, and soybean meal. I also like a feed supplement for extra protein, Horsecharge, and alfalfa and timothy hay. I'm getting their percentages of protein, calcium, and phosphorus from the charts. *(The expert ceases thinking aloud, records the percentages of these feeds, and begins using a calculator.)*

The expert writes a list of feeds and weights and leans back in the chair, signaling that the exercise has finished.

■

- **Frequently, the expert will stop thinking aloud and need prompting to resume. Several types of prompts are given below.** One reason for using different prompts is to avoid repetitiveness that can be irritating to both expert and interviewer. Another reason is to use the most subtle prompt possible to remind, rather than to distract, the expert. Distraction

can cause experts to lose their place and then solve the problem differently than they would have otherwise. The prompts are given in order of most-to-least subtle.

1. Look intently at the expert and lean forward slightly.
2. Look at the expert, lean forward, and clear your throat.
3. State *please try to think aloud.*

The first prompt would be used with an expert who only needs the occasional reminder to think aloud. The expert perceives the interviewer's forward movement almost subliminally and then responds. The second prompt is slightly stronger because it triggers associations concerning the expert's throat (e.g., my throat should also be making sounds). The third prompt would be used on the expert who failed to notice or respond to the first two prompts.

- **Generally if the expert is counting or otherwise performing calculations, a reminder to think aloud would be inappropriate.** The expert could lose his place. If it is necessary to prompt the expert at this time, the first prompt would be better than the third. The nonverbal aspect of the first prompt allows the expert to note it without being distracted.

Verbal probe. To review, the verbal probe is questioning done at a particular time and in a specific manner. The type of verbal probe discussed here is used immediately after the expert has reached a solution, focuses on only one problem, and is indirectly phrased so as to minimize influencing the expert's thinking. For example, immediately after the expert has solved the problem and given the answer, the verbal probe is used to learn why that answer was given.

Interviewer--Why did you give that answer--that feed program?
Expert--Well, it provides the right amount of protein, calcium, and phosphorus for a horse to grow at this age.

- **If the verbal probe is used in an individual, face-to-face interview, set up as mentioned earlier for the verbal protocol.** As a general rule, individual interviews are used for pursuing more detailed information than can be obtained in group settings.
- **If there will be multiple experts in a group setting, the group moderator mentions, in advance, that the experts will be asked to provide their reasons for giving their answers.** The experts should be asked to provide this information verbally because people, in general, do not provide good written records of their reasoning. Their notes are usually sketchy and, for this reason, more likely to be misinterpreted by the interviewer than their more complete verbal counterparts.

Example 10.2: *Illustration of the Verbal Probe*

Assume that the expert has used verbal protocol, as previously illustrated, and is at the point in time where the problem has just been solved. In this example, the expert will be asked for an answer and then administered the verbal probe to learn why that answer was given.

Interviewer--What was your answer?
Expert--The diet, per day, that I would recommend for this horse is 4 lbs. of oats, 1.5 lbs. of corn, 0.5 lbs. of soybean meal, 1 lb. of Horsecharge, 4 lbs. of alfalfa, and 2 lbs. of timothy. This is, of course, only the starting point.
Interviewer--Why did you give that answer--that feed program?
Expert--Well, it provides the right amounts of protein, calcium, and phosphorus for a horse to grow at this age.

■

- **Word the verbal probe in the past tense (e.g., *Why did you give this answer?*) to emphasize your wish to know how the expert actually solved the problem.** Some experts will begin, even after solving a problem, to describe how they or a colleague could have solved it. The interviewer's response to such a beginning should be something like: *I am interested in your thinking and how you solved this problem.*
- **Check for tautological responses.** Tautologies are reasons that do not truly provide the *why* information that is being sought. For example, *I gave that feed program because it seemed right* is a tautological response. Although, the tautology here is obvious, tautologies can be difficult to discern in an unfamiliar domain, vocabulary, and/or when fatigue sets in. Frequently, the desired information can be obtained by using the ethnographic technique to probe on the tautology.

Ethnographic technique. The ethnographic technique involves restating the expert's words into questions. For example, the ethnographic method could be used to probe on one of the expert's responses to obtain an operational definition that could then be entered into the expert program. The expert has just said that the colt's feed program may need to be adjusted if the colt is not keeping his weight on.

Interviewer--Not keeping his weight on?
Expert--Yes, not gaining as he should at this time.
Interviewer--At this time?
Expert--At his age, 6 months, he should be gaining between 1.5 and 2 lbs. per month.

- **How the ethnographic technique is set up is usually determined by the setup for the other techniques used or by the situation in which it is used alone.** For example, if the ethnographic were used with the verbal protocol, the verbal protocol's setup would be used. If the ethnographic technique is to be used by itself, no special set up would be needed. For example, if it is to be used on one expert at a time, it can be set up as a conversation would be (e.g., with the interviewer sitting adjacent to or across from the expert).

Example 10.3: ***Illustration of the Ethnographic Technique***

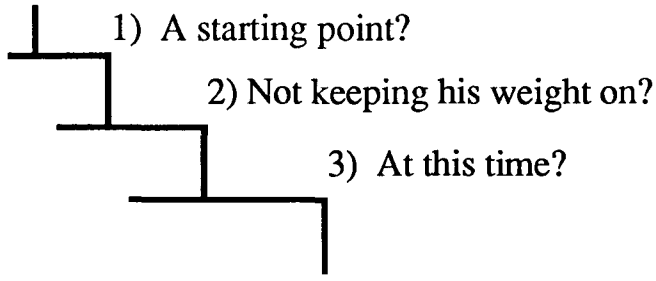
The expert has just answered the verbal probe explaining why that answer was given. The ethnographic technique could be used either to investigate this information or information given earlier. Earlier, the expert had given the feed program and added, *Of course, this is only the starting point.* For the purpose of this illustration, the ethnographic technique will be applied to the reply to the previous verbal probe and then to the subject as it develops.

Interviewer--A starting point?
Expert--Yes, because the rations may need to be adjusted, if the colt is getting fat or if he's not keeping his weight on.
Interviewer--Not keeping his weight on?
Expert--Yes, not gaining as he should at this time.
Interviewer--What do you mean by "at this time"?
Expert--At his age, 6 months, given his expected total weight, he should be gaining between 1.5 to 2 lbs. per day.

■

The ethnographic technique was used to follow a line of the expert's thoughts to a detailed level where an operational definition of *not keeping his weight on* was obtained. The interviewer stepped down three levels in questioning, as shown in the figure below. The ethnographic technique is effective at taking the questioning to the desired level of granularity, in this case to a definition that would be represented and programmed. In another case, the ethnographic technique could be used to branch across several topics at a more general level.

Example 10.4: ***Illustration of a Series of Ethnographic Questions***



■

- **Vary the length of the ethnographic questions so that they do not become tedious.** For example, if the expert said, *The books give samples and I start with these,* a short ethnographic query would be, *samples?* For variety, a longer query would be, *What samples do the books give?*

- **If the ethnographic technique is used in problem-solving sessions, plan to use it only after the expert has reached the point in problem solving that satisfies the interviewing goals.** Otherwise, the expert is likely to lose what he has in short-term memory while responding to questioning. Then, if the interviewer resumes questioning where the interruption occurred, the expert will have to guess at his previous thoughts. Bias can result. For this reason, the ethnographic technique is best used after other techniques or by itself. Occasionally, the ethnographic technique can be inserted during the use of another technique but only if it is asked quickly and not pursued to fine granularity.

Monitoring and Adjusting for Bias During the Elicitation

During the gathering and recording of data mentioned above, bias can intrude. This section includes steps three and four of the program that we propose for handling bias. The program includes the following steps:

1. Anticipate which biases are likely to occur in the planned elicitation (chapter 8)
2. Make the experts aware of the potential for introducing bias and familiarize them with the elicitation procedures (chapter 10)
3. Monitor the elicitation for the occurrence of bias (chapter 10)
4. Adjust in real time to counter the occurrence of these biases (chapter 10)
5. Analyze the data for the occurrence of these biases (chapter 14)

How to perform steps 3 and 4 is described below.

Monitor the elicitation for the occurrence of particular biases--Step 3.

For many of the selected biases, there are signs that indicate their occurrence. The interviewer or a trained observer can watch for these signs during the elicitation. In general, monitoring biases, as described in this book, requires that the experts verbalize their thoughts and answers. Without this feedback, we have found the monitoring to be much more difficult.

Signs of selected biases (group think, wishful thinking, inconsistency, availability, and anchoring) are given in chapter 3 in *Steps in a program for handling bias*; *Signs of selected biases*.

Adjust, in real time, to counter the occurrence of particular biases--Step 4.

This step is perhaps the most delicate one in the program for handling bias because if done carelessly it could confuse the results and any later analyses. The interviewer needs to decide in advance on the timing of the adjustment because these adjustments change the conditions under which the data is gathered. If a condition leading to bias is corrected before the analyzable data has been gathered, there is no problem. If, however, the analyzable data has been gathered under two conditions, when bias was occurring and then when it was corrected, the data will be mixed. Unless the situations can be clearly

separated, such as before and after correction of a particular bias, testing for the presence of this bias is not possible.

Mentioned below are some of the techniques that we have used to adjust for bias. In general, our approaches have been to (1) impede those factors contributing to a particular bias, or (2) to employ the opposite bias. For example, the interviewer's tendency to lead or influence the expert contributes to social-pressure bias. When using the first approach, we have advocated the use of elicitation methods--verbal protocol, verbal probe, and ethnographic technique--that curb this tendency (Meyer et al. 1989). Another example of this approach is focusing on the fatigue and confusion that contribute to our natural tendency to be inconsistent. The elicitation sessions can be scheduled to stop before the point when most of the experts become fatigued. The basis of the second approach, fighting bias with bias, comes from the grandfather of survey design, Stanley Payne (1951). Payne believed that all interviewing was biased and that one should therefore aim for equal but opposite biases. An example of this technique is to try to have experts anchor to their own judgments in attempts to counter a group-think situation.

Suggestions on how to adjust for selected biases (group think, wishful thinking, inconsistency, availability, and anchoring) are given in chapter 3 under *Steps in a Program for Handling Bias; Suggestions for Countering Selected Biases.*

Common Difficulties--Their Signs and Solutions

Difficulty: ***The experts resist the elicitation process or resist giving judgments under uncertainty.*** We have seen this situation develop in a few interactive group situations, especially among engineers who were unaccustomed to thinking in terms of uncertainty. One of its first signs is the experts' reluctance to give their judgments. They may mutter amongst themselves or they may openly criticize the elicitation procedures and refuse to give their data.

An expert who is reluctant to give his judgment should not be forced for two reasons: (1) he is probably not an expert in this area if he does not feel qualified to give his judgments; and (2) his reaction to being forced is likely to be negative and to detrimentally affect his view of the entire study. This reaction is illustrated by reviewer's statements about a study: "The participants were forced to provide unsubstantiated guesses as input" (Benjamin et al. 1987:F-5,6).

Solutions: In general, the solution to this problem must rest on addressing the experts' reasons for resistance. We recommend that the objecting experts be individually taken aside and asked in a pleasant manner why they are reluctant to give their judgments. Ask the experts separately because asking them as a group may not reveal the individual reasons and it may reinforce their resistance. We have been given the following reasons for experts' resistance: (1) that they had misgivings about an elicitation process that they were involved in developing, (2) that the experts did not trust the use to which their judgments would be put, (3) that they thought there would be new data related to the question and that they would they would not be able to use it in making their judgments, (4) that they feared that their judgments would be misinterpreted, taken out

of context, and unfairly criticized, and (5) that they did not think they could give reliable estimates given the high uncertainty of the subject area.

The means for resolving this difficulty lies in addressing its underlying causes. For instance, if the experts implicitly feel that they did not have sufficient input into developing the questions as described in (1), allow them to refine the questions now. In addition, let them know of the design considerations and constraints that led to the creation of the current methods. This information will help convince them that the elicitation was not ill conceived and will save you from having to accept everything that they suggest. That is, you will be able to point out which of their proposed modifications would not meet the project's constraints.

The other reasons for resistance mentioned above can be similarly resolved:

If the experts do not trust the use to which their judgments will be put (2), show them an outline of the project report or arrange for the eventual users of their judgments to brief them.

If the experts were concerned about using the most current information in making their judgment (3), help them circulate this information before the elicitation sessions. Explain that their judgment is a snapshot of their knowledge at the time, that it is likely to always be in a state of change, and that there must be some cutoff point for writing up the results of the study.

If the experts were worried about misinterpretation of their judgments (4), explain that they will be asked to review the documentation of their judgments after the elicitation situations. If possible, show them an outline on how their judgments will be presented in the report or model.

Address the experts' doubts about giving judgments in areas of high uncertainty (5) by acknowledging the validity of this concern. Explain that the high uncertainty in the field is one of the reasons why their judgment must be elicited--that other data is lacking. Also summarize how the elicitation practices will help them give more reliable estimates.

Questioning the experts on their reasons for resistance also helps identify those experts who are the most vocal in their reluctance. Sometimes only one or a few experts are causing the rest to question the elicitation. We have found it effective to first address the concerns of these few natural leaders. Again, we suggest talking to these experts separately and privately. If you cannot resolve the leaders' concerns, you may be able to balance them with the positive things that can come out of the project. For example, one expert was very concerned that the funder of the project would use the expert data inappropriately, something which we could not absolutely prevent (i.e., people can always pull data out of context and misuse it). Aside from emphasizing the good that would come from the study (the opportunity for the experts to meet together, to pool their most current data, to identify gaps, to conduct further research, and to cause the field to progress), we could not completely assuage this expert's worries. However, we were able to convince this expert to proceed, and thus the others, by privately appealing to him for his assistance. We explained that he was a natural leader in this group, that the others were being influenced by his views, and that we needed him to agree to be elicited.

If all else fails, consider doing as much of the elicitation process as possible with one expert at a time. For example, perhaps the experts' data can be obtained from

individual interviews rather than from group situations. In this way, you only deal with one uncooperative expert at a time and can be more flexible in responding to their individual needs.

Difficulty: ***The experts seem confused about what they are to do or how they are to do it.*** The experts may progress very slowly through the elicitation procedures and look puzzled. Sometimes it may be hard to discern if they are confused about the elicitation or stalling because in both cases they will be slow to provide their judgments. If the experts were introduced to the process by working sample questions and using the response mode, these exercises will provide the first evidence of their confusion. If they have not been given any explanation of what they are to do, then the existence and source of their confusion can almost be assumed. Intensive pilot tests, if performed, will have given some warning as to which parts of the elicitation were likely to cause confusion.

Solution: The first step, as in the above-mentioned difficulty, is to identify the cause of the expert's reaction. The cause may stem from the general procedures for the elicitation, such as when the experts are to meet together and when separately for individual interviews; it may stem from the response mode that they are to use; or it may be confusion over some instructions. One of the special problems with confusion is that people are often unable to express what they are confused about One way of identifying the source is to quickly go through the information that was already presented and ask the expert to signal the points that are unclear. Then, it should be conveyed that the problem has resulted from a lack of clarity in the presentation rather than from a lack of understanding on the part of the expert.

We urge working with the expert until his confusion is resolved because not doing so can have severe ramifications. In the one situation where we saw experts pushed to provide data when they were confused, the experts later criticized the elicitation methods and raised doubts as to the data's validity.

Difficulty: ***The final statement/representation of the question or the expert's last data were not documented.*** The main cause of this common difficulty is the evolutionary quality of elicitations. The phrasing of the question often evolves gradually during a group session, and there is usually no special sign that signals its entry into its final form. Then too, in the individual elicitations, the expert may try to solve the question in different ways, backtracking a few times, before settling on a process and arriving at a final judgment. Thus, the failure to record the final form of the question or its solution often goes undetected--unless it is caught by those who worked on it. To detect this difficulty, the involved persons must review the documentation while their memory of this information is still sharp. We recommend that the interviewer request the group or the expert to review the final form while it can still be easily corrected.

Solution: If this difficulty is detected immediately following an elicitation session, the group or the expert and interviewer can update their copy using their memories. If not, the interviewer will have to replay the mechanical recordings to update the documentation. The latter is usually very time consuming and difficult because the communications are not as clear in retrospect as they were when they occurred.

Difficulty: ***Bias may have occurred but its presence was not monitored during the elicitations.*** Bias has not usually been considered in designing or conducting elicitations. For the majority of studies, the possibility of bias is not considered until the end of the elicitation and only then because some one has raised this question. When the question of bias is raised so late, there are fewer options for detecting its presence than there would have been earlier (e.g., anticipating which were likely to occur and then monitoring the session for their intrusion). However, if sufficient data was gathered, it may be possible to test for the presence of a particular bias.

Solution: A detailed description of how to analyze the expert data for bias is given in chapter 14. If this analysis is being performed at the request of others, they may identify the biases that they want to test for. In our experience, people have been most worried about three biases: the one that occurs because the interviewee's thinking was led by the interviewer; the one that occurs when the expert is led by other experts, such as in a group-think situation; and the one that arises from a conflict of interest, from the expert's wishes or interests influencing his judgment. These three can be considered motivational biases because they arise from human needs and desires.

The data can also be analyzed for the presence of particular cognitive biases, and often more easily (Meyer and Booker 1989:13). For example, on one project, we tested for the underestimation of uncertainty (Meyer and Booker 1989). The experts had estimated the likelihood of achieving national magnetic fusion milestones within particular time periods. They gave probability estimates, such as 0.90, and ranges, such as ± 0.10. The experts' ranges were analyzed and found to be within one standard deviation of the set of probability estimates. This result indicated that the experts thought they were adequately accounting for uncertainty when they were only accounting for about 60% of uncertainty (Meyer, Peaslee, and Booker 1982).

As a general rule, we recommend that the data be analyzed for the presence of whatever biases are possible and that these results be included in the project report.

Difficulty: ***There is wide disagreement between the expert's data.*** Although differences in the expert data may not be a problem, they are frequently perceived as being such. Some view interexpert disagreement as an indication that the elicitation process failed--that it did not produce the one *right* answer or means of solving the question. However, as mentioned in chapter 1, experts can legitimately solve the question in different ways, and the ways that they solve the question affect the answer that they reach (Ascher 1978, Booker and Meyer 1988a, Meyer and Booker 1987b). Handling varying expert data can pose problems if this data is to be brought together such as in a program for a knowledge-based system or if it is to be mathematically aggregated.

Solutions: Before a solution can be offered, it must be determined whether there is truly a problem--whether the expert disagreement was caused by a weakness in the elicitation or by the natural differences between the experts. To answer this question, data is needed on the experts' problem-solving processes, in particular their definitions and assumptions. If the expert's questions were only loosely defined, the experts will often make their own definitions and assumptions in clarifying the question. In so doing, they can create different questions, use different problem-solving processes, and give

correspondingly different answers. If this has happened, there is indeed a problem. If, however, records show that the expert's used the same definitions and assumptions (i.e., answered the same questions), any differences can be assumed to result from their use of acceptable but different means of solving the question. In this case, if there is a problem, it is one of perception.

If the expert's differences were found to be caused by their answering different questions, there is little that can be done. One option is to separate the expert's answers according to the questions that they answered. Sometimes, this operation is used to declare the answers of one or more experts as being inapplicable to the question. That is, their answers differed from the rest because they answered a different question, one that is outside the study's selected questions. Another option is to recontact those experts who used the discrepant definitions and assumptions and ask them to use the agreed-upon definitions or assumptions and solve the question again. Recontacting the experts to have them solve the problem again can be very time consuming and can, unless carefully done, cause others to question the competence of the project staff.

If the experts legitimately differed, you may still find that you have problems--those of other's perceptions. Frequently, those funding the project or the outside reviewers interpret expert differences as a negative sign. We recommend that they be made aware that expert disagreement is valid and that the differences arise from variation and uncertainties in the experts' answers. Also, they should be provided with any evidence that the differences were not induced by the elicitation (e.g., the experts used the same definitions of the questions).

For detailed information on how to aggregate experts' answers, see chapter 16. Suggestions on how to integrate differing expert's knowledge is given in chapter 9 of *Knowledge Acquisition* (McGraw and Harbison-Briggs 1989).

PART III

ANALYSIS PROCEDURES

11

Introducing the Techniques for Analysis of Expert Judgment Data

The analysis of expert judgment data is viewed by some analysts as *ad hoc* at best and impossible at worst. There are no standard forms of modeling and analysis that are applicable to all types of problems in the world in general, and there are no standard forms for analyzing all expert judgment data. Part III, therefore, contains a compendium of techniques whose applications vary depending upon the design of the elicitation and upon the goals of the study. In this chapter (11) a detailed explanation of these techniques for those unfamiliar with them is provided. In the following chapters, 12 through 18, we mention the various statistical and computational techniques and make suggestions for their use.

Some statistical concepts are needed to understand the analyses and techniques suggested. The glossary provides basic definitions of these concepts and some discussions are presented below. Special attention is given below to the concepts of **random variables** and **probability distributions**.

Techniques are discussed in the remaining sections of this chapter. These techniques cover three basic areas of statistical methods: (1) **simulation** techniques, (2) data analysis techniques, and (3) **Bayesian** techniques. The simulation section describes the uses, advantages and disadvantages of Monte Carlo and **bootstrap** simulation methods. The data analysis section describes multivariate techniques such as **correlation, cluster, factor,** and **discriminant analyses; analysis of variance**; and a **decision analytical** tool called Saaty's method. The methods developed from Bayes Theorem and its applications are discussed in the third section. In all sections the descriptions are brief and limited to the applications for expert judgment data analysis. References are furnished for outside works where further details and applications can be found.

Random Variables and Probability Distributions

The basic premise for the concept of the random variable is that for all quantities of interest which are being measured or observed, there is a set of possible values that

quantity of interest can assume or *take on*. That quantity of interest is called a random variable, or simply **variable**, and the values that it can assume come from this possible set. When a measurement is done, an observation is made, or a datum is gathered, the random variable is assigned one of those possible values. The assignment process is represented by a real-valued function. The function that does the assigning is a **probability distribution function** for that random variable.

Probability distribution functions (pdfs) or, more simply, probability distributions are nonnegative functions that allocate unit mass to points on the real line. The two properties of pdfs are that each value of the pdf is greater than or equal to zero and that the area under the entire curve is equal to 1.0. In terms of **probability**, the interpretation of $f(X)$ is not the frequency or probability of x, it is the probability per unit interval of a small increment of x, dx.

An easier probability interpretation is available by accumulating the areas under the curve of the pdf for a range of values of the random variable. The function resulting from such an accumulation or integration is the cumulative distribution function or cdf. The notation for the cdf is $F(x)$, and the interpretation for $F(x)$ is the probability that $X \leq x$. In other words, the cdf gives values for the probability that the random variable is less than or equal to a value of the random variable.

The pdf and cdf have the following relationships. The cdf is the integral of the pdf. The pdf is the derivative of the cdf. Therefore one can be calculated from the other using calculus.

Examples of random variables are (1) the probability of an event, (2) the failure rate of a component, (3) the amount of time to repair a component, (4) the time to failure of a component, (5) the number of failures for a component, (6) the age of an expert, (7) the colleagues working with the expert, (8) the **rank** of one alternative relative to another, (9) the odds of winning a race, and (10) just about anything else that is determined in an experiment, an observation, an elicitation, or other data-gathering process. The symbol for a random variable is usually a capital letter such as X. The values that X can have are generically listed as x.

Examples of probability distribution functions are (1) the normal, Gaussian or bell-shaped curve with one **mode** (hump) at the center and with half of the curve cut at the mode forming an exact mirror image of the other half; (2) the lognormal distribution with the $\log(X)$ distributed as a normal and with a single mode shifted to the left (skewed right) such that the right tail is long and drawn out; (3) the beta distribution with possible values of X restricted to the range of 0 to 1.0 and with many shapes possible such as U-shaped, horizontal line, decreasing curve, skewed right curve, and bell-shaped curve; (4) the uniform with the equal probability assigned to each value of X such that the distribution is a straight, horizontal line; and (5) the exponential with the values decreasing in a decay function shape. The symbol $f(x)$ is used to denote the distribution function and to designate the value of that function at $X = x$.

To this point, the pdfs and random variables discussed are continuous pdfs and continuous random variables. This means that x and $f(x)$ can be any value along the real number-line. For random variables that can only *take on* a finite number or discrete number of possible values, the corresponding pdf is a discrete distribution function. For discrete distributions, the value of the function is the probability that $X = X$, $\Pr(X=x)$.

Cumulative distribution functions for discrete distributions are found by summing all the probability values rather than by integration.

Examples of discrete distribution functions are the binomial and the Poisson distributions. The binomial characterizes the number of success as x in a given set of n trials where each trial has the probability of success as p. The Poisson characterizes the number of occurrences (failures), x, in a given, fixed t units of time where the failures are independent and at a constant rate, λ.

Descriptions and Uses of Simulation Techniques

With the availability of fast and personal computers, the use of simulation techniques has recently grown. New data sets can be formed from the original data set (**bootstrap** simulation) or from specified distributions representing the data (standard **Monte Carlo** simulation). Both are useful in gaining new insights about the data set and for forming estimates that might not be available.

Monte Carlo Techniques

What is Monte Carlo simulation?

Traditional simulation, as described here, is often referred to as Monte Carlo simulation. The basic idea is to form new **samples** or distributions of data either from existing samples or from specified distributions. The formation process is done by randomly selecting values from the existing samples or specified distributions, making some calculation, performing this several hundred or thousand times, and collecting the hundreds or thousands of calculated values into a table or distribution for **inference** purposes.

The following steps illustrate how to perform a simple Monte Carlo simulation to solve a problem that has no tractable mathematical solution. The steps may be summarized as follows:

Step 1: Determining the desired quantity to be estimated.

For example, the product of two random variables is desired. Each variable has a specified distribution; however the product of these distributions is not in a closed or known form.

Step 2: Determining the distributions from which sampling is done.

For example, each of two random variables is distributed as normal--one with mean 0, **variance** 1 and the other with mean 1, variance 0.5.

Step 3: Finding or code a computer program that randomly selects a value from each of the specified distributions from step 2.

A random number generator is needed to randomly select values, and an algorithm is needed for mapping that value onto the specified distributions. Many such codes are

available. In appendix B, a code is given that selects values from beta distributions. In appendix C, a code is provided that forms empirical, or data-based, distributions for simulation. For a more complete guide on simulation techniques see Ripley (1987) or Johnson (1987).

Step 4: Determining the number of samples, N, to be taken.

With most modern computers, 1000 samples is not too expensive or time consuming and gives accuracies for nearly two decimal places.

Step 5: Constructing a program for taking the N samples.

For each sample, the random values are chosen and mapped onto the function or distribution. The desired quantity (e.g., the product of the two normals from step 1) is then calculated. Upon completion of the N samples, there will be N values of the desired quantity.

Step 6: Collecting the values of the desired quantity for making inferences.

The N values of the desired quantity form a distribution of possible values for that quantity. Estimations are possible using this distribution. For example, estimates for the mean and the variance of the products of the two normals are available. First, the 1000 (N=1000) values for the 1000 products of two normals are collected and ordered. The mean of these 1000 values is the estimate of the desired mean quantity, and the variance of these 1000 is the estimate for desired variance quantity. Percentiles are also available from this distribution. Sometimes the interest is purely in the resulting distribution rather than in an **estimator** (such as a mean, variance, or **percentile**). Quantities of interest in the resulting distribution are the shape, center, tails, spread or range, and modes (humps).

Example 11.1 uses Monte Carlo simulation in a **reliability analysis** application.

EXAMPLE 11.1: ***Monte Carlo Simulation***

The reliability of a component, r, is often characterized by the beta probability distribution function:

$$f(r) = \frac{1}{B(x,n-x)} r^{x-1}(1-r)^{n-x-1} \quad ,$$

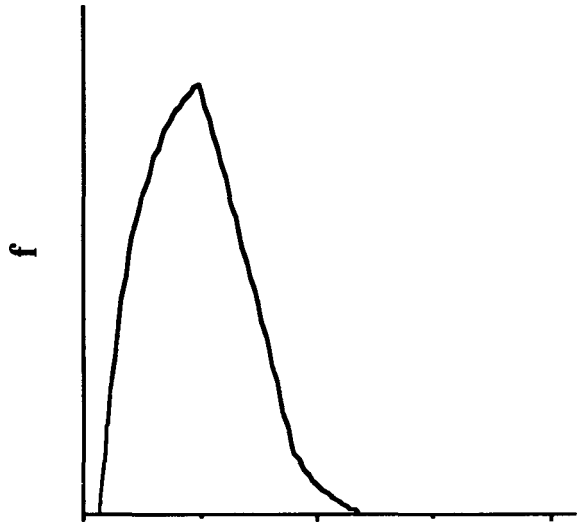
where the parameters $n - x$ and n are interpreted from the number of component failures, x, in n trials, and B is the beta function (Martz and Waller 1982).

For a system composed of two components, r_1 and r_2, in series, the system reliability is the product of the reliability functions. If the reliability function for each component is distributed as a beta distribution, then the system reliability is the product of two beta distributions. Simulation is used to find the distribution of this product that represents the system reliability.

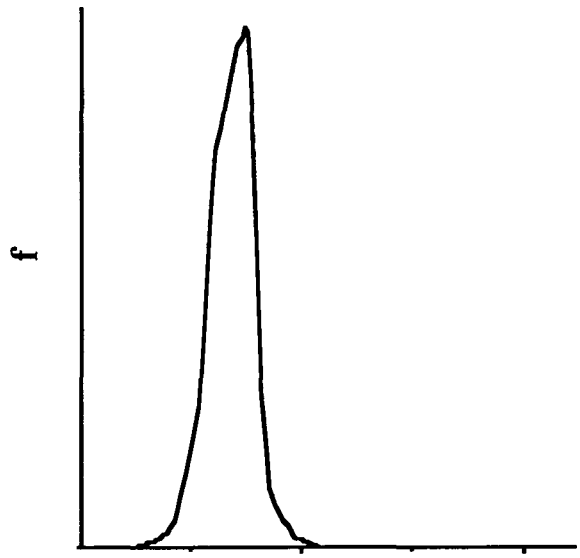
It is known that the first component failed in four out of seven trials; therefore, x_1 = 3 and n_1 = 7. There is no data on the second component, but an expert estimates that its average reliability is 0.90 and that a 95th percentile value for the reliability is 0.99 (i.e.,

there is a 5% chance that the reliability is greater than 0.99). The parameters of the beta with that mean and 95th percentile are $x_2 = 11.7$ and $n_2 = 13.0$. The notation for this beta is beta $(x,n\text{-}x)$ or, in this case, beta (11.7,1.3). These parameters are found using the code in appendix B.

Comparing the two components, the reliability of the first is much worse than the second. Their beta distributions below reflect this:



Beta 1



Beta 2

Using the steps for simulation, the system reliability was determined:

Step 1: The desired quantity is the product of the two reliabilities, $r_1 \cdot r_2$.

Step 2: The reliability for component 1 is distributed as a beta (3,4), and the reliability of component 2 is distributed as a beta (11.7,1.3).

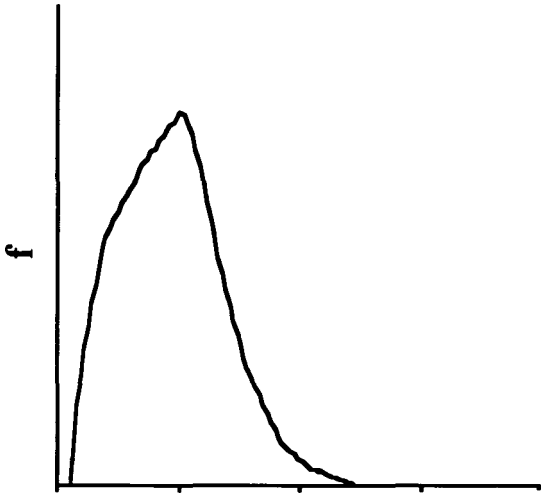
Step 3: A computer code was written using a uniform random number generator to select values of r_1 and r_2 from the two betas.

Step 4: The chosen value of N, the number of **samples**, is 1000.

Step 5: A computer code was written to determine the random values of r_1 and r_2 from their respective beta distributions. A uniform random number generator chooses a random value for the probability of the reliabilities of r_1 and r_2. The code determines the values of r_1 or r_2 by mapping the chosen probabilities through each beta distribution. This is an inverse process; i.e., the probability values for the beta are chosen as random numbers; then they are mapped through the beta function to determine the reliability value, r. This is done for r_1 and again for r_2 in each of the N samples. In each sample, the chosen value of r_1 is multiplied by the chosen value of r_2. The code for the beta sampling and simulation is given in appendix B.

Step 6: The 1000 product values are stored and sorted. They are plotted to indicate the shape of the system reliability distribution. The results follow:

The mean of this simulated distribution is	0.36
The standard deviation is	0.16
The 5th percentile is	0.12
The 95th percentile is	0.65



Beta 1 • Beta 2

Therefore, the system reliability is not very good. It is dominated by the poor reliability of the first component.

The product of two beta distributions does not usually produce a known type of distribution function. However, research (Bruckner and Martz 1987) indicates that this final distribution is not too far from being in the beta distribution family.

■

Advantages and disadvantages

The major advantage of Monte Carlo simulation is that it allows the analyst to determine estimates of desired quantities that would be either impossible or extremely difficult to determine by theory or mathematical computation. (Some analysts rely on the phrase, "when in trouble, simulate!") This major advantage is very important in **risk analysis** and reliability applications where many system components interact, but the performance of the entire system is needed for study. It is also advantageous for uncertainty calculations where uncertainties are often characterized by distribution functions and these functions are combined to determine the uncertainty of a larger system.

Monte Carlo simulation, especially the sampling and collecting steps, is relatively easy to program. Many such programs already exist, and most computers have random number generators available.

Simulation is a tool for increased understanding of the data. Many times new features in the data can be identified. Through simulation, sensitivity studies can be done to determine which characteristics or components are most important to the system.

The disadvantages of simulation are related to the advantages listed above. Programming can be difficult if complex distributions are required or if packaged distribution programs are not available. Bad random number generation or improper use of sampling techniques can wildly distort results. Care is needed in programming and in the use of packaged routines.

Simulation can be expensive and consume computer time. However, that is less of a problem today than it was even 5 or 10 years ago.

Uses for Monte Carlo simulation

Widespread use of this technique can be found in risk and reliability studies. As mentioned above, in these studies there are a large number of components (random variables) comprising a system, and system behavior is required for making inferences. Each component can affect the system or other components. The result is a complicated expression describing the system behavior in terms of the various components.

In addition, each component may have a behavior that is random or uncertain in nature. In this case, each component can be characterized by a probability distribution function. The system is then a complex combination of these component distribution functions. This is the type of situation where simulation provides a way of obtaining information about the system. This situation also describes how simulation can be used in an environment characterized by uncertainties (chapter 17 gives more details on this).

Monte Carlo simulation is the backbone of other data-based analysis techniques, such as the bootstrap. These techniques have there own advantages such as uses for exploratory data analysis and the handling of small sample-size problems.

Bootstrap Sampling and Estimation

What is the bootstrap?

One of the more recent developments in the simulation community is the increased use of empirical distributions. Empirical distributions are distributions resulting from the original data set. Constructing distributions from the data without making assumptions

about the form of the distributions (e.g., that the data is distributed as a normal) falls under the general heading of **nonparametric** techniques. Because the original data is used over and over again in a simulation method, construction of empirical distributions is often referred to as resampling techniques. Both descriptions apply to the technique known as the bootstrap (Efron 1979).

Advantages and disadvantages

The bootstrap has the advantage of relying solely on the original data itself without assuming that the data follow a particular probability distribution. Bootstrap simulation allows the formation of distributions of any desired quantity (such as the sample **median**). The resulting simulated distribution of that quantity provides estimates of the variance and percentiles of that quantity. Statistical biases resulting from the bootstrap estimation process are possible but are also estimable. (Statistical bias means that the expected value of the estimated quantity does not equal the parametric or theoretical value.) The extremely small samples that are common in expert judgment data are troublesome with many techniques, parametric or nonparametric. However, the bootstrap does reasonably well for smaller samples (Efron and Gong 1983).

The major disadvantage of the bootstrap technique is that the formation of samples in the simulation is limited to the range of the original data. This limitation tends to form empirical distributions with truncated (chopped off) tails (at both ends). Another caution in using the bootstrap deals with a statistical bias that may be induced in the estimation process. That is, the bootstrapped value for a chosen estimator may not be unbiased. However, the statisical bias can and should be monitored as part of using the technique.

Uses for the bootstrap

In general, the bootstrap technique is used for providing estimates of parameters that would normally be obtained by assuming a distributional form of the data or of the parameter. The bootstrap avoids the necessity of these assumptions and therefore can be used in any application or problem setting. In expert judgment problems, its use is ideal because information regarding distributions of expert judgment data is lacking. The bootstrap is also useful for the small sample sizes that are usually found in expert judgment applications.

In this book, the bootstrap will be used in three different ways. First, it will provide ways of investigating the correlation structures and biases (motivational and cognitive) in the data. Second, it will be used as a simulation method to characterize and analyze uncertainties in the experts' estimates. Finally, it will provide a convenient way for aggregating the multiple expert's estimates into a single estimator and corresponding distribution.

How to implement the bootstrap

The implementation of the bootstrap is easily done on any programmable computer or calculator. The original sample data, that is, the responses to a single question, is denoted as $\{x_1, x_2, ..., x_n\}$ for the n experts. The parameter to be estimated is denoted by θ. Then the estimate of θ from the bootstrap results will be $\hat{\theta}$. A simulation is done by randomly forming many samples, N, (usually of the original size, n) from the original data.

The samples are formed by replacing sampled data points back into the original data set so that they can be sampled again. In other words in forming the first sample, if x_3 is chosen at random, it could be chosen at random again, appearing twice in the same sample. This method is called sampling with replacement. The basic idea of the simulation is to calculate the estimate of θ for each sample formed. The various values of $\hat{\theta}$ denoted as $\hat{\theta}_j$ (j = 1, 2, ..., N), are then ordered to form an empirical distribution for θ in the same manner that the resulting distribution of values is formed in Monte Carlo simulation. The following steps summarize the bootstrap technique.

Step 1: Determining the desired quantity to be estimated, estimator

For example, the estimator or quantity of interest is the sample median of a sample of size n.

Step 2: Deciding on the number of bootstrap samples to be taken, N

One thousand will usually give an acceptable standard error (to within two decimal places.)

Step 3: Forming of N random samples of size n from the original data by sampling with replacement (replacing each sampled value back into the data set so that it is available to be chosen again)

Step 4: Doing N times: Calculate the desired estimator for each sample (e.g., the median)

These individual estimators are $\hat{\theta}_j$.

Step 5: Calculating the overall bootstrap estimator of θ, $\hat{\theta}$ using

$$\hat{\theta} = \frac{\sum_{j=1}^{N} \hat{\theta}_j}{N}$$

Step 6: Calculating the standard error of the estimator, $\hat{\sigma}$, using

$$\hat{\sigma} = \left[\frac{\sum_{j=1}^{N} (\hat{\theta}_j - \hat{\theta})^2}{N-1} \right]^{1/2}$$

Step 7: Ordering the N estimates of $\hat{\theta}_j$ to find (1-2α)% putative central coverage intervals for θ

These intervals represent the central (1-2α)% area of the bootstrap distribution. In the case of the median, for α = 0.5, then this area corresponds to the middle 90% of the median values generated in the N bootstrap samples.

Step 8: Finding the two values of the $\hat{\theta}_j$ that correspond to the α% and (1-α)% in the ordered list

These two values, $\hat{\theta}(\alpha)$ and $\hat{\theta}(1-\alpha)$, are the α-th and (1-α)-th percentiles of the bootstrap distribution for θ, and they form the (1-2α)% putative central confidence interval.

Step 9: Calculating the statistical bias induced by the bootstrap technique by calculating the value of $\hat{\theta}$ from the original sample, $\hat{\theta}_0$

Statistical bias in the technique is the difference between the value of the estimator calculated from the original sample, $\hat{\theta}_0$ and $\hat{\theta}$ calculated from step 5. If the bias is large (more than 10% of either of the two values), then the use of the bootstrap technique might not be appropriate.

These steps are easily programmed into a code. A FORTRAN version is included in appendix D, and an illustration of this code is given in example 11.2

EXAMPLE 11.2: ***Bootstrap Simulation***

Fourteen experts provided value estimates for a quantity on a continuous **linear scale** from 0 to 1:

(0.07, 0.50, 0.28, 0.63, 0.95, 0.70, 0.62, 0.70, 0.58, 0.78, 0.4, 0.68, 0.43, 0.60) .

If the data are plotted, it can be seen that the distribution of these values is not symmetric. The median is a commonly used **aggregation** estimator to represent such asymmetric data. The median of this data is 0.61. However, since the exact form of the distribution of this data cannot be assumed, there is no available estimator for the variance of the median. The bootstrap sampling method provides such a variance estimate, and more.

Following are the steps outlined in the bootstrap method:

Step 1: The parameter of interest is the median.
Step 2: One thousand bootstrap samples will be taken.
Step 3: One thousand random samples of size 14 were taken from the original set of data using the bootstrap code in appendix D.
Step 4: To form each sample, 14 data points were selected from the original set, with replacement; e.g., a single value could be chosen more than once in a given sample. For each sample, the sample median was calculated; therefore, a set of 1000 medians resulted.
Step 5: The average of these 1000 medians was 0.597.
Step 6: The **standard deviation** of these 1000 medians was 0.058.
Step 7: The 1000 medians were sorted in order.
Step 8: To form the 90% putative interval for the median, the 5th and 95th percentiles were calculated by finding the 50th and 950th values from the ordered median set. These values were 0.465 and 0.680, respectively.

Step 9: The bias is 0.610 - 0.0597 = 0.013, which is less than 10% of either value, making the bias induced by the bootstrap method acceptable.

Thus the median is 0.597 with a variance of 0.003, and the 90% putative intervals for the median are 0.465 to 0.680. This forms a complete description of the chosen aggregation estimator for this data set.

■

Descriptions and Uses of Data Analysis Techniques

Multivariate Techniques

Multivariate analysis techniques refer to statistical methods designed for the analysis of data sets with many random variables (multivariate). Most data sets are multivariate in structure; however, the study of the possible variate relationships is often ignored or assumed without analysis. In such cases important results are not considered in drawing conclusions.

The variates are of two types: (1) **answer**, response, or **dependent variates,** and (2) **ancillary,** conditional, or **independent variates.** The names differ, but the relationship between these two types is the same; the independent variates are measured or fixed variables that influence or are functionally related to the dependent variates. The independent variates are usually thought of as variables that can be controlled by the analyst in the data-gathering process, and the dependent variates are usually thought of as variables that are unknown and uncontrolled and are the values being gathered in the study. The dependent variables are dependent upon the independent variables.

Correlation analysis

Correlation refers to the linear relationship between two variables. If the two variables have values that are completely identical, their correlation is 1.0. A graphical interpretation is that if the values of the two variables are plotted and they fall exactly on a line with positive slope, then the correlation is 1.0. If the values fall exactly on a line with negative slope, then the correlation is -1.0. In most applications, the values do not all lie exactly on a line. Instead the values of the two variables form a general scatter with either a positive trend, negative trend, or no trend. For these cases the correlation is near 1.0, near -1.0, or near 0.0, respectively. The closer that the correlation values are to 1.0 or -1.0, the stronger the linear relationship between them is.

The most common measure of correlation uses the Pearson product-moment correlation coefficient, r. For two variables, x and y, that coefficient, r, is calculated by

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\left[\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}\right]} \quad ,$$

where $\bar{x}$ and $\bar{y}$ are the **mean** values of x and y, and the summations are overall values of x, y, and xy pairs. Example 11.3 illustrates how the correlation coefficient is calculated.

EXAMPLE 11.3: ***Correlation Analysis***

The following answers to two questions were elicited from 10 experts. Are the two answers correlated?

Expert	Answer 1, A_1	Answer 2, A_2
1	0.10	0.15
2	0.05	0.00
3	0.15	0.20
4	0.11	0.10
5	0.10	0.12
6	0.14	0.10
7	0.00	0.05
8	0.00	0.10
9	0.15	0.20
10	0.09	0.09

$\Sigma x = 0.89$ $\quad \Sigma y = 1.11$

$\Sigma x^2 = 0.11$ $\quad \Sigma y^2 = 0.16$

$\Sigma xy = 0.12$

A calculator formula for the correlation coefficient is

$$r = \frac{\Sigma xy - \Sigma x \Sigma y/n}{\sqrt{[\Sigma x^2 - (\Sigma x)^2/n]} \cdot \sqrt{[\Sigma y^2 - (\Sigma y)^2/n]}} = \frac{0.12 - 0.89 \cdot 1.11/10}{\sqrt{(0.11 - 0.89^2/10)} \cdot \sqrt{(0.16 - 1.11^2/10)}} .$$

Using the above values, $r = 0.64$. For 10 experts (n=10), an r value greater than or equal to 0.63 is considered different from $r = 0$ (no correlation) using a 5% **level of significance**. Therefore, this x, y relationship is strong (significant at 5%) and is positive in nature (as x increases, y increases). Tables are available in most statistics books indicating the cutoff values for r for various significance levels and sample sizes.

■

Other correlation measures exist, such as nonparametric ones based on ranks and agreement between x, y pairs. For the Pearson correlation, if the data for x and y are normally distributed, then a zero correlation can be interpreted as x and y being two variables that are statistically independent. More discussions on the correlation or **dependence** of variables are available in chapter 14.

Cluster analysis

Clustering refers to the grouping structure of the data. Clusters can be formed as (1) how the values from a single variable (quantity of interest) are grouped, or (2) how the different variables are grouped. The groups are formed based upon how closely related the values or variables are to each other. Cluster analysis traditionally refers to methods of determining how the values of one or more variables can be grouped together. Most methods form hierarchical clusters -- beginning with all the data in one cluster, splitting it into two, then three, etc., until each datum forms its own cluster.

Cluster forming is determined by many different methods (Duran and Odell 1974). Basically, the methods compute and then compare some measure of the distance between clusters. For example, some methods form clusters by maximizing the distance between cluster means. If the squared Euclidean distance is used, then the method is the centroid method. Other methods, called linkage methods, form clusters by maximizing distances of individual observations in the clusters. Still other methods minimize variances to determine clusters.

For a multivariate data set, clusters formed by grouping similar variables are of interest. Variable cluster analysis uses the correlation or covariance matrix of the variables to determine the clusters of the variables. Again, many techniques for variate cluster determination are available.

In either the data or variable hierarchical clustering, the clustering process by the various methods begins with the formation of one large cluster and ends with each observation or variable in its own cluster. It is up to the analyst to decide which set of clusterings between these extremes is to be used for interpretations and conclusions. Many packaged programs provide graphical trees to aid in this decision. The distance measurements are plotted against the cluster groupings to help determine which grouping to use. Example 11.4 illustrates how this is done using the SAS® software for data and variable clustering.

EXAMPLE 11.4: ***Cluster Analysis of Variables and of Data***

The following data was gathered on 11 experts.

YRSEXP	--	total years of professional experience
YRSED	--	total years of college education
YRSJOB	--	number of years on the current job/project
DEGAREA	--	code (1-3) for discipline of highest degree/education
EXPAREA	--	code (1-5) describing major area of experience
ANSWER	--	answers to the question on a continuous number scale [0,1]

Expert	*ANSWER*	*YRSEXP*	*YRSED*	*DEGAREA*	*YRSJOB*	*EXPAREA*
1	0.11	6.00	4	1	0.75	5
2	0.12	4.50	6	2	1.00	5
3	0.90	8.50	8	3	1.25	5
4	0.78	3.00	6	3	2.25	4
5	1.00	5.00	6	2	3.00	5
6	0.17	4.75	5	2	1.00	4
7	0.14	4.25	6	2	3.25	3
8	0.83	5.00	6	2	3.00	3
9	1.00	7.00	6	2	2.50	2
10	0.88	8.25	7	1	3.00	2
11	0.20	4.00	5	1	1.00	1

I: Cluster analysis of all six variables

Cluster analysis of the six variables indicates the following cluster formations. Each formation is based on the proportion of variance explained by the clustering--the higher the proportion, the tighter the individual clusters and the larger the separation between clusters.

Proportion of Variance	Cluster Formation	
0.37	All 6 variables in one cluster	---
0.63	*YRSEXP YRSED YRSJOB ANSWER*	Cluster 1
	DEGAREA EXPAREA	Cluster 2
0.76	*YRSEXP YRSED*	Cluster 1
	DEGAREA EXPAREA	Cluster 2
	YRSJOB ANSWER	Cluster 3
0.85	*YRSEXP YRSED*	Cluster 1
	EXPAREA	Cluster 2
	YRSJOB ANSWER	Cluster 3
	DEGAREA	Cluster 4

The decision of which cluster formation to use can be based upon the proportion of variance values. The proportion is doubled from the all-in-one cluster formation to the two-cluster formation. This increase makes the two-cluster formation an attractive choice. The remaining formations do not indicate as great a change in the proportion.

Cluster interpretation is just as important as deciding which cluster formation to use. If the clusterings in a particular formation do not make sense, then using that formation makes no sense no matter how good the clustering is (in this case, how large the proportion of variance is). In the example, the two- cluster formation does make sense. Cluster 1 contains the years-related variables and the answer. Cluster 2 contains the

variables related to the experts' areas. Based on interpretation and proportion of variance, the two-cluster formation is the logical choice.

II: Cluster analysis of the experts using the answer variable

It is notable that the answers given by these experts cover the entire available range on the offered number-line from 0 to 1: some experts are high, and some experts are low. A cluster analysis on the answer variable can be done in a number of different ways. One way is to use only the answer variable itself, disregarding the other variables. Another way is to use the other variables to help determine how the answers cluster. Yet a third way is to use a subset of the other variables. A natural choice for such a subset would be the set of variables that provide similar information about the answers. From the above variable cluster analysis, this subset would include the three variables describing the number of years for various items.

Again, a number of viable cluster formations results from the cluster analysis of the answer variable. Criterion, such as distance measures between clusters, can be used to determine which formation is a logical choice. Again, logical or reasonable interpretation is equally important.

Only the answers are used to determine the following cluster formations from an analysis based on the centroid method. The expert numbers are listed, and the centroid distances are listed for the different cluster formations.

Cluster Formation (clusters are connected by underlines)	Centroid Distance
5 9 3 10 4 8 1 2 7 6 11	1.33
5 9 3 10 4 8 1 2 7 6 11	0.27
5 9 1 2 7 6 11 3 10 4 8	0.15
5 9 1 2 7 6 11 3 10 4 8	0.11
5 9 1 2 7 6 11 3 10 4 8	0.09
5 9 1 2 7 6 11 3 10 4 8	0.05
5 9 1 2 7 3 10 6 11 4 8	0.04
5 9 1 2 3 10 7 6 11 4 8	0.04
5 9 1 2 3 10 7 6 11 4 8	0.02
5 9 1 2 3 10 7 6 11 4 8	0.00

The greatest change in the distance measure is from the single-cluster formation to the two-cluster formation. This change delineates the major division between the high-answer experts and the low-answer experts. All other cluster formations differ very little in the distance measure and do not have any better interpretative value.

III: Cluster analysis of the experts using three variables

A cluster analysis on the answers was performed including the three variables describing years of items. Again, the centroid method was used, and the distance measures were used for cluster formation determination.

Cluster Formation (clusters are connected by underlines)	Centroid Distance
5 4 8 7 6 11 2 1 9 10 3	1.14
5 4 8 7 6 11 2 9 10 3	0.84
5 4 8 7 6 11 2 9 10 3 1	0.70
5 4 8 7 6 11 2 9 10 1 3	0.63
5 4 8 7 6 11 2 9 10 1 3	0.59
5 8 7 6 11 2 9 10 1 3 4	0.51
5 8 7 6 11 2 1 3 4 9 10	0.34
5 8 6 11 2 1 3 4 7 9 10	0.31
5 8 6 11 1 2 3 4 7 9 10	0.23
5 8 1 2 3 4 6 7 9 10 11	0.05

These cluster formations are different from the ones in part II of this example. The reason for this difference is that the information from the other three variables is being used to determine clusterings. The new information forms clusters differently from those formed using just the answers.

Again, there is a substantial change in values from the single-cluster formation to the two-cluster formation. This change indicates that the two-cluster formation is reasonable. However, the experts in these two clusters do differ from those in the two clusters in part II. This discrepancy indicates that the three added variables make a difference in the results. For this example, other modeling is indicated to use the other variables in combination with the answer variable. Chapter 15 discusses model formation and uses. However, valuable information has been gathered from the cluster analyses about variable relationships. Cluster analysis can be a useful tool for such investigation.

■

Factor analysis

A related method to variable clustering is factor analysis. Factor analysis analyzes the relationships among a set of variables through their correlation and covariance to determine what information is shared among subsets of the variables (common factors) and what information is unique to each variable (unique factors). A common factor is an unobserved, imaginary variable that shares information with (or contributes to the variance of) at least two of the original variables. A unique factor is an unobserved, imaginary variable that represents information from only one of the original variables. It is assumed that all the unique factors are uncorrelated with each other and that the unique factors are uncorrelated with the common factors.

Factor analysis can be done using many different methods. Some methods rely on the use of principal components (Kshirsagar 1972). Other methods use least squares methods or maximum likelihood techniques. Still others formulate scores that are based on correlations. There are many ways of transforming the variables, called rotation, so that the formation of the common and unique factors is both optimized and logical. Example 11.5 illustrates the principal factor method for an unrotated set of variables.

It is not recommended that factor analysis be used casually. To select the proper factor method and use of rotation requires experience and knowledge of the various factor

methods. To interpret the results requires even more knowledge. Only if the variable set has a good underlying structure can a simple factor analysis reveal important information. A successful result can then be used to reduce a large number of variables to a smaller number of variables containing the common factors. However, in most applications, the (common) factors produced are not interpretable and are useless.

EXAMPLE 11.5: ***Factor Analysis***

A principal components factor analysis using SAS® was done on the set of six variables from example 11.4. The (common) factor loadings listed below indicate how each variable maps onto each factor. A successful result would be for each variable to map nearly completely (e.g., 0.8 or greater) onto one factor and for the set of factors to be much smaller than the original number of variables. Then success must also be measured in the interpretation. Can meaning be attached to each of the factors based on the variables that constitute (load onto) them? This is the same problem faced in interpreting the results of a cluster analysis.

The principal components factor analysis resulted in six factors for the six variables. This result does not indicate that a successful reduction in the number of variables is possible. Each factor represents a portion of the original variation of the six variables. Examining these portions can help determine which factors are the most important:

	Factor					
	1	2	3	4	5	6
Portion	0.42	0.25	0.18	0.08	0.05	0.02
Cumulative	0.42	0.67	0.85	0.93	0.98	1.00

The first three factors account for 85% of the variation. While 85% of the variation would be considered very good for one factor and good for two factors, it is not a good result for three factors.

The six variables map or load onto these three (common) factors as follows:

	Factor		
	1	2	3
ANSWER	0.88	-0.13	-0.06
YRSEXP	0.56	-0.20	0.79
YRSED	0.90	0.10	0.13
YRSJOB	0.60	-0.43	-0.54
EXPAREA	0.02	0.88	0.16
DEGAREA	0.53	0.71	-0.33

This mapping of the variables is not very good. *ANSWER* loads well onto the first factor, and so does *YRSED*. *EXPAREA* loads well onto the second factor. However, *YRSEXP*, *YRSJOB*, and *DEGAREA* spread their loadings over all three factors. If an

interpretation were attempted, the first factor would represent some combination of *ANSWER* and *YRSED*, the second factor would represent the two AREA-type variables, and the third factor would represent some combination of *YRSEXP*, *YRSJOB*, and *DEGAREA*. Such an interpretation is not very clear. Thus, this example would not represent a successful factor analysis. For a more successful analysis, see chapter 15 on model formation.

■

Discriminant analysis

In cluster analysis, variables could be clustered, or observations based on variables could be clustered. Discriminant analysis determines how well other (ancillary or independent) variables predict the classifications or groupings described by the dependent or classification variable. If the ancillary variables do a good job of determining the classifications of the dependent variable, then they can be used to predict into which class or group that a new observation (a new value of the dependent variable) will fall.

A discriminant function is calculated from the predictor variables based on the distances between classes and the variation within classes. The theory is based on the data for the variables following a multivariate normal distribution. This assumption is highly restrictive. It is unlikely that the data from any expert elicitation would be multivariate normal. Therefore, this method is offered as an exploratory data analysis and premodeling tool. It is only a means to other analyses, and not the sole analysis tool. Example 11.6 illustrates how discriminant analysis can be useful.

EXAMPLE 11.6: ***Discriminant Analysis***

The eleven experts from examples 11.4 and 11.5 were asked to make recommendations based on the answers (scaled 0,1) they gave. The recommendations (1-4) were given as follows:

Expert	Recommendation
1	1
2	1
3	4
4	3
5	4
6	2
7	1
8	3
9	4
10	4
11	2

It is hypothesized that the original answer variable and some of the other variables might be good discriminators in determining the recommendations. A discriminant analysis was done. The results of this analysis indicate whether the six other variables are good

discriminators if they properly classify the recommendations into the four classes. A discriminant function for each class is calculated using the six other variables. These functions are used to determine the existing classifications and are also used to predict classifications of experts who did not make the recommendation but provided information on the six variables.

The classification was a success based on the following results:

Expert	Recommendation	Predicted Class Based on Variables
1	1	1
2	1	1
3	4	4
4	3	3
5	4	4
6	2	2
7	1	1
8	3	3
9	4	4
10	4	4
11	2	2

No misclassifications occurred. The discriminant functions for the classes are as follows:

Recommendation 1 = -109.2 + *ANSWER* • 382.6 + *EXPAREA* • 9.2 + *DEGAREA* • -14.0 + *YRSED* • 19.2 + *YRSEXP* • 9.0 + *YRSJOB* • 5.0.

Recommendation 2 = -107.8 + *ANSWER* • 483.0 + *EXPAREA* • 7.5 + *DEGAREA* • -14.4 + *YRSED* • 20.2 + *YRSEX* • 8.3 + *YRSJOB* • 1.5.

Recommendation 3 = -611.1 + *ANSWER* • 1145.9 + *EXPAREA* • 15.2 + *DEGAREA* • -32.0 + *YRSED* • 43.3 + *YRSEXP* • 17.7 + *YRSJOB* • -1.4.

Recommendation 4 = -908.1 + *ANSWER* • 1398.8 + *EXPAREA* • 18.4 + *DEGAREA* • -40.6 + *YRSED* • 52.4 + *YRSEXP* • 22.7 + *YRSJO* • -2.5.

The main result from this example is that the recommendations given are functions of the values of the other six variables. A careful look at the recommendations versus the answer variable reveals that there is a very strong positive correlation between them. This correlation is reflected in the above discriminant functions as well.

Discriminant analysis assumes a multivariate normal distribution for all the variables. It is likely that this assumption is not valid for expert judgment data. Therefore, this technique should only be used as an exploratory analysis tool to gain information about the variate relationships. Results from the discriminant analysis should be cross validated by the use of other analysis and modeling tools as described in chapters 13, 15, and 18.

■

Analysis of Variance

Analysis of variance is a broad-based methodology for analyzing data from an experiment where the dependent or response variables are considered functions of the independent or ancillary variables. Each independent variable or factor is tested (controlled) at specified values (levels) of that variable. In most analysis of variance usage, the experiment is carefully designed using techniques that specify the levels of all the factors to be studied (Snedecor and Cochran 1978) so that a minimum number of experimental tests or observations is required to yield information on the importance of all the factors.

The major inferences made and hypotheses tested in analysis of variance concern the equality (or lack of it) of the means for the various values or levels of the factors. If the factor means differ for the responses, then that factor is said to be significant in determining the response.

Multiple factors are designed and tested in a single set of experiments. Each factor is tested individually for its influence or effect upon the response. Combinations of two or more factors can be tested at a time for their combined effect upon the response. These combinations are called interactions. Interactions are important because the factors individually may not be significant but their interaction may be significant.

Using analysis of variance for expert judgment studies is not recommended because there can be no controlled design of the study (experiment). Most of the factors are gathered during the elicitation and cannot be controlled prior to the study to produce a good experimental design. At most, only single factors can be analyzed, each in a separate analysis, as illustrated in example 11.7. Performing several analyses of variance such as this is not recommended. In doing so, the analyst loses control over the chance of detecting differences in the factor means when no differences exist. This is the alpha level or **type-I error**. Multiple factors must be tested using the analysis of variance technique in a single analysis and that requires good experimental design before the elicitation. Because good design is not possible in expert judgment applications, analysis of variance is used as an exploratory tool for examining simple *between* versus *within* experts' values as is suggested in chapter 14.

EXAMPLE 11.7: ***One Factor Analysis of Variance***

In the previous examples in this chapter, the expert's degree area or discipline was described using a variable *DEGAREA* which had three values (levels). These values were 1 = mechanical engineering, 2 = nuclear engineering, and 3 = physics. *DEGAREA* is therefore a single factor with three levels that could be analyzed using analysis of variance.

There are three experts with *DEGAREA* = 1, six experts with *DEGAREA* = 2, and two experts with *DEGAREA* = 2. For a single factor analysis of variance (AOV), this imbalance of the number of experts across levels is fine. However, if two or more factors were to be analyzed, a balance of the numbers of experts in each level for both factors would be necessary in most software programs for a conventional analysis of variance.

For example, suppose the two factors were *DEGAREA* and *EXPAREA*. *DEGAREA* has three levels and *EXPAREA* has five levels. For a balanced design of the experiment, 15 experts would have to be interviewed, one for each combination of all the levels of the two factors. With such a balanced experiment, tests on the effects of *DEGAREA* and *EXPAREA* would be possible. To have a test for the *DEGAREA/EXPAREA* interaction, more than one expert for each combination would be necessary. To find experts necessary to fit these combinations would not be very practical or even possible and illustrates the difficulties in designing experiments for expert elicitation.

The data for this example is as follows:

	Factor: *DEGAREA*		
	1	2	3
Answers	0.11	12.00	0.90
	0.88	1.00	0.78
	0.20	0.17	
		0.14	
		0.83	
		1.00	
Sum	1.19	3.26	1.68

To test the differences between the three means for the factor, variance components are calculated for between (across) the three categories and within the three categories. If the variance between the three is significantly greater (using an F-test statistic) than the within variance (which acts as the noise level), then the factor means are not the same. The idea behind this comparison of variations is that the levels of the factor will influence the response if the means of the three levels are different. To test if the three means differ, the variation between the three levels is compared to some noise level. If the between variation is large compared to the noise level, a difference in the three levels is indicated. This noise level is determined as the variation within the three levels, the within variation. Thus the name analysis of variance implies what is actually tested. It is the variations that are compared and tested to determine if the means of the factor levels differ.

The following analysis of variance table outlines the steps of the variance calculations and the test (F-test) used to compare the variances:

Term	df	Sum of Squares	Mean Square	F-Statistic
Between	*dfB*	*SSB*	*MSB*	*F*
Within	*dfE*	*SSE*	*MSE*	
Total	*dfT*	*SST*		

MSB represents the variance between the three levels of the factor. *MSE* represents the error or noise level calculated using the within-levels' variance. The F-statistic measures the significance of the between variance relative to the within or noise variance. If this value for F is larger than the critical value for an F distribution with parameters of dfB and dfE, then the between variance is significantly larger than the noise. The conclusion would then be that the means of the factor levels differ and that the factor significantly affects the answers given by the experts.

The following formulas indicate how to calculate the quantities in the analysis of variance table:

N = number of observations = 11
T = grand total = sum of all data = 1.19 + 3.26 + 1.68 = 6.13
SST = sum of squares of all 11 answers = $0.11^2 + 0.88^2 + ... + 0.78^2 = 5.00$
SSB = sum of squares of categories (between) = $1.19^2/3 + 3.26^2/6 + 1.68^2/2 = 0.47 + 1.77 + 1.41 = 3.65$
C = correction factor = the grand mean = $T^2/N = 6.13^2/11 = 3.42$
SST = total sums of squares = SST - C = 1.58
SSB = between sums of squares = SSB - C = 0.23
SSE = within (error) sums of squares = $SST - SSB$ = 1.35
dfB = degrees of freedom for between = number of categories - 1 = 2
dfT = degrees of freedom for total = N - 1 = 10
dfE = degrees of freedom for error = $dfT - dfB$ = 10 - 2 = 8
MSB = mean square between = SSB/dfB = 0.23/2 = 0.12
MSE = mean square error = SSE/dfE = 1.35/8 = 0.17
F = F-test statistic = MSB/MSE = 0.12/0.17 = 0.71

To determine if this F value is larger than the critical value for an F distribution with 2 and 8 degrees of freedom, a table or program of F distributions is required. These are available in all statistical packages and textbooks (Snedecor and Cochran 1978).

To use the tables, a significance level is needed and is determined by the analyst. The level represents the chance that the analyst is willing to accept for making the following error: declaring that between variance is larger than the within variance when, in truth, they are the same. Usually a 5% value is commonly chosen for the chance of making this error (called a type-I error). This chance is called the level of significance or α. Sometimes an extremely safe or conservative value of 1% is chosen. Sometimes a liberal value of 10% is chosen.

If the significance level is chosen at 5%, then the critical value for this F distribution (with degrees of freedom two and eight) is 4.46.

F(2,8,0.05) = 4.46
F = 0 .71
F(2,8,0.05) < F (no factor effect indicated)

The F value calculated must be larger than 4.46 for the factor to have a significant effect. Here the F value is only 0.71; therefore, the factor means are considered the same, and the factor itself has no effect on the answers provided by the experts.

■

Saaty's Technique for Pairwise Data Analysis

What Is Saaty's Method?

In lieu of asking experts to compare multiple items simultaneously either by numerical or **qualitative** evaluations, experts can be asked to make relative comparisons and evaluations. It is very difficult for humans to simultaneously examine and evaluate many items. However, with **pairwise comparisons**, it is only necessary to examine two items at a time. Paired comparisons can be made by evaluating items relative to each other in a qualitative evaluation such as better, worse, or equal. Comparisons can be made in a **quantitative** evaluation using specified numerical scales. Either way, the comparisons made by the experts are then quantified using a matrix algebraic approach resulting in relative numerical weighting factors for all the items being compared. The paired comparisons technique and one of the scales designed for this technique are part of the Saaty Analytical Hierarchy Process (AHP) (Saaty 1980).

The AHP has been widely applied in many decision analysis problems. Its basic appeal for these applications is its ease of use by the experts and its ability to easily quantify qualitative evaluations.

A simple example illustrates the usefulness of the technique. An expert is asked to determine which of the following meteorological conditions would be the most likely to cause a loss of off-site power in a power plant:

1. Flash flooding at plant site with 0.5 to 2 inches of water
2. Flash flooding with 2 to 4 inches
3. Flash flooding with more than 4 inches
4. Lightning (direct hit to power lines)
5. Direct hit by a tornado
6. Winds between 20 to 40 mph
7. Winds higher than 40 mph

The expert begins by thoroughly defining and clarifying the seven conditions. The expert then decides upon an evaluation scheme. If a qualitative evaluation is to be made, the expert only needs to compare all possible pairs of the seven items (making 21 comparisons) using the terms *better*, *worse*, or *equal*. If a quantitative evaluation is to be made, then the choice of scale, such as the one below designed by Saaty, is made. The Saaty scale is listed below with descriptions of the numerical evaluations for the comparisons:

Number	Description
1	The two items are of equal importance or likely
3	A slight favoring of the first item over the second
5	A strong favoring of the first item over the second
7	A demonstrated dominance of the first over the second
9	An absolute affirmation of the first over the second
2,4,6,8	These are used when compromise is needed
1/3, 1/5 1/7, 1/9	These values are used to indicate the above relationships when the first item is worse or less likely than the second

The expert begins comparing all possible pairs of conditions. The numerical comparisons are recorded in a 7 x 7 matrix of values with 1s on the diagonals and the comparisons in the upper triangular portion of the matrix. The lower portion is filled later with the reciprocal values of the upper portion. That is, if condition 3 is strongly more likely than condition 4, the value assigned in row 3 column 4 is 5 (from the scale). Then, the value for row 4, column 3 is 1/5.

When the comparisons are made and the matrix is completely filled, the relative weights of the seven conditions are obtained from matrix theory. Specifically, these weights are the normalized eigenvectors of the maximum eigenvalue of the 7 x 7 matrix. The reason why these weights are formulated in such a fashion may not be obvious; however, the mathematical theory behind it is sound.

Another advantage of using the Saaty method is its ability to monitor the consistency of the expert's evaluations. For instance, if an expert evaluates condition 1 versus condition 2 as a 4, and evaluates condition 2 versus condition 6 as a 3, and evaluates condition 1 versus 6 as a 1, then his three evaluations indicate an inconsistency. Using matrix theory, the Saaty technique provides an index of consistency for the comparisons made in a single matrix. The expert is warned when his consistency is lacking by a high value for this index of consistency. When this happens, the experts should re-examine the definitions and evaluations that were made and resolve the inconsistencies. Example 11.8 illustrates the pairwise comparisons, the relative weights and the inconsistency measures for the meteorological conditions described above.

EXAMPLE 11.8: ***Saaty's Pairwise Comparison Method or AHP***

The seven meteorological conditions important for affecting loss of off-site power (LOSP) in a reactor are as follows:

1. Flash flooding at plant site with 0.5 to 2 inches of water
2. Flash flooding with 2 to 4 inches
3. Flash flooding with more than 4 inches
4. Lightning (direct hit to power lines)
5. Direct hit by a tornado
6. Winds between 20 to 40 mph
7. Winds higher than 40 mph

By comparing these seven using all possible pairs (number of pairs = 7(7-1)/2 = 21), a set of relative weights can be found. The weights are interpreted according to the pairwise comparisons. In this case, comparisons are made by determining which of the pairs is more likely to cause LOSP. The pairs are evaluated using the Saaty scale listed above as follows:

1 vs 2----------1/3
1 vs 3----------1/4
1 vs 4----------1/5
1 vs 5----------1/5
1 vs 6----------1/3
1 vs 7----------1/4

2 vs 3---------- 1/2
2 vs 4---------- 1/3
2 vs 5---------- 1/3
2 vs 6---------- 1/2
2 vs 7---------- 1/3

3 vs 4----------1/2
3 vs 5----------1/2
3 vs 6----------1
3 vs 7----------1/2

4 vs 5---------- 1/3
4 vs 6---------- 1/5
4 vs 7---------- 1/4

5 vs 6----------5
5 vs 7----------4

6 vs 7---------- 1/2

These comparisons form the upper triangle of a matrix. The diagonal terms are 1s and the lower triangle contains the reciprocals of the upper triangle:

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	1	1/3	1/4	1/5	1/5	1/3	1/4
2	3	1	1/2	1/3	1/3	1/2	1/3
3	4	2	1	1/2	1/2	1	1/2
4	5	3	2	1	1/3	1/5	1/4
5	5	3	2	3	1	5	4
6	3	2	1	5	1/5	1	1/2
7	4	3	2	4	1/4	2	1

The principal eigenvalue of this matrix is 8.001. The weights for the seven factors are formed by normalizing (so that they sum to 1.0) the seven terms in the eigenvector for this eigenvalue. These normalized weights are

(0.03, 0.06, 0.11, 0.11, 0.35, 0.15, 0.19)

The Saaty method provides a consistency check in the form of a ratio value, called the consistency ratio, that indicates the deviation of the principal eigenvalue from the theoretical eigenvalue of a perfectly consistent matrix. The ratio is also adjusted for the number of factors, the dimension of the matrix. If a consistency ratio is greater than 0.10, inconsistency is indicated.

In this example the consistency ratio is 0.13, indicating some problems. Upon closer examination of the meteorological conditions comparisons, the following results are indicated:

1 vs 4 is the same as 1 vs 5
2 vs 4 is the same as 2 vs 5
3 vs 4 is the same as 3 vs 5
6 vs 4 is the same as 6 vs 5
7 vs 4 is the same as 7 vs 5

These results imply that 4 and 5 are the same; however, 4 vs 5 is given as 1/3. Also, 6 < 7, but examining 4 vs 6 and 4 vs 7 indicates that 6 > 7. There may be other minor inconsistencies in the magnitudes of the relationships; however, three major corrections are made as follows:

Comparisons	Correction	Objective
4 vs 5	1	To make 4 and 5 the same
4 vs 6	5	To match 5 vs 6
4 vs 7	4	To match 5 vs 7

Now the consistency ratio becomes much more acceptable at 0.06. The weights become

(0.03, 0.07, 0.11, 0.28, 0.28, 0.08, 0.15)

The interpretation of these weights indicates only relative comparisons. Direct hit lightning (4) and tornado (5) are the most likely to cause LOSP. The least likely is the flooding with 0.5 to 2 inches of water (1). It is incorrect to draw conclusions based on the numerical values of the weights; such as, flooding with 2 to 4 inches; (2) is only one-fourth as likely as a tornado (5).

■

The example problem above consisted of a single matrix evaluation. As the name AHP implies, most problems using this technique are hierarchical in structure. In the above example, loss of off-site power may be one of many plant conditions that are of critical concern to operations. Another matrix could be formed comparing all such critical concerns. For each of those other critical concerns, a matrix of meteorological conditions could be attached. These condition matrices do not have to be identical to the one for the loss of off-site power concern. Thus, an entire hierarchy of as many levels as are needed can be constructed. Usually, the hierarchy is constructed from the top down, with the top levels being the more general environmental or scenario factors. The middle levels are usually the more specific criterion or characteristics under consideration. The bottom level is usually the list of competing alternative decisions, actions, or choices that must be decided upon to answer the question.

Each matrix is evaluated at each level resulting in a set of relative weights. The weights are multiplied down the levels to form a final set of weights for the bottom level

items. This final set of weights is then used to make the decisions regarding the choices of the bottom level items. The higher the weights, the more desirable that item is.

There are many codes and packages that perform this technique at various levels of user interaction. Some codes merely provide the algorithms for the technique; others take the user through the entire problem from the initial building of the hierarchy to the final set of weights (Booker, Bryson, McWilliams 1984). For general references on the technique, Saaty has two books that provide codes and instructions (Saaty 1980 and 1982). A FORTRAN user-interactive code for a single matrix evaluation taken from the second of these books by Saaty (Saaty 1982) is given in appendix A.

Advantages and disadvantages of Saaty's method

The main advantages of this technique are its ease of use for the expert, its ability to monitor consistency of the expert's evaluations, and its ease of quantifying highly qualitative information. These advantages make it suitable for use in expert judgment problems.

The major disadvantage is that for application the problem must be structured in a hierarchical formation. Incorporating feedback cycles and pathways other than straight up or down the hierarchy are difficult to implement. A single level structured problem can be used to avoid the hierarchy; however, a single level formation is usually an over-simplification of the problem.

Uses for Saaty's method

The primary use for the hierarchical design is in decision analysis problems. Here the weights are used to aid in a decision maker's choice of the competing alternatives at the bottom level of the hierarchy. Therefore, the major disadvantage in applying this technique in expert judgment problems is that usually expert judgment problems cannot be neatly formulated into a hierarchical structure. Also, this type of forced structure formation is not consistent with the data analysis philosophy and model formation advocated in this book. The analyses and models are suggested by the data, not the forced fitting of the data to the analyses and models chosen.

In expert judgment applications, the scale and **quantification** features of this method can be used as a chosen **response mode** and as a quantification technique, respectively. It is these limited uses that are the reason for introducing Saaty's method.

Descriptions and Uses of Bayesian Techniques

What is the Bayesian Philosophy?

There are two different statistical philosophies for analyzing data and for interpreting the roles of probability distributions. These two different approaches are the classical or frequentist approach and the Bayesian approach.

The classical statistical approach assumes that the data or sample is representative of the **population** (the universal set of possible values) for the random variable. It is common practice to characterize the population as a probability distribution with certain

features (mean, variance, range, percentiles, mode, median, etc) called parameters. The parameters are fixed but unknown quantities. The estimates of these parameters are values calculated from the sample (data), and these estimates are called statistics. For example, if the population is represented by a normal distribution, then a sample of 20 values randomly chosen from that normal will represent that population. The population mean was 2.0, but this is not known. The sample has a mean of 1.82 which is the best available estimate for that unknown population of 2.0. Using the sample mean (statistic) to draw conclusions about the population mean (parameter) is the process of inference that is further discussed in chapter 18.

The Bayesian approach of philosophy is different in interpretation. The population parameters are not fixed quantities. Instead, they follow probability distributions, called **prior distributions**, just as the random variables do. This prior distribution represents the state of knowledge or information about the parameter before the sample is taken. The sample (data) also forms a distribution called the likelihood which represents how likely it was for that sample to be taken from the population. After the sample is taken, the likelihood distribution can be combined with the prior distribution to form a final combined distribution called the **posterior distribution**. The posterior represents the combined state of knowledge or information from before and after the sample data is taken. The analytical tool (equation) used to perform this combination is Bayes Theorem. Hence, the mathematics, the approach, and the philosophy are all labeled as Bayesian.

The philosophy is a logical one. It is common to have information about the problem before any data or experiment is done. It makes sense to use all available information to draw conclusions. The Bayesian approach provides a method for doing just that: combining different sources of information. Application of the technique involves representing the previously known information as a prior distribution, gathering the sample data, and using Bayes Theorem to combine the distributions into the resulting posterior distribution. Bayes Theorem is as follows:

$$g(\theta|x) = f(x|\theta)\, g(\theta) / f(x)$$

where

$g(\theta|x)$ is the posterior distribution,

$f(x|\theta)$ is the likelihood or data distribution,

$g(\theta)$ is the prior distribution for the parameter θ, *and*

$f(x) = \int f(x|\theta)\, g(\theta)\, d\theta$ is the marginal distribution that can be considered as a normalizing constant in the denominator of the above theorem equation.

Therefore, the theorem can be stated as follows: the posterior is equal to the prior times the likelihood divided by the marginal, or the posterior is proportional to the prior times the likelihood (Martz and Waller 1982).

Advantages, Disadvantages, and Uses of Bayesian Methods

The major advantage of taking the Bayesian approach is that it provides a means for combining or pooling information from different sources. The philosophy of using all the available information is a logical and reasonable approach especially when information from a single source is sparse or lacking as in most reliability applications.

Some short examples illustrate the uses of Bayesian methods--a pooling mechanism: (1) Expert estimates could provide the information for the prior, and that could be combined with sparse data; (2) Expert estimates could be combined, one at a time, to form an aggregation estimate; (3) An expert aggregation estimate (prior) could be combined with information from a decision maker; (4) Older information from an expert (prior) could be combined with his new assessment to update his judgment in view of different conditions or information; (5) Generic data such as an overall failure rate of all check valves (prior) could be combined with data on a specific check valve; and (6) Uncertainties (prior) could be modeled with the data. Bayesian methods are suggested and discussed in more detail for aggregation (chapter 16), for characterizing uncertainties (chapter 17), and for updating (chapter 7).

The major disadvantage of Bayesian methods lies in the requirement of transforming all the available information, regardless of its source or form, into probability distributions. For qualitative data, this transformation is an especially difficult task. (Chapter 12 discusses ways of handling qualitative data.) Transformations may not be any easier for quantitative data. Once distributions are formed, the second disadvantage of Bayesian methods emerges. These various distributions are combined using Bayes Theorem. This combining may not be a mathematically easy task. However, with modern simulation techniques, using the theorem for combining distributions is not as difficult as it was a decade or so ago.

12

Initial Look at the Data--The First Analyses

After the elicitation is completed, the information gathered will seem like a large, complex mass of words and numbers. The first step is to become familiar with the information of this mass: that is, examine the data that has been gathered; then, focus on some important data features, investigate transforming the data or quantifying it, and formulate a data base for further analysis.

What Data Has Been Gathered?

The components of the elicitation, methods for formulating the questions, response modes and documentation, are described in chapter 7. Tailoring these schemes for the particular elicitation is described in chapter 8. Implementing these is described in chapter 10. Having implemented the chosen schemes for questions, response mode, and documentation, the post-elicitation information base should consist of large amounts of qualitative and quantitative information from each expert on each question. Following the documentation guidelines helps reduce some of the volume of information to a more compact and efficient form at chosen levels of detail. However, this is not much help to the analyst faced with the qualitative/quantitative data mixture containing a potentially large number of variables.

The information gathered at the post-elicitation stage consists of two major parts: (1) the answers to the the questions, and (2) the ancillary information. This ancillary information is in two groups: (1) the information about the expert such as his background and experience, and (2) the information called **expert data** in chapter 1 about how the expert solved/answered that question and how long since he had seen such a problem.

The qualitative or quantitative structure of the answer data and the ancillary data depends on the choices of the response mode and the documentation. Usually the answer data is quantitative and the ancillary data is a mixture. Regardless of the original structure of either, some quantification of some of the qualitative information becomes necessary for analysis. One criterion for determining the necessity and method of quantification is to consider the level of generality or granularity of all the data and of the analyses.

Overview of the Data

In the process of analyzing the data, often two important overall features of the data are taken for granted and therefore forgotten by the analyst. These two features of the data set are important to the analyst at all stages of the analysis and have important effects on the conclusions reached. The two features are **granularity** and **conditionality**.

Granularity is the level of detail defined or chosen for the data, the analysis, and the conclusions. Two examples of the information recorded on an expert's problem-solving may be (1) in the form of detailed steps, equations, heuristics, definitions, and descriptions, or (2) in the form of a general categorization of this problem-solving stating simply that the expert used a pessimistic approach. The above are two different granularities for the information regarding the expert's problem-solving process.

Conditionality refers to the inescapable fact that all of the elicited data is conditioned on many other factors. Some of these factors are controlled, some are not controlled, and some are unknown.

The *Pitfalls* section in chapter 2 discusses the importance of these two features for expert judgment applications in more detail. There are sections on granularity and conditionality in all the chapters dealing with the analysis of the data.

Establishing Granularity

To some extent, the granularity will have already been established in the selections of the response mode and documentation recording schemes. The granularities chosen for each could be different, and the analyses can have a third level or even more. However, for interpreting the results and drawing conclusions, one level, the most general of all, must prevail. That level is the only one applicable to the results and conclusions. Therefore, it is wise to establish that one desired level of detail in the initial planning phase of the study before the elicitation. If that is not possible, then at least establish the level at the preanalysis phase and use it throughout all the analysis steps. Otherwise, analyses may have to be repeated at the proper granularity.

If levels are mixed in the analyses, conclusions can change. For example, comparisons of variables at the granularity gathered, raw form could reveal some significant correlations among the variables that vanish if they are compared after being combined or reduced to a more general level. Of course, such a combining or collapsing process might not change the significant results among the variables from the raw form, but it could produce significance where none existed in raw form or it could lose significance where it existed in raw form. The effects on the results of changing granularity is not known beforehand.

In the chapters that follow, many stages of data analysis are described. In each, the granularity is important. At each stage, the results can change if the granularities are changed. Examining how results can change with different granularities within each stage and across stages is an exercise consistent with the spirit of investigative data analysis advocated by this book. However, great care must be taken not to confuse such an academic exercise with the goal of determining conclusions for the problem at hand. There is another point to consider when playing with granularities. It is possible to take detailed

information and make it more general, but it is not possible to do the reverse. To avoid confusion and problems, it is recommended that one granularity be chosen and used throughout all analysis stages.

Establishing Conditionality

Often the conditional structure of the data, especially the answer data, is ignored in the analysis and in the conclusions. Disregarding the conditional structure of the data produces conclusions that are a mixture of differing effects, or more simply, a mixture of apples and oranges. Most analysts would agree that such conclusions are worthless. Indeed, this lack of care in analysis may be the reason why many do not trust expert judgment data or claim that expert information can not be analyzed.

A recent example will help illustrate the problems in dealing with conditionality. In an effort to revise the probabilistic risk assessment (PRA) methodology for nuclear reactors, the Nuclear Regulatory Commission has invested time and money in the NUREG-1150 project (U.S. NRC 1989). As part of this task, several panels of the world's top experts were gathered for eliciting their data on many rare, and undefined events affecting reactor safety. These events were decomposed into decision-type event trees; decision trees are briefly described in chapter 15 and also by Raiffa (1970). The tree structures and probabilities for each branch were elicited from the experts. The final answers came from multiplying these probabilities through the tree. Each answer is therefore conditioned on the tree and its estimates. This conditioning cannot be ignored. Two experts could arrive at exactly the same final answer but for very different reasons, or two experts could arrive at different answers for exactly the same reasons.

Analyses, such as the ones described in this chapter and in chapters 13-15 are needed to determine what effects, if any, such conditioning has on the final answers. If the answers are not dependent upon the conditions, then conditioning can be ignored; however, this determination is necessary before setting conditionality aside.

How to Quantify

Quantification can be useful for preparing the complex post-elicitation data set for the data base. Both the qualitative data and the quantitative data may require transformations and cleanup for the data base. Transforming words such as descriptions or preference scales (*worst, worse, bad, neutral, good, better, best*) into numerical values will often be required for analysis of the data using numerically based techniques (such as statistical techniques). Also, some data that is already numerical in raw form may require additional numerical transformation to more convenient scales or to the chosen granularity. In both cases, the transformation process is quantification, transforming the raw information into a desired numerical form.

Several commonly used methods of quantification are described below. The major problem with quantification is to not impose additional assumptions about the information in order to *fit* the data into the desired form. In most instances this is difficult or impossible

to avoid. However, sometimes other available information elicited from the experts can aid in the quantification process as illustrated in example 12.1.

EXAMPLE 12.1: ***Using Definitions to Quantify***

The analyst is attempting to convert a statement of the expert's preference about how good a reactor system design might be. The analyst can refer back to the expert's elicited definition of *good* and use that definition to compare to definitions from other experts. This comparison can form a consistent numerical scale across experts as follows:

Expert's Definition of *Good*	Numerical Scale Value
System functions outside specifications at all times	5
System functions within specifications at all times	3
System functions within specifications 90% of the time	1

At this stage of the analysis, the reasons for eliciting information from the experts about definitions, assumptions, and problem-solving processes become obvious.

■

When Is It Necessary to Quantify?

Ideally, all the information gathered during the elicitation should be quantified to a common numerical scale for comparisons using statistical analyses. However, this form of quantification is not feasible, nor is it entirely necessary. Many times information gathered is redundant. The expert will state the same information repeatedly in different forms as illustrated in example 12.2.

EXAMPLE 12.2: ***Detecting Redundant Information***

An expert gives a lengthy explanation of a physical phenomenon. Five minutes later, he realizes that he was simply applying a basic principle or law. The information provided by the expert and the usage of that information has not changed. The expert has just given the same information in two different forms. Many times the expert does not realize this redundancy, but the analyst can find it in the course of his analysis if proper documentation was done.

■

The first items that need quantification are the answers to the technical question. In most instances the answers will already be in the desired numerical form from the chosen response mode designed in the elicitation. (See chapter 7, *Selecting from Response Modes* and *Selecting from Dispersion Measures*.).

It may be difficult to quantify assumptions, definitions, and problem-solving processes initially. Yet, some assumptions about physical quantities such as temperature are easily converted to a scale of values or ranges of values. Ranges of values should not be reduced to a single value. The process of such a reduction imposes assumptions on the

involving problem-solving information, experts' background information, and expert answers.

EXAMPLE 12.3: ***Dichotomous Quantification***

Experts are asked if they applied the first law of thermodynamics. To quantify the simple *yes* or *no* responses, set *yes* = 1 and *no* = 0.

Experts are asked if they consider themselves engineers or not. To quantify the responses, set *engineer* = 1 and *nonengineer* = 0.

Experts are asked if the probability of an event is greater than 0.001 or less than 0.001. To quantify the responses, set greater than 0.001 = 1 and less than 0.001 = -1.

■

Rank or rating quantification

If there are more than two choices for representing some information, multiple integer or numbered values can be used. The values can be in ascending or descending order (ranks) or the values can be chosen from a specified scale. In either case, these values reflect an ordering of the information and should not be used unless the information has a logical ordering. The order implied by ranks is linear, implying equal spacings between the ranks and relationships, such as a rank of 4 is twice a rank of 2. Example 12.4 illustrates the proper use of ranks.

EXAMPLE 12.4: ***Rank Quantification***

In gathering background information, the experts are asked if they have had any reactor operator experience. The responses to that question are given in verbal terms such as none, some, and extensive. The ranks 0, 1 and 2 can be assigned to the answers none, some, and extensive. The ascending order implied by the ranks is logical.

The experts are also asked to describe their major discipline area. The responses are given in terms such as nuclear engineer, civil engineer, mechanical engineer, physicist, and mathematician. The ranks 1, 2, 3, 4, and 5 should *not* be assigned to these answers because they do not have a logical order and should not be given an order through the use of ranks.

■

Number line quantification

As mentioned earlier, accuracy of information content is important in the transformation process, especially for transformations to the continuous number line. Accuracy is an issue related to granularity. It is usually considered when determining how many significant digits can be used to represent the information. The number of significant digits used in the analysis and in the results is a granularity issue.

original data by the analyst that the expert might not have had in mind. In general, the assumptions, definitions, and problem-solving information should be kept in raw form until the modeling stage of analysis (chapter 15).

The ancillary information or data, such as data on the expert's background and experience, also needs quantification. This data can be in many different forms and usually has been elicited without much advanced planning or designing of the analysis. Therefore, this data may range from completely descriptive information to strictly numeric values and to everything in between.

Quantification Schemes

The application of the following quantification schemes should be done with extreme caution. It is so easy for the analyst to impose assumptions on the data to make it *fit* into the desired quantification scheme. The higher the degree of qualitative structure, the more such assumptions are required to transform the raw information to numbers useful for standard analysis techniques. The *Pitfalls* section in chapter 2 discusses in more detail the interviewers, analysts, and knowledge engineers as sources of this bias.

The following methods of quantification cover ways of transforming qualitative information to quantitative information. Examples of application are included in each method.

Dummy variables

This method is the one most people think about when dealing with quantification. The raw data is transformed into artificial or dummy numerical values. Many times the transformations are done without proper logic or thought. Examples of this transformation follow:

1. Transforming information where only two options are available into dichotomous values such as (0,1) or (-1,1).
2. Transforming information to integer values such as in the use of scales or ranks (1, 2, 3, etc.).
3. Transforming information to the real number line, a continuum of values for a specified interval (e.g., 0.0, 0.1, 0.15, 0.27, 0.96 in the [0,1] interval). The choice depends on the information gathered, its potential use, and its accuracy (the number of significant digits).

The reason for transforming to number values is for use in both the data base formation (discussed later in this chapter) and for the modeling process. Numerical variables are easy to analyze in most statistical and analytic procedures and are desirable for that reason. The appropriate formation of the various types of dummy variables is given in more detail in the following sections.

Dichotomous quantification

Most information, whether qualitative or quantitative, can be transformed to a 0,1 or -1,1 dichotomy (two choices) with little or no assumptions required for the transformation. Example 12.3 illustrates dichotomous quantification for the cases

EXAMPLE 12.5: ***Significant Digits***

Experts are asked to provide the number of years experience that they have had in a field. The number of digits offered by the experts will differ across experts. Some say *about* 2 years and mean greater than 1 year but less than 3 years. This is only one significant digit. Some said 2 years and 6 months or 2 1/2 years. This is 2 significant digits. Some say 2 years and mean exactly 2.0 years (2 digits) or mean *about* 2 years (1 digit). Because the number of digits differs across experts, the most general level of detail must be used for all experts. In this case that means the lowest number of significant digits (1) would be used for all experts.

Using only 1 significant digit results in a loss of information at the finer level of detail offered by some of the experts. To avoid this loss, the elicitation of the information should be more thorough. The first and second experts should be queried for the desired level of detail using the verbal probe or ethnographic methods described in chapter 7. The elicitation process can help guarantee that the level of information content is consistent among experts, thereby minimizing the problems with quantification.

■

Proper elicitation planning and execution also involves understanding why this information is being gathered and what potential use it will be in the analysis. Knowing this, the analyst should make sure that the the information is being elicited at the desired granularity (e.g., number of significant digits) rather than getting mixed levels of detail and having to transform information to some other level in the post-elicitation phase.

Number-line quantification can be used to combine information from two or more variables, provided these variables have a common basis of accuracy. One such application could be the following:

EXAMPLE 12.6: ***Combining Number Line Quantifications***

Experts were asked how much thermodynamics training and experience each had. The answers were as follows:

Expert No	School Training (yrs)	Job Experience (yrs)
1	1.50	0.00
2	0.50	5.50
3	0.00	2.33
4	2.25	3.10

If training is only half as valued as job experience, then the results would be

1	0.50(1.50) + 1.00(0.00) = 0.75 years
2	0.50(0.50) + 1.00(5.50) = 5.75 years
3	0.50(0.00) + 1.00(2.33) = 2.33 years
4	0.50(2.25) + 1.00(3.10) = 4.23 years

It is not recommended that information of a more descriptive structure (words) be transformed onto the real, continuous number line. Only information gathered in a continuous, numerical form should be treated in this way.

■

Ordinal ranks

Ordinal ranks are usually used on descriptive information (nonnumeric) that has some relative ordering. The rank values assigned are also descriptive (words) in nature. The variables formed using this quantification can be incorporated into the data base and can be used in many of the analytic procedures for such analyses as correlation detection and for understanding the conditional nature of the data base.

Because the ranks are relative comparisons, care must be taken regarding the use of definitions to achieve and maintain consistency of application. Assuming that the information from the experts was elicited in the proper fashion, clarifications are available in the documentation to help the analyst or decision maker assign relative comparisons for certain quantities.

EXAMPLE 12.7: ***Assigning Ordinal Ranks***

In solving a problem, most experts use a basic principle from thermodynamics but in varying degrees of emphasis of use. This information can be added to the data base using the following ordinal rank variable:

Expert No.	Descriptive Use	Rank
1	Did not use the principle at all	1
2	Extensively used the principle	6
3	Only mentioned the principle	2
4	Used the principle a few times	4
5	Used the principle once	3
6	Used the principle several times	5

■

The Saaty pairwise comparison technique (chapter 11) is a way of making relative comparisons using the technique's own quantification to a numerical scale. However, these resulting numbers or weights can only be interpreted in a relative sense because the information used in the method is only relative comparisons. A resulting relative weight from this technique of 0.25 cannot be interpreted as half as good as a weight of 0.50. The relative interpretation is that the value 0.25 is less important or less preferred than 0.50. This technique is best used as an elicitation method for obtaining responses from the experts. However, it can be used as a quantification scheme for post-elicited data. One major advantage of this method is that the relative comparisons are made in pairs and do not have to be made simultaneously.

EXAMPLE 12.8: ***Ordinal Ranks From Pairwise Comparisons***

Of the five experts solving a problem, some used a rule of thumb or a modification of that rule. The pairwise comparisons on the usage of this rule are determined by answering the following question: Did expert *i* apply the rule more completely than expert *j*? The answers follow:

Expert *i*	Expert *j*	Comparison
1	2	Same
1	3	Yes
1	4	No
1	5	No
2	3	Yes
2	4	No
2	5	No
3	4	No
3	5	No
4	5	Yes

Using Saaty's method the resulting relative weights for these 5 experts follow:

Expert	Weight
1	0.138
2	0.138
3	0.079
4	0.387
5	0.257

These weights indicate that expert 4 applied the rule more completely than the others. Expert 3 applied the rule less completely than any of the others. Experts 1 and 2 applied the rule in the same manner. No further interpretation is possible with such a relative comparison.

■

Categorical variables

In cases where qualitative information cannot be ranked or ordered in preference, the information can be transformed and stored into groups or categories, usually according to verbal descriptions. It is not recommended that these verbal categories be coded into an arbitrary numerical code (a dummy variable). When this is done there is a great temptation to analyze the numeric data as if the numbers reflected ranks or orderings. Most modern software can easily handle character (word) information as a part of the information data base.

Educational information on the experts are usually put into the data base as categorical or classification variables. Degree titles such as BS, MS, MBA, etc. are

examples of these categories. Degree disciplines such as civil engineering, mechanical engineering, nuclear physics, and thermodynamics are other examples.

During some of the later analyses, categories or classes can be consolidated or collapsed in later analysis stages if there are too many classes and not many experts (less than 3) in each . Such collapsing does change granularity from specific to more general.

EXAMPLE 12.9: ***Collapsing Categories***

For example, if there are 10 experts with the following disciplines in their highest degree, these 10 different disciplines might be collapsed into 3 disciplines:

Expert	Elicited Discipline	Transformed Discipline
1	Mechanical engineering	Engineering
2	Nuclear engineering	Engineering
3	Thermodynamics	Engineering
4	Hydrodynamics	Physics
5	Material science	Physics
6	Nuclear physics	Physics
7	Mechanics	Engineering
8	Computer science	Computation
9	Simulation science	Computation
10	Knowledge engineering	Computation

■

The definitions and rationale for the above transformation should be recorded and consistently applied throughout the study. The above process is one of changing the granularity of the discipline information from a more detailed description to a broader one. The results of any analysis done on the collapsed version will only be valid for the more general categorization. For this reason, it is recommended that collapsing be done only during the later analysis stages and that the categorical data be stored untransformed in the data base.

Description Variables

It may not be possible to quantify some of the information gathered. This information should be condensed to as few words as possible and kept for further analysis uses.

Descriptions of how the experts solved problems can be analyzed as conditioning variables or can be used in model formation (chapter 15). Usually this information is not easily translated into numbers or even categories without making some assumptions or without changing the level of detail. At this initial analysis stage, neither is recommended.

Description variables such as the definitions and assumptions used by experts have another important use. These variables need to be retained for documentation purposes and for purposes of updating the experts' answers if new information is considered (chapter 10).

Forming a Data Base of Information

Once the quantification steps from the previous section are done, the information can be placed in a data base. This data base could be a computer file or a paper listing of all the information gathered about each expert including their answers. A suggested list follows:

Expert ID	Name and Number
Interview information	Date, time, place, duration, environment
Expert's background:	
Education	Degrees, dates, schools, disciplines
Experience	Years, organizations, colleagues, nature and type of work
Expert's problem solving	Definitions, assumptions, steps, cues, heuristics
Answers	Values and comments

Each expert will then have many quantities (variables) associated with him. Some information could be missing for some experts. A code word or number is needed to denote missing information (e.g., *miss*), and one is also needed for nonapplicable information (e.g., *na*) to distinguish these from 0 values. Many statistical and data-base packages have their own designations for missing information.

If the guidelines for tailoring the elicitation in chapter 8 were followed, the data gathered will reflect the design chosen, and the reasons for the choices are already documented. There is little that needs to be done to the data base to conform to the elicitation method used.

Because the data base and the elicitation method are so closely connected, the monitoring and use of granularity is important. The results and conclusions that will be found from the data base should be either at the same or at a more general level of detail than the information in the data base. In the analyses suggested in the next chapters, granularity is continuously monitored. Examples are given where the results change if the granularity changes. It is important at this data-base stage to be aware of the detail of the information content in the data base. A quick review of the variables and information in the data base is usually sufficient. This review can be aided by listing each answer in ascending order and listing other variables beside the answers for each expert.

During this review, the analyst can get some ideas about model formation (chapter 15). If many variables have missing values, more general-leveled models are suggested. If the information is complete in the data base, models at the current granularity can be formed and analyzed.

Also during this review, the analyst can see which and how many variables need testing for possible conditioning effects on the answers. In addition, some variables may require testing as sources of correlation or bias among the experts (chapter 14).

If many variables are in the data base (more than 20), there is a strong possibility that information is being repeated among the variables. Many statistical techniques (from

chapter 11) can be used to monitor for this redundancy of information among the variables. In some cases, variables can be eliminated from the data base. These steps are presented in the next chapter (chapter 13) on understanding the data base.

13

Understanding the Data Base Structure

In this chapter analyses are suggested to gain understanding of the relationships existing among the ancillary variables, among the answer (response) variables, and between these two sets of variables. The information and knowledge gained from the results of these analyses are merely for understanding the data base, are not to be considered as the final results, and conclusions should not be drawn from them. The analyses suggested are standard, statistical techniques, most of which require assumptions about the data that would not necessarily hold under expert judgment applications. Instead, these techniques are suggested as tools for understanding variate relationships and are not to be used to determine final or significant results. The use of several techniques is suggested for the purpose of cross-verification of suspected relationships among the variates.

The words of caution regarding the use of the statistical techniques presented below are serious words. It is very uncomfortable for the authors to recommend applying a technique when there is good reason to believe that assumptions required for its use are being violated. It is also very difficult to recommend using a technique and then strongly urging not to rely on the results. These statistical tools are used to explore variate relationships in the data base. If the results from these analyses do not make sense, or if the results of one test contradict results from another, there could be a very good reason; namely, the techniques were not used properly. We strongly advise using these techniques with the help of a statistician.

The analyses presented in this chapter begin with the investigation of potential relationships between the ancillary variables and the answers. Then a separate analysis of the answer data is presented for two purposes: (1) to present analyses of the answers when there are no suspected conditional effects, and (2) to gain additional information about the answer variables. The separate analysis of the ancillary data is also provided. Finally analysis techniques are given for analyzing the ancillary data with the answer data.

The analysis techniques used are commonly found on statistical and data analysis software packages. The software used for most of the examples is the SAS® product.

Conditionality--Examining Relationships Between Answers and Ancillary Data

All of the variables formed from the ancillary data have the potential of being **conditional variables** that could have affected the answers given by the experts. This relationship is called conditionality. Some very basic statistical tests and graphic procedures can be used to begin the investigation of any potential relationships. These are referred to as bivariate analyses where each ancillary variable is checked against each answer variable. The multivariate investigative procedures are described in the later section on *Analyzing the Ancillary Data with the Answer Data*.

Correlations

The easiest starting place for bivariate analysis of the ancillary data and the answer data is to calculate Pearson pairwise correlation coefficients for all pairs of numerical ancillary variables and numerical answer variables. Most statistical and data analysis packages have correlation routines. If none are available, the formula in chapter 11 can be used.

In order to determine if any of the correlations indicate potential relationships between the pairs of variables, a significance level must be specified. If there are a total of n pairwise correlations calculated, the significance level used to determine if any pair is correlated should be $< 1/n$ for $n > 20$ or the customary values of 0.05 or 0.01. Example 13.1 illustrates this determination for a large number of experts.

EXAMPLE 13.1: ***Correlations and Significance Level***

The following 10 correlation coefficients were calculated for 31 experts answering two questions (Q_1 and Q_2) compared to five ancillary variables from their background (*Ys* = years since they worked on this type of problem; *Ya* = years worked as an assessor; *Yp* = years worked in applications; *Yd* = years worked as a developer; and *Yn* = years worked on documentation):

	Ancillary Variables				
Correlation (level)	*Ys*	*Ya*	*Yp*	*Yd*	*Yn*
Q_1	0.095	-0.484	0.167	0.153	0.035
	(0.61)	(0.006)	(0.37)	(0.41)	(0.85)
Q_2	0.322	-0.452	-0.095	-0.089	-0.309
	(0.08)	(0.011)	(0.61)	(0.63)	(0.09)

The significance levels are listed in parentheses below the correlations. A significance level of 0.05 or 0.01 is indicated because the number of correlations calculated is less than 20. A conservative (minimizing the chances of an incorrect conclusion) level of significance for this example would be 0.01. Any correlation whose level is less than 0.01

would be considered significant, and a relationship would be suspected between those two variables. In this case, only the Ya / Q_1 relationship is indicated. The conclusion is that large values of Ya correspond to small values of Q_1 (because the correlation coefficient is a negative value.).

■

At this point in the analysis, all the significant relationships should be recorded along with any possible reasons or explanations. The ancillary variables that are significantly correlated to the answers are the beginning of a list of potential conditional variables. However, these significant relationships may not hold when the multivariate analyses are performed later. Nevertheless, for now, they offer some understanding about the data base, the reasons for the experts' answers, and some directions for future analyses.

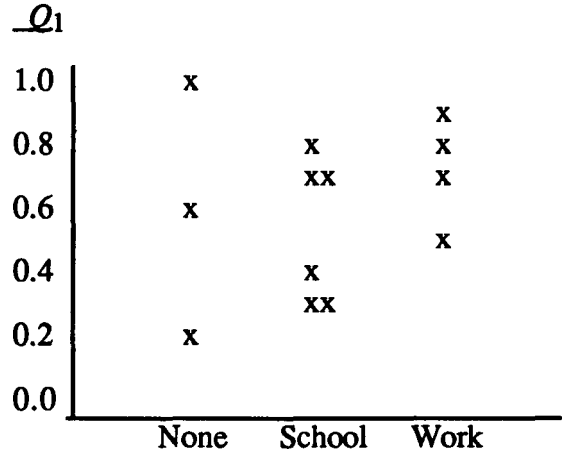
Graphs

The correlations can only be done for numerical variables. Graphs can be used to plot potential bivariate relationships between the ancillary variables and the answers. Graphs can also indicate nonlinear relationships; whereas correlation analysis is only good for linear trend detection.

To graph qualitative or categorical data, equally spaced intervals may be used on the axis using the ordering or ranking inherent in the categories. If there is some reason for using unequal spacing, that should also be tried. If nonlinear relationships are indicated, transformations of the data by taking logarithms can sometimes produce linear relationships. Example 13.2 illustrates that such a graph of categorical data can be done.

EXAMPLE 13.2: ***Graph of an Ancillary Variable and an Answer Variable***

The following is an example of a graph of a categorical variable describing the reactor experience of 13 experts versus the answer variable Q_1. The three categories of the experience variable indicate an order of importance. That ordering is used to place the categories on the axis.



From the above graph, no linear or nonlinear relationships between the two variables are indicated.

■

Plots like the one in example 13.2 should be done on all possible pairs of the ancillary variables with answer variables. If the graphs indicate a relationship between the variables, then the ancillary variable is added to the list of potential conditional variables. Any explanations or reasons for the relationship should also be kept with the list.

As mentioned above, additional analyses (e.g., multivariate analyses) will be used to add to and to change the list of the potential conditional variables. Prior to these analyses, some basic investigation of the answer data is useful especially if conditionality does not appear to be a problem.

Analyzing the Answer Data

The answers to the technical questions are the prime reason for electing expert judgment. It has been emphasized that these responses can be highly conditioned on other (ancillary) information such as the experts' problem-solving processes and response environment. It has also been emphasized that any scientific investigation must be done with the understanding of the granularity used at the various stages of the problem: the information gathering, analysis, and conclusions. These issues, granularity and conditionality, are considered here in the analysis of the answer data by examining the between/within variance and multimodal structures of the answers.

Investigating Multimodality

Empirical evidence has shown that the answers given by multiple experts form a multimodal distribution (Booker and Meyer 1988a; Meyer and Booker 1987a; Baecher 1979). This multimodality was partially responsible for the widespread belief that experts must be correlated or dependent upon one another. Therefore, the explanation of the modes or clusters of their answers reflected membership of the experts into groups based on common backgrounds, educations, or experiences. Until recently this belief and the reasons for these clusters had not been investigated (Meyer and Booker 1987a).

In many cases the number of distinctive modes formed by the data will be obvious to the eye. For example, a bimodal case with one mode at high values and the other at low values with a gap in the midrange obviously splits the data set into two groups or clusters (as seen in example 13.3). However, some cases may not indicate such obvious groupings or clusterings. For these cases, a formal cluster analysis is useful for determining the structure of possible groupings.

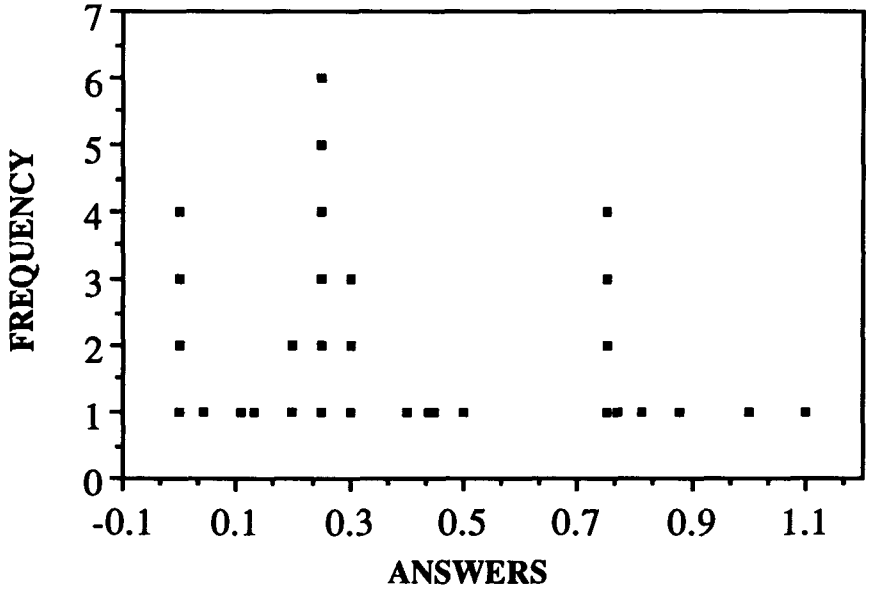
As described in chapter 11, cluster analysis forms the clusters from a data set according to some distance criteria separating the individual data points. Most cluster analysis programs have several options available for determining this distance criteria, the most commonly used is the centroid method. The results of the cluster analysis can change depending upon the method chosen. In any case, the interpretation of the cluster analysis

results are left to the user. Most cluster programs printout results according to a hierarchical clustering scheme where clusters are formed beginning with one cluster containing all the data and ending with clusters containing only one data point each. The user must decide which of these possible cluster formations to use to characterize the data set.

There are ways that the user can decide on which cluster structure to use. The analytically based way is to choose a cluster formation that shows the largest change in distances between clusters (examples 13.3 and 13.4). Example 13.3 is a frequency data plot of 31 experts' answers to a reactor safety question with a continuous response scale from 0.0 to 1.1 (Meyer and Booker 1987a). At first glance, the data appears bimodal in nature. The results of a formal cluster analysis using the centroid method on SAS® are given in example 13.4. This graph shows the distance between the cluster centers plotted for the different numbers of clusters formed (the different cluster formations). The detailed clusters are also given for all possible cluster formations.

EXAMPLE 13.3: ***The Frequency Plot of a Raw Data Set***

Thirty-one expert responses to one question are plotted below. The responses were elicited on a continuous scale from 0 to 1, where a 1 was considered the highest likelihood. One expert (value assigned as 1.1) felt that the event was even more certain than the likelihoods presented on the scale. Of course, these values could have been transformed from a 0 to 1.1 scale to a 0 to 1.0 scale for analysis.



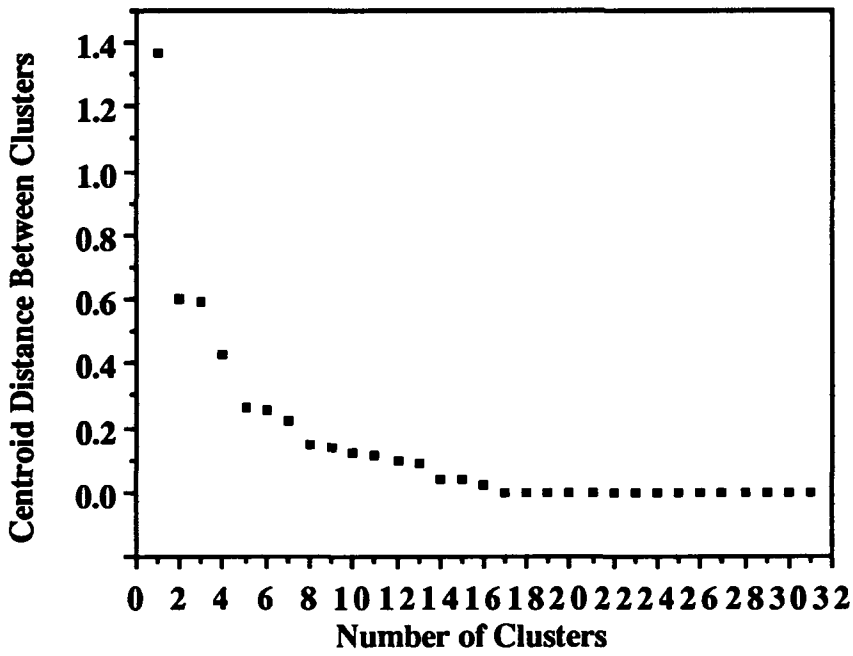
The data form two distinctive groups or distributions. The formal cluster analysis should also indicate these two major clusters.

■

EXAMPLE 13.4: ***Cluster Analysis Graph***

The following graph indicates the distances between all possible cluster formations for the 31 responses to one question. The distances are based on the centroid method of cluster analysis.

Cluster formations range from 17 clusters where each value given by an expert forms its own cluster to 1 cluster where all 31 observations form a cluster. (Some experts gave the same value, making only 17 distinctive values in the data set.) From the plot, there is a dramatic change in cluster distances with the two-cluster formation. The next major breaks in cluster distances occur at the four- and five-cluster formations. By the time that 17 clusters are formed, the distance measure is at zero value.



The following table gives the values from the experts belonging to each cluster formation where individual clusters are marked by square brackets:

No. of Clusters	Cluster Formations with Members
1	[0.0, 0.0, 0.0, 0.0, 0.04, 0.11, 0.13, 0.20, 0.20, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30, 0.30, 0.30, 0.40, 0.44, 0.45, 0.50, 0.75, 0.75, 0.75, 0.75, 0.77, 0.81, 0.88, 1.0, 1.1]
2	[0.0, 0.0, 0.0, 0.0, 0.04, 0.11, 0.13, 0.20, 0.20, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30, 0.30, 0.30, 0.40, 0.44, 0.45, 0.50] [0.75, 0.75, 0.75, 0.75, 0.77, 0.81, 0.88, 1.0, 1.1]

3 [0.0, 0.0, 0.0, 0.0, 0.04, 0.11, 0.13, 0.20, 0.20,
0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30, 0.30, 0.30,
0.40, 0.44, 0.45, 0.50]
[0.75, 0.75, 0.75, 0.75, 0.77, 0.81, 0.88] [1.0, 1.1]

4 [0.0, 0.0, 0.0, 0.0, 0.04, 0.11, 0.13]
[0.20, 0.20, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30,
0.30, 0.30, 0.40, 0.44, 0.45, 0.50]
[0.75, 0.75, 0.75, 0.75, 0.77, 0.81, 0.88] [1.0, 1.1]

5 [0.0, 0.0, 0.0, 0.0, 0.04, 0.11, 0.13] [0.20, 0.20,
0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30, 0.30, 0.30]
[0.40, 0.44, 0.45, 0.50] [0.75, 0.75, 0.75, 0.75,
0.77, 0.81, 0.88] [1.0, 1.1]

6 [0.0, 0.0, 0.0, 0.0, 0.04, 0.11, 0.13] [0.20, 0.20,
0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30, 0.30, 0.30]
[0.40, 0.44, 0.45, 0.50]
[0.75, 0.75, 0.75, 0.75, 0.77, 0.81] [0.88] [1.0, 1.1]

7 [0.0, 0.0, 0.0, 0.0, 0.04] [0.11, 0.13] [0.20, 0.20,
0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30, 0.30, 0.30]
[0.40, 0.44, 0.45, 0.50]
[0.75, 0.75, 0.75, 0.75, 0.77, 0.81] [0.88] [1.0, 1.1]

8 [0.0, 0.0, 0.0, 0.0, 0.04] [0.11, 0.13] [0.20, 0.20,
0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30, 0.30, 0.30]
[0.40, 0.44, 0.45, 0.50]
[0.75, 0.75, 0.75, 0.75, 0.77, 0.81] [0.88] [1.0] [1.1]

9 [0.0, 0.0, 0.0, 0.0, 0.04] [0.11, 0.13] [0.20, 0.20,
0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.30, 0.30, 0.30]
[0.40, 0.44, 0.45] [0.50]
[0.75, 0.75, 0.75, 0.75, 0.77, 0.81] [0.88] [1.0] [1.1]

10 [0.0, 0.0, 0.0, 0.0, 0.04] [0.11, 0.13] [0.20, 0.20,
0.25, 0.25, 0.25, 0.25, 0.25, 0.25] [0.30, 0.30, 0.30]
[0.40, 0.44, 0.45] [0.50]
[0.75, 0.75, 0.75, 0.75, 0.77, 0.81] [0.88] [1.0] [1.1]

11 [0.0, 0.0, 0.0, 0.0, 0.04] [0.11, 0.13] [0.20, 0.20,
0.25, 0.25, 0.25, 0.25, 0.25, 0.25] [0.30, 0.30, 0.30]
[0.40, 0.44, 0.45] [0.50] [0.75, 0.75, 0.75, 0.75,
0.77] [0.81] [0.88] [1.0] [1.1]

12 [0.0, 0.0, 0.0, 0.0, 0.04] [0.11, 0.13] [0.20, 0.20]
[0.25, 0.25, 0.25, 0.25, 0.25, 0.25] [0.30, 0.30, 0.30]
[0.40, 0.44, 0.45] [0.50] [0.75, 0.75, 0.75, 0.75,
0.77] [0.81] [0.88] [1.0] [1.1]

13 [0.0, 0.0, 0.0, 0.0, 0.04] [0.11, 0.13] [0.20, 0.20]
[0.25, 0.25, 0.25, 0.25, 0.25, 0.25] [0.30, 0.30, 0.30]

	[0.40] [0.44, 0.45] [0.50] [0.75, 0.75, 0.75, 0.75, 0.77] [0.81] [0.88] [1.0] [1.1]
14	[0.0, 0.0, 0.0, 0.0] [0.04] [0.11, 0.13] [0.20, 0.20] [0.25, 0.25, 0.25, 0.25, 0.25, 0.25] [0.30, 0.30, 0.30] [0.40] [0.44, 0.45] [0.50] [0.75, 0.75, 0.75, 0.75, 0.77] [0.81] [0.88] [1.0] [1.1]
15	(The distance from 14 to 15 clusters was the same.)
16	[0.0, 0.0, 0.0, 0.0] [0.04] [0.11] [0.13] [0.20, 0.20] [0.25, 0.25, 0.25, 0.25, 0.25, 0.25] [0.30, 0.30, 0.30] [0.40] [0.44, 0.45] [0.50] [0.75, 0.75, 0.75, 0.75] [0.77] [0.81] [0.88] [1.0] [1.1]
17	[0.0, 0.0, 0.0, 0.0] [0.04] [0.11] [0.13] [0.20, 0.20] [0.25, 0.25, 0.25, 0.25, 0.25, 0.25] [0.30, 0.30, 0.30] [0.40] [0.44] [0.45] [0.50] [0.75, 0.75, 0.75, 0.75] [0.77] [0.80] [0.88] [1.0] [1.1]

■

The centroid distance measure can be used to determine which clusters are reasonable. The strongest cluster separation splits the data set into two clusters of sizes 9 and 22 experts. This corresponds to the bimodal structure indicated in example 13.1. The next cluster structure suggested by the analysis forms four clusters breaking off the two largest answers and the seven smallest answers for the original two clusters. Again this separation is visible on the graph in example 13.3.

A second way that the decision on clusters can be made is to use the ancillary information gathered on the experts. For example, if there are two clusters in the data, which ancillary information corresponds to the experts in each cluster? Perhaps, all the experts in the first cluster made very optimistic assumptions when answering the question, and the experts in the second cluster made very pessimistic assumptions. The cluster dividing lines could then be drawn based on the experts' assumption-making in conjunction with the statistical clustering results from the formal cluster analysis. Investigating the relationships between the answer data and the ancillary data is discussed further in the section below. However, it is always important to keep potential conditional relationships in mind and to be watchful for them.

Determining the number of modes and clusters will be useful later in the bootstrap simulation applications for investigating correlation and bias, and for forming aggregation estimates. The cluster formations that look reasonable and any possible ancillary information or explanations relating to the clusters should be documented for these later investigations.

Investigating Between/Within Variation Structure

In most problems where expert judgment data is to be used, several experts are asked more than one technical question. The multiple technical questions may be either

totally different, or some could be quite similar in content and structure. In either case, information regarding any differences in the experts can be gained by examining the variation in the answer data *between* and *within* the experts. A commonly used technique for analyzing between variation versus within variation is analysis of variance (see chapter 11). Example 13.5 illustrates the mechanics of the analysis of variance technique for calculating these two sources of variation. (For further details, introductory statistics or analysis of variance textbooks such as Snedecor and Cochran 1978, chapter 10, are useful.)

EXAMPLE 13.5: ***Between and Within Response Variation Calculation***

Eight experts were asked four technical questions on scales from [0,1]. The between expert variance is MSB, and the within expert variance is MSE. X_{ij} is the response of the ith expert to the jth question. The overall mean of all 32 responses is C, the number of experts is t (= 8); the total number of responses is N (= 32); and n_i is the number of questions asked of each expert (= 4).

	Expert							
Question	1	2	3	4	5	6	7	8
1	0.90	0.50	0.75	0.65	0.80	1.00	0.22	0.44
2	0.95	0.50	0.75	0.65	0.80	1.00	0.22	0.40
3	0.80	0.50	0.75	0.75	0.30	0.38	0.06	0.40
4	0.94	0.48	0.75	0.75	0.65	0.65	0.06	0.38
Total, $X_{i.}$	3.59	1.98	3.00	2.80	2.55	3.03	0.56	1.62
Mean, $\bar{X}_{i.}$	0.90	0.50	0.75	0.70	0.64	0.76	0.14	0.41

$$MSB = \sum_i n_i(\bar{X}_{i.} - C)^2/(t - 1)$$
$$= 1.63/7 = 0.23$$

$$MSE = \sum_i \sum_j (X_{ij} - \bar{X}_{i.})^2/(N - 5)$$
$$= 0.49/24 = 0.02$$

$$MSB/MSE = 0.23/0.02 = 11.50$$

The MSB/MSE is a ratio of variances and it measures the relative difference of between experts versus that of within experts. This ratio is also an F statistic and follows an F probability distribution. If the ratio is large, then the differences of the between-experts values are large relative to those of the within experts. In this example 11.50 is large because it is in the far upper-right-hand tail of the F distribution with (t-1) and (n-t)

degrees of freedom. Therefore, the variation of between experts is significantly larger than the variation within experts.

■

If two or more questions are similar, then the variation of the responses to these questions can be used as a source of random or background variation for the experts. This random source provides a gauge with which to measure the variability between the experts. The section on *Using Analysis of Variance* in chapter 14 indicates how this variation comparison can be used to investigate interexpert correlation. Basically, if the variance between experts for the similar questions is much larger (e.g., four times larger) than the variance within experts, then the experts are giving quite different responses, and correlation among them is not suspected. If not, then the variation from one expert to another is similar to the individual experts' variation, and correlation among experts might be a problem.

If all the technical questions are vastly different in either content or structure, then it is expected that the within-expert variability would be larger than it would be if the questions were similar. The within expert variability could even be larger than the between expert variability (Meyer and Booker 1987a). In this case, it is necessary to investigate why such large within-expert variation is present. Perhaps, some questions were quite familiar to the experts whereas others were never seen before. Perhaps, the experts had difficulty in using the response mode. Perhaps, the experts used an anchoring/adjustment heuristic on some questions and simply guessed on others. Perhaps fatigue was a problem. Perhaps there was some inconsistency in the elicitation process. With proper recording of the experts' rationale and monitoring of the elicitation, these possibilities can be traced and understood.

Even though the multimodality and between/within investigations were done only on the answer data, explanations and reasons for the results found incorporated the ancillary information. At this point some investigation into the ancillary data is needed.

Analyzing the Ancillary Data

The ancillary variables and information in the data base can be analyzed separately from the answer data. However, the conditional relationships between the answers and this data should not be ignored. The primary purpose for the separate analysis is to investigate any redundancies in the ancillary variables, thereby reducing the number of variables in the analysis with the answers later on. A secondary purpose is to gain insight into the structures and relationships among these variables.

The suggested analysis of the ancillary data is given in a series of steps using standard statistical multivariate techniques. More detailed descriptions of these techniques are given in chapter 11.

Step 1: Factor analysis of the ancillary variables

Factor analysis produces a new set of factors from the original set of variables. The original variables are mapped onto the new factors according to their common information

content (based on variability). The mappings are sometimes difficult to interpret, and thus the results may not be very useful. However, if the ancillary variables map well into a set of new factors, either the new factors can be used as the ancillary variables or the original number of variables can be reduced.

Example 13.6 illustrates two cases of factor analyses. The first case indicates how a successful factor analysis can be used. The second case illustrates a factor analysis that is not useful in trimming down the set of ancillary variables.

EXAMPLE 13.6: ***Use of Factor Analysis for Ancillary Variables***

Case I: Successful factor analysis

There are 12 numeric, ancillary variables gathered from an elicitation of 20 experts. A factor analysis on the 12 variables resulted in the following factor loadings on four new factors:

	Factors			
Ancillary Variable	1	2	3	4
A_1	0.054	0.112	0.802	0.032
A_2	0.556	0.236	0.080	0.128
A_3	0.754	0.001	0.026	0.219
A_4	0.011	0.218	0.099	0.672
A_5	0.832	0.064	0.002	0.102
A_6	0.000	0.000	0.347	0.653
A_7	0.107	0.883	0.000	0.001
A_8	0.256	0.495	0.202	0.017
A_9	0.109	0.001	0.389	0.501
A_{10}	0.000	0.049	0.076	0.875
A_{11}	0.672	0.256	0.003	0.069
A_{12}	0.000	0.837	0.006	0.157

The interpretation of these loadings is that variables A_2, A_3, A_5, and A_{11} comprise (load onto) factor 1; variables A_7, A_8, and A_{12} load onto factor 2; variables A_4, A_6, A_9, and A_{10} load onto factor 4; and only variable A_1 loads onto factor 3. It happens that A_2, A_3, A_5, and A_{11} are the only set of variables containing information on the experts' education. Variables A_7, A_8, and A_{12} refer to the experts' recent work experience. Variables A_2, A_3, A_5, and A_{11} refer to the years of work on various related projects. Variable A_1 indicates how long the experts took to interview. With this clean breakdown of variables, the 12 original ancillary variables can be restructured using the four factors. However, the four new factors have interpretations that are a little more general than the original variables. Thus, the granularity has changed.

Case II: Useless factor analysis

Suppose the factor loadings were as follows:

Ancillary Variable	Factors 1	2	3	4
A_1	0.054	0.112	0.802	0.032
A_2	0.256	0.236	0.080	0.328
A_3	0.754	0.001	0.026	0.219
A_4	0.011	0.218	0.099	0.672
A_5	0.832	0.064	0.002	0.102
A_6	0.000	0.300	0.347	0.353
A_7	0.107	0.883	0.000	0.000
A_8	0.256	0.495	0.202	0.017
A_9	0.109	0.001	0.389	0.501
A_{10}	0.000	0.049	0.875	0.076
A_{11}	0.672	0.256	0.003	0.069
A_{12}	0.400	0.437	0.006	0.157

Here there is no clear indication as to which factors variables A_2, A_6, and A_{12} belong. Also, in factor 3, the variables A_1 and A_{10} have nothing in common, making the interpretation for factor 3 difficult. Furthermore, factor 1 does not include all the educational variables, factor 2 does not include all the experience variables, and factor 4 does not include all the work variables. These results are not very helpful in gaining understanding about the ancillary variables relationships or structure.

■

Step 2: Graphical analysis

Factor analysis can only be used on the numeric ancillary variables. Relationships among qualitative variables or among mixed qualitative/quantitative variables can be examined using plots or graphs as suggested in the section above on conditionality.

Because the goal in these analyses is to search for interesting and redundant relationships among the ancillary variables, any graphs that show all the data points falling on or near a line indicate possible redundant information among the two variables plotted. A list of such variables should be made and checked against the correlation analysis suggested in the next step (3). The pairs of variables on this list should be either positively correlated (for a line indicating positive slope) or negatively correlated (for a line indicating a negative slope).

Step 3: Correlation analysis of the ancillary variables

Another useful step in understanding relationships among the ancillary variables is the Pearson pairwise correlation coefficient for all possible pairs of ancillary variables. A level of significance is needed for deciding whether any correlation is significant (important enough). The level is based on the number of pairs, n, for which correlations are calculated. If n is less than 20, then the standard 0.05 or 0.01 levels can be used. If n is greater than 20, then the level should be less than $1/n$.

Even though correlation analysis is considered a bivariate analysis technique, it is useful in gaining understanding of all the ancillary data relationships. It is also useful in verifying graphic results for numeric variables. If any correlation is highly significant (a

significant level of less than 0.0001 or a correlation greater than 0.90), then redundancy is suspected. Only one of the pair of variates needs to be considered for further analysis.

Step 4: Categorical analysis

The ancillary variables can be modeled among each other. One way to do this is by modeling all variables that have integer (ranks, dummy variables, or quantifiable categories) as dependent variables with the numerical ancillary variables as independent variables. **Categorical analysis** techniques will indicate any potential relationships among the dependent and independent variables. This analysis technique is based on linear modeling (Grizzle, Starmer, and Koch 1969) and is not discussed or used extensively in this handbook. Any significant relationships indicated by categorical analysis should be noted and added to the list for potential redundancies.

Step 5: Cluster analysis for variables

Like factor analysis, cluster analysis can be used to examine how much information is shared among the numerical ancillary variables. If the variables cluster in distinct and tight groups, then information is shared among the variables in the group. The variables from any tightly formed groups should be added to the list of potentially redundant variables.

By following all or some of the above steps, lists of potentially redundant ancillary variables are available for interpretation. The following flow chart indicates how the steps can be used, and the example in example 13.7 indicates how results can be interpreted and used. Only results indicating the strongest information redundancies should be used for trimming the set of ancillary variables. Additional similar tests will be done on the combined ancillary/answer variables data set in the next section.

Summary of Steps for Ancillary Data Analysis

<u>For Numeric Data:</u>

(1) Factor analysis
indicates shared information

(3) Correlation analysis
indicates possible relationships

(5) Cluster analysis
indicates shared information

<u>For Descriptive Data:</u>

(2) Graphs
indicates possible relationships

<u>For Integer Data:</u>

(4) Categorical analysis
indicates possible relationships

EXAMPLE 13.7. ***Ancillary Variables Analysis***

The 12 ancillary variables from example 13.6, case II, plus 5 more ($A_{13\text{-}18}$) descriptive variables were analyzed using the five steps. The results are indicated below:

Step 1: Factor analysis, part II, example 13.6: Indicated no clear redundancies of information; no new factors could be used in lieu of other variables.

Step 2: Graphs of all pairwise combinations of the 18 variables: Indicated a strong relationship between A_1 and descriptor A_{13} and between A_3 and A_5.

Step 3: Correlations of all 12 pairwise numerical variables: Indicated one pair (A_3, A_5) strongly correlated with a significance level of 0.0001 and that several other pairs were barely significant.

Step 4: Categorical models of all integer variables (A_7, A_8, A_9) with the other numerical variables: Indicated A_7 is influenced by A_2, A_8, and A_{12} but not strongly.

Step 5: Cluster analysis of the numerical variables: Indicated a weak clustering of A_7, A_8, A_2 and A_{12} and a weak clustering of A_3, A_5, and A_9.

Interpretation

Weak clusterings and no clear factor analysis results indicate little shared information.

The only strong result is the A_3, A_4 correlation. Either of those two could be eliminated from the ancillary variables set.

All other variables should be kept for further analysis.

■

Analyzing the Ancillary Data with the Answer Data

Many of the steps for analysis presented in this section are identical to those described in the previous section, *Analyzing the Ancillary Data.* The main objective in this section is to compile lists of possible multivariate relationships--specifically, which answer variables are related to which ancillary variables. Ancillary variables that are related to the answers are called conditional variables. The model formations and analyses use multivariate techniques. Results from these analyses should be consistent with results found in the previous sections of this chapter.

Step 1: General linear models (GLMs)

As the descriptor *general* implies, it is tempting to try to formulate one giant model of all the numerical variables with the ancillary variables as the independent variables in the model and the answer variables as the dependent variables in the model. Such a **general linear model** would describe relationships existing between the answer and ancillary variables and determine which of the answer variables were conditioned on which ancillary variables. However, this temptation should be avoided for a couple of reasons. First, there is a strong possibility that the same information is shared by many variables. Including all of them in a single model results in erroneous variable relationships. Second, the variables may be of different types (dummy variables, etc.) and different granularities. The model would then be a mixture of levels of detail in the information. Finally, some variables may have missing values. Many computer packages cannot perform the analysis or will give erroneous results.

Some logical general linear models can be formed by using some of the information gained in the bivariate analyses and the ancillary analysis. For example, if any of the graphs of the answer data versus the ancillary data indicated a trend either in the positive or negative direction, these variables should be analyzed using the GLM procedure known as **regression**. The answer variables are the *Y*'s, or dependent variables, and the ancillary variables are the *X*'s or independent variables. Likewise, models can be formed using any of the variables, indicating significant correlations; and models can be formed by using any other information, such as suspected relationships between the variables.

To assist in regression model formulations, many software packages have procedures, called stepwise procedures, that indicate which models produces significant relationships between the independent and dependent variables. However, stepwise procedures do not make model choices. They only give a set of models. Model choices should still be based on the information already gathered from previous analyses and logical variable selections.

Once the models have been run on the regression analysis code or a linear model package, a list of significant models and relationships among the variables should be compiled. These significant results should be consistent with the information already gathered in the previous analyses; however, some differences will result because (1) bivariate results do not necessarily hold for multivariate models, and (2) weak significant relationships can change to no significance or to stronger significance depending upon the procedure used.

Model formations are not restricted to using only ancillary variables for the *X*'s. One answer variable may be modeled in terms of the other answer variables if it appears that the answer variables are related or correlated to each other.

The model formations at this stage of the analysis are solely for the purpose of investigating variate relationships. Model formation for the purpose of obtaining final results and interpretations is described in more detail in chapter 15.

Step 2: Discriminant analysis

This procedure can be used in addition to or in place of the GLM for investigating variable relationships. The objective of this modeling is to find a set of ancillary variables that best discriminates among the values of an answer variable.

Studies have indicated that answer variable values tend to be distributed with multiple modes, multimodal, (Booker and Meyer 1988a, Meyer and Booker 1987a, Baecher 1979). In order to determine which, if any, ancillary variables are responsible for this clumping of values, discriminant analysis can be used. The results of a discriminant analysis indicate a list of variables that best discriminate among the values of the chosen answer (dependent) variable. In order to set up the discriminant analysis, the values of the dependent variable need to be grouped or classified into categories. If there are several modes, this grouping is obvious. If no value groupings are evident, then discriminant analysis is not indicated. If a discriminant analysis is done for each answer variable, then any significant (effective) discriminating ancillary variable indicates a potential variable relationships. New ancillary/answer relationships should be added to the list, and any old relationships should be noted as confirmation of a previous result. The variable relationships found from the discriminant analyses will also be useful in chapter 14 for correlation and bias detection.

Step 3: Mulitvariate correlation

Pairwise correlations have been suggested for the ancillary data, the answer data, and the combined set of both. The information from all these pairwise correlations can be combined graphically to form an *ad hoc* multivariate correlation structure. Such a graph depicts many intertwined relationships among the variables.

One way of constructing this graph is by forming a *distance* between the pairs of variables based on their correlations and plotting the variables according to how far apart they are. Distance measures based on variance and covariance (related to correlation) are used in cluster and discriminant procedures. A simple distance based on the correlation is calculated by 1-r, where r is the correlation coefficient. The effect of this calculation is that the higher the value of r, the closer the distance will be for any pair of variables.

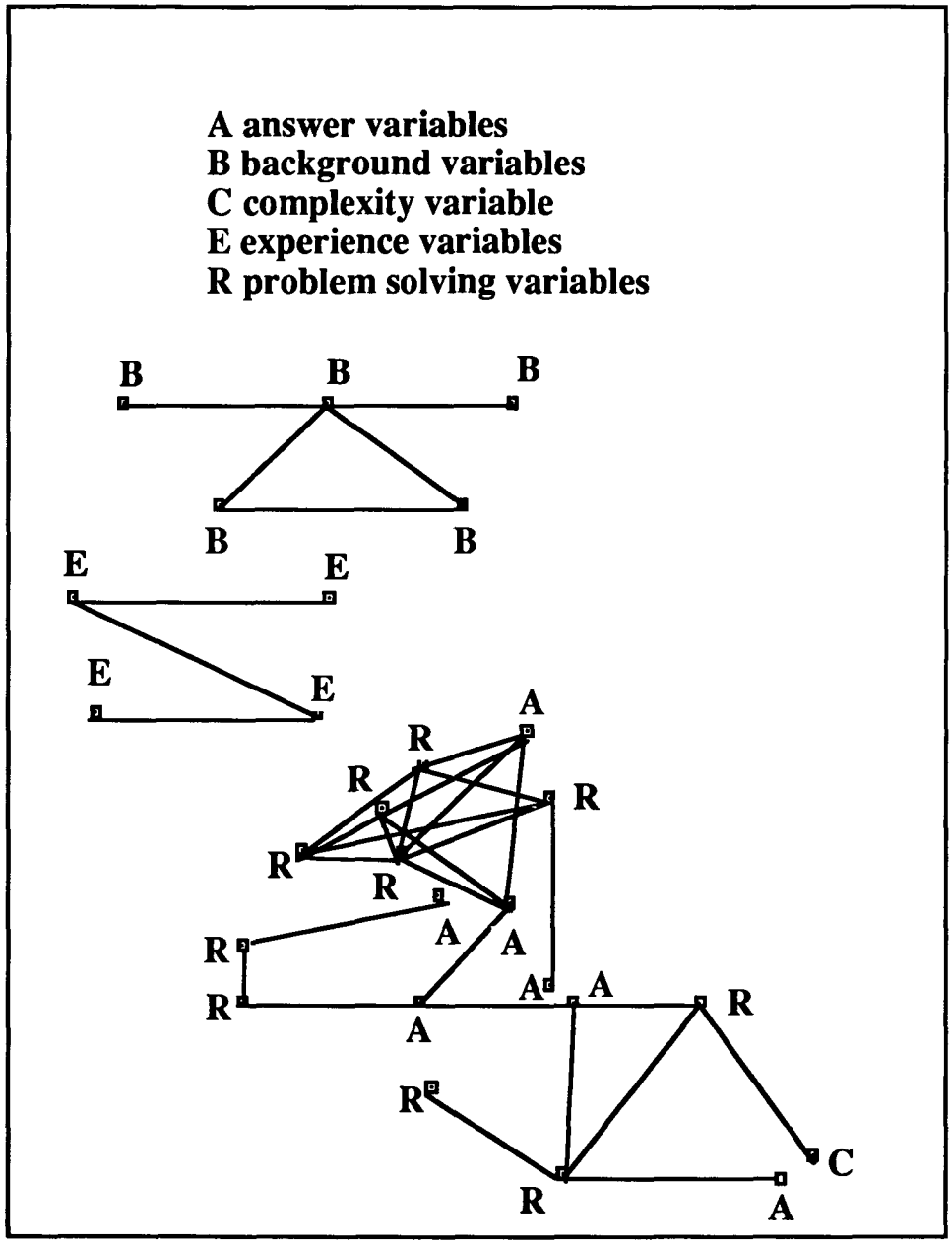
For the graph, distances should only be calculated for pairs of variables that have significant correlations. The rules for significance are outlined above in the bivariate correlation step.

Example 13.8 illustrates a graph from the Meyer and Booker study (1987a). Three separate conglomerates of variables are evident from the significant correlations among the variables. The first group of variables comes from five different ancillary variables describing the experts' backgrounds. The second group of variables comes from four different ancillary variables describing the experts' experiences. The third group of variables comes from all the answer variables (labeled *a*), 10 of the variables describing the experts' problem-solving processes (labeled *r*), and the single variable which described the experts' evaluation of the complexity of the technical question (labeled *c*).

Many interesting results can be seen from such a diagram:

1. There is no connection between the background variables, the experience variables, and the answer/problem-solving variables. The separation of the answer variables from the experience and background variables implies little evidence for conditionality.
2. There is a close connection between the answer variables, the problem-solving variables, and the complexity variable. This connectivity implies strong evidence for the answers being conditioned on the problem-solving variables and the complexity variables.
3. The answer variables are closely interconnected among themselves. The problem-solving variables are also closely interconnected.

Example 13.8: ***Multivariate Correlation Analysis***



■

The sole purpose of performing GLMs, discriminant analysis, and mulivariate correlation steps for ancillary and answer data is to search for strong, consistent relationships among the variables in general and, in particular, to search for answers conditioned on ancillary variables. From the results of these analyses, an expert judgment model can be constructed, as discussed in chapter 15.

14

Correlation and Bias Detection

In this chapter, the concept of correlation is defined and discussed as it is used in expert judgment applications.

Correlation among experts is closely related to the concept of dependence, and distinctions and similarities of both concepts are discussed below in *Defining Correlation and Dependence.*

Correlation is also closely related to the various forms of bias, discussed in detail in Part II; the affinity of these two concepts is addressed in *Bias and Correlation Relationships.*

Because correlation among experts is often considered a problem area in analysis of expert judgment data, the third section, *Detecting Correlation in the Analysis,* focuses on various methods of detection. This section is organized as a series of 14 steps. Each step relates to the usage of several different analysis techniques. These steps and techniques may appear to be redundant. Indeed, they are meant to be redundant. Comparing results from different techniques is the only way to verify conclusions about correlation. In *Analysis Summary and Conclusions,* we summarize both the 14 steps and the results from the examples in the steps.

Defining Correlation and Dependence

Correlation among expert judgment answers has typically meant dependence or lack of independence among expert answers. Thus, to discuss correlation, the concepts of independence and dependence need to be clarified.

One concept of independence comes from a mathematical definition and is referred to as probabilistic independence. The mathematical procedures for combining data from multiple experts require that the data have this type of independence. In probabilistic terms, two events, A and B, are said to be independent if the probability of A is unaffected by what happens to B. Stated another way, the unconditional probability of A, $P(A)$, is unaffected by B such that $P(A) = P(A|B)$, the probability of A given B (Feller 1957).

In the context of examining independence in data, the same definition can be used, but the process of identifying independence is not so straightforward. Analysts tend to

think of two pieces of data as being independent if the occurrence of one datum is unaffected by the other. One way of defining *unaffected* is to examine ways in which the data were collected. This examination involves investigating conditionality. If data are collected under various conditions, they may be unconditionally dependent because those conditions are affecting the data and the data are affecting each other. Thus, the observation of one datum is affected by another such that $P(A) \neq P(A|B)$. However, if the data are collected under conditions, C, mutually affecting both A and B, they could be conditionally independent such as $P(A|C)$ is independent of $P(B|C)$. In either case, conditionality becomes the focal point of investigating independence or dependence.

The terms correlation and dependence have been used interchangeably and synonymously in expert judgment problem settings. This usage is also mathematically valid and is in keeping with the probabilistic definition. However, the terms zero correlation (uncorrelated) and independence can only be used interchangeably when the data are normally distributed. Only in the normally distributed cases does zero correlation or uncorrelated guarantee independence. In nonnormal cases, zero correlation could imply either independence or dependence. Thus, for most of the analyses in this book, the dependence/independence problem is discussed from the dependence (or correlated) viewpoint.

In collecting expert judgment data, analysts have historically speculated that the data were not independent (Baecher 1979, Winkler 1981). The reasoning was that dependence is likely because the experts had many *conditions* in common that would affect their estimates. Analysts considered such conditions as shared training, common work experiences, and exposure to the same data bases. Through time the speculation about the effects of these conditions became identified as sources of correlation among the experts.

A simple example can be used to illustrate how this line of reasoning developed. The following is a sample of probability estimates from five different experts for an event: (0.1 0.15 0.1 0.6 0.65). The bicluster structure of the estimates is commonly seen (Baecher 1979, Booker and Meyer 1985, Meyer and Booker 1987b). The analyst looking at the clustering of the answers tends to arrive at the conclusions that the first three experts are giving the same answer and the last two are giving the same answer, that there are really only two *independent* answers, and that this is not a sample of five independent pieces of information. The data *appear* to have a correlation structure with the first three experts being correlated to each other and the last two experts being correlated to each other. If the analyst has assumed that experts should be correlated, then the clustering of the data supports that assumption. However, it should be noted that the clustering is the only reason for suspecting dependence. No conditions have been examined to support the dependence idea, nor has any reason been given for dependence.

The reasoning used in the above example was responsible for the development of a new body of literature on how to deal with data from dependent sources (Winkler 1981, Lindley and Singpurwalla 1984). The focus of this research was to establish methods for handling dependent data, assuming that the dependencies existed and in many cases assuming that the correlation structures were known. The analytical focus concentrated on these assumptions and not on the real issues. First, the correlation structures are generally not known. Second, correlation among the experts may not exist or may not be an analytical problem if it does exist. Third, conditionality and granularity need consideration in determining and interpreting correlation.

Two studies (Booker and Meyer 1985, Meyer and Booker 1987b) had the goal of identifying possible sources of correlation among experts. In both studies, it was not assumed that experts were correlated nor that clusters of answers implied correlation. Instead the approach was to investigate any possible sources of correlation using the definition of dependence based on conditions and monitoring the effect of granularity. Sources (conditions) were sought from many different forms of information gathered in intensive interviews of the experts.

For the model formations of each study, the granularities were chosen with the level in the first study more detailed than in the second. For the chosen granularity, the results of each study indicated that the answers were conditioned on the problem-solving processes of the experts. These conditions appeared to be a reasonable source for possible correlation or dependence among the answers. However, an equally valid result would be that the answers were conditionally independent, being conditioned on the sources in the same way. Therefore, discovering conditions affecting answers does not automatically imply dependence.

Granularity must also be considered in interpreting the results. In both studies, at a granularity finer than the ones chosen for analysis, the experts had nothing in common in background, experiences, or problem solving (i.e., there were no conditions that could induce dependence among them). Applying the conditioning argument at a different granularity results in the conclusion being that there is no dependence among the experts. Of course, another equally valid possibility is that sources of correlation could exist at an even finer level of detail than was gathered. Gathering information at such an extremely fine granularity might be difficult or impossible because the subjects might not be capable of providing information at such a level or by providing it the interview would be prohibitively long.

Therefore, in defining correlation or dependence among experts, the conditioning argument from the definition of probabilistic independence can be used as a guide. In its application, care must be taken to use the predecided granularity for the entire problem because conclusions about correlation and dependence can change if levels are changed.

Bias and Correlation Relationships

In the broadest definition, **bias** can be related to a number of sources. Bias can be induced from the interview environment through factors such as the interviewing technique, the question phrasing, and the interactions with others including the interviewer present at the interview (**motivational bias**). Bias can be related to the internal consistency of the experts' reactions, conditioning, and thinking (**cognitive bias**). This bias occurs when the expert is not consistent in his own reasoning or when the expert is not consistent with fundamental rules of logic. Bias can be induced by faulty memory retrieval (**availability bias**). Therefore bias can be considered a conditional phenomenon. It can be initiated by certain external or internal conditions during the elicitation.

Because bias can be found in any of the above forms, it becomes difficult to monitor and to control. However, monitors and controls such as those discussed in Part II should be used and considered an integral part of the experimental design. The various

ways of handling and minimizing bias are discussed in chapters 3, 7, 8, and 10 in conjunction with the elicitation techniques.

In the analysis of the data from experts, bias manifests itself as conditions responsible for correlated data. Therefore, application of bias-minimizing methods is important for minimizing possible dependence. Of course, answer data can be correlated for reasons other than bias; however, the original postulated sources of correlation can all be traced to the different types of bias.

The correlation in the data appears in the forms of interexpert correlation or between expert correlation. If an expert has a motivational bias that drives him to be consistently optimistic, then his answers will reflect that bias by being more optimistic than the other experts. This expert's answers could indicate interexpert correlation but not between-expert correlation. If all the experts shared the common goal for a project, then their answers might exhibit between-expert correlation and little or no interexpert correlation. The remainder of this chapter describes how to detect and analyze correlated data in both forms regardless of whether a bias or some other condition is the source of correlation.

Detecting Correlation in the Analysis

Several steps using several methods are given below to investigate any potential correlation among the experts' answers. The main emphasis in each step/method will be on detecting correlation or dependence in the answers given by the experts using the answer data itself and the conditionality definition of dependence. In other words, dependence is sought in terms of possible conditions that may be influencing the answers of multiple experts or that may be biasing the answers given by a single expert. It is not assumed, *a priori,* that dependencies exist. The investigation is done using only the data (except for step 14) and minimizing the assumptions necessary for using the analysis techniques.

The methods for detection include using the results from correlation analysis, multivariate analysis, analysis of variance, and simulation techniques, such as the bootstrap. Detection is also done by using the features of the chosen elicitation method and using assumed correlation structures and distribution forms. Any one of these steps or methods can be used individually. However, to maximize the information inherent in the data, it is suggested that all steps, except the last one (step 14), be done for a data analysis approach. Because the last step is an assumptional approach, it may not always be appropriate to use, but it can be done in conjunction with the others or by itself.

The following 14 steps fall into various categories according to usage. These categories are listed as subsection headings. There may be one or more steps under each heading.

Using Granularity

Step 1: Defining a level of detail to be used in the analysis and in the interpretation of the results.

Even after having assumed and established a definite granularity for the problem, the application of the definition of dependence must be done considering granularity. What

does it really mean to say the experts are correlated? With a fixed granularity, the question becomes, What does it really mean to say the experts are correlated for this granularity of data and analysis? It has already been shown that the answer to this question can be different if the granularities are allowed to change. Therefore, the first step and most important step is to fix the level in the analysis and in the interpretation of the results. The results of the rest of the steps in the correlation investigation are given in terms of the granularity chosen in this first step.

Using Hypothesized Sources of Correlation

The next few steps are data preparation steps for some of the remaining methods of correlation detection. They do offer the analyst an opportunity to hypothesize what *conditions* in the ancillary data might be potential sources of correlation. They also offer an opportunity to do some *hands-on* data analysis gaining additional insight into the data set for formal model formation later (chapter 15).

Step 2: Compiling a list of conditions from the ancillary information that are suspected or hypothesized as being potential sources of correlation.

This list could include suspected sources from the literature (Booker and Meyer 1985), such as, the experts' educational background, commonly shared work experiences, how recently they worked on a similar problem, assumptions made in the problem-solving process, and heuristics or rules used in solving the problem. Likely candidates for this list can come from the results of the multivariate analyses done on the answer data and from the ancillary data in chapter 13. Any ancillary variables that were significantly correlated to answer data should be added to this list.

Step 3: Examining the raw data clusterings.

In chapter 13 on understanding the data-base structure, the multimodal structure of the data was investigated. A raw data frequency plot, such as in figure 13.1, was helpful in determining the raw data clusters. The important point here is that these clusters of the raw data do not necessarily imply a correlation or dependence structure. For one reason, the data may be conditionally independent on one or more ancillary variables. For another reason, there may not be any rationale for the clusterings. However, it is important to try and find reasons for the raw clusterings. Also do any of the ancillary variables group the data according to these clusters? This question is partially answered in the ancillary data analysis in chapter 13, but a more complete answer to this question is needed in the correlation investigation. The steps below pursue this answer.

Step 4: Forming clusters of the data using the suspected conditions.

New data clusterings can be formed using the *conditional* variables listed in step 2. This formation can be done by hand by simply splitting the values of the conditional variables into categories or clusters and listing the corresponding answer data for each category. If the answers within these clusters are numerically similar, then the question posed in step 3 is answered affirmatively. An alternative formation can be done by plotting

the categories of the conditional variable against the numerical answers to see if clusters form. In either case, if no cluster formations are evident, information is still gained regarding the relationships of the data and the ancillary information. This information should be consistent with the results from the analysis in chapter 13.

Step 5: Compiling a list of unsuspected variables.

This list should be a random selection of the ancillary variables not hypothesized as sources of correlation. Ideally, this list should be as long as the list in step 2; however, that may not be possible in some cases either because there may not be enough ancillary variables left or because such a list would be too long for the analyses in the following steps.

The purpose of this unsuspected list is twofold. First, a random selection provides a chance of discovering conditions that might be sources but were not previously hypothesized as possible sources. Second, this list provides a comparison to the list of suspected sources from step 2.

The lists compiled in steps 2 and 5 should not be considered final at this point in the investigation. The results from the remaining steps/methods will help determine if these hypothesized lists are correct.

Step 6: Forming clusters of the data using the unsuspected variables.

Graphs or hand listings of the unsuspected ancillary variables and the answers should be examined to see if the answer data clusters similar to the raw clustering (from step 3). The clustering for an unsuspected variable could be better (closer to the raw clustering) than the clustering for the suspected variables (from step 4).

Steps 2 through 6 will give three different clustering mechanisms to interpret and compare: (1) the raw data clusters, (2) the data clustered by the various conditions suspected of inducing dependence, and (3) the data clustered by other unsuspected conditions. If any of the variables forms clusters of approximately the same size and of similar values to the raw data clusters, then there is reason to suspect that the variables are sources of correlation. The remaining steps will help verify this conclusion. If none of the variables produces good clusterings, there may still be correlation among the experts and the other steps are necessary. Example 14.1 illustrates hand listings of cluster formations from ancillary variables.

Example 14.1: ***Clusterings Using Different Variables***

From the list of background variables on the eight experts, years of work experience in the technical area (*YRWORK*) and one of the variables describing the expert's problem-solving process (*PSRATE*) were selected as potential sources of correlation. Two other variables, not suspected, were also chosen: discipline of highest degree (*DEGREE*) and whether or not the expert had any experience in a nuclear experimental facility (*EXPFAC*). The answers to the technical question are listed as *P*.

Expert	*P*	*DEGREE*	*EXPFAC*	*YRWORK*	*PSRATE*
1	0.90	Engineer	Yes	1.5	1
2	0.50	Engineer	Yes	0.5	-1
3	0.75	Engineer	Yes	3.0	0
4	0.65	Physics	No	3.0	-3
5	0.80	Engineer	yes	1.2	-1
6	1.00	Engineer	No	10.5	-1
7	0.22	Physics	No	7.6	-2
8	0.44	Engineer	Yes	4.9	2

Two clusters for each variable can be formed by hand listing as follows:

Variable	Clustering Description	Cluster 1	Cluster 2
P	[0,0.50] & [0.50,1.0]	0.22, 0.44, 0.50	0.65, 0.75, 0.80, 0.90, 1.0
DEGREE	Physics & engineer	0.65, 0.22	0.90, 0.50, 0.80, 0.75, 0.44,1.00
EXPFAC	Yes & no	0.65, 1.00, 0.22	0.90, 0.50, 0.75, 0.80, 0.44
YRWORK	> 5.0 & [0,5.0]	1.00, 0.22, 0.44	0.90, 5.00, 0.65, 0.75, 0.80
PSRATE	[0,2] & [-3,0)	0.90,0.44, 0.75	0.50, 0.65, 0.80, 1.00, 0.22

Here none of the ancillary variables clusters the answer data similar to the raw data clusters (line *P*), even though the cluster sizes (2 or 3 for cluster 1 and 5 or 6 for cluster 2) are similar for the ancillary variables.

■

Using Correlation Analysis

Step 7: Calculating the Pearson correlation matrices of the answers.

Calculating the Pearson pairwise correlation matrix (chapter 11) for the experts' answers to all questions is helpful for finding experts that are numerically correlated in their answers. This correlation is strictly a numerical correlation, and it does not necessarily imply dependence of the experts under the definition. However, it is good to know which pairs of experts' answers are significantly correlated using the Pearson correlation coefficient. Any significant correlation coefficients can indicate potential expert correlation that can be further tested under the dependence definition using the remaining steps/methods.

The correlation coefficients are calculated for pairs of experts by comparing the numerical similarity between the experts answers to more than two technical questions. There is an implicit assumption made regarding the similarity of the technical questions asked. The coefficient calculation assumes that the answers to these different questions represent repeated measures of each experts' knowledge and problem-solving processes. This assumption is not a bad one to make if the technical questions are similar in structure (form and response mode) and if they are similar in content. One way to help verify similarity among the questions is to check the results of the analysis done in chapter 13. If

the answer variables were found to be related to each other from Pearson correlation coefficients, from graphs, from GLMs, or from cluster analysis, then similarity among the questions is plausible.

Example 14.2 illustrates two different correlation matrices for the data from example 13.1. The first matrix is the expert correlation matrix of coefficients calculated across answers to the questions. The second is the answer matrix of coefficients calculated across experts to help verify question similarity through correlation of the answers.

EXAMPLE 14.2: ***Correlation Matrices of Experts and Answers***

Eight experts were asked four questions that were similar in subject matter and used the same response mode, a continuous scale from 0.0 to 1.1. The correlation matrix for the experts is formed using the Pearson correlation coefficients for each pair of experts as follows:

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8
E_1	1.00	-0.41	0.00	-0.46	0.87	0.74	0.46	-0.15
E_2		1.00	0.00	-0.58	-0.03	0.24	0.58	0.66
E_3			1.00	0.00	0.00	0.00	0.00	0.00
E_4				1.00	-0.80	-0.93	<u>-1.00</u>	-0.69
E_5					1.00	<u>0.96</u>	0.80	0.35
E_6						1.00	0.93	0.52
E_7							1.00	0.69
E_8								1.00

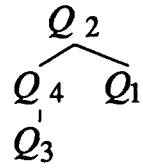
The only significant correlations in this matrix (using a 5% significance level) are the ones underlined. The lower triangular values are not listed because they are identical to the upper triangle values folded over at the diagonal of 1.00 values. One interesting result here is the perfect negative correlation of expert number 4 and expert number 7. Usually, analysts tend to be concerned with positive correlations among experts because that has the effect of underestimating the true expert variation if ignored. However, if experts are negatively correlated, the expert variation is overestimated if the correlation is ignored. Neither effect is desirable and both produce statistically biased estimates of the expert variation. In any case, correlations (positive or negative) should be investigated.

The correlation matrix for the answers follows:

	Q_1	Q_2	Q_3	Q_4
Q_1	1.00	<u>0.99</u>	0.51	<u>0.86</u>
Q_2		1.00	0.52	<u>0.87</u>
Q_3			1.00	<u>0.85</u>
Q_4				1.00

The significant correlation coefficients (using a 5% significance level) are underlined. Here the answers to the questions are very well connected into a structure.

This structure indicates that using the answers could serve as repeated measures of the experts' processes of answering similar technical questions. Therefore, the correlation results from the first matrix (experts) are also usable. The answer structure can be diagramed as follows, with the connecting lines representing significant correlations:



■

Using Multivariate Analysis

Step 8: Determining which clusterings match the raw clusters.

The objective in this step is to determine which of the ancillary variables cluster the data in a manner most similar to the raw data clusters by using multivariate analysis methods such as discriminant analysis and by using some of the results obtained from the analyses done to gain understanding of the data base in chapter 13. If an ancillary variable clusters the answers well, and this clustering has a basis in terms of cognitive theory, then a potential source of correlation is identified (Meyer and Booker 1987b). The word *potential* is important here because it is not true that the raw clusters of the answer data indicate the existence of a correlation structure. Additional analysis such as using simulation techniques and examining variances (in steps 9 and 10) is needed to make such a determination.

Discriminant analysis could be used to complete the investigations done in steps 3 and 6 where the answer data are clustered according to the ancillary variables from the lists compiled in steps 2 and 5. Discriminant analysis can be used in two different ways. First, it can be used to determine which of the ancillary variables best predicts to which raw cluster each datum belongs. Second, it can be used to determine which ancillary variables predicts the clustering behavior of any other ancillary variable, such as the variables from the lists in steps 2 and 5.

To implement the first application, a variable is chosen from either the list in step 2 or in step 5. Categories of values for the clustering are formed using the values of that variable. For example, the variable *PRSOLV* in example 14.3, could be categorized as positive (1) or negative (-1) values. *PRSOLV* is the variable describing the experts' problem-solving score accumulated over several problem-solving characteristics. This variable was from the list of suspected correlation sources (step 2). Another variable, *DEGREE* categorizes the discipline of the highest degree earned by the experts. It is not suspected as being a source of correlation (from the list in step 5).

EXAMPLE 14.3: ***Using Ancillary Variables as Discriminators***

The answers of 8 experts are listed below as Q_1. Two other ancillary variables, *PRSOLV* and *DEGREE*, represent choices from the lists in steps 2 and 5, respectively. Discriminant analysis is done with each of the ancillary variables to determine if they might

be potential discriminators for the answers. The results given below indicate how many answers would be misclassified if the ancillary variable was used as the discriminator.

Expert ID	*PRSOLV*	*DEGREE*	Q_1
1	1	1	0.90
2	-1	1	0.50
3	1	1	0.75
4	-1	0	0.65
5	-1	1	0.80
6	-1	1	1.00
7	-1	0	0.22
8	1	1	0.44

Discriminant analysis results using where the asterisks indicate misclassification by *PRSOLV*:

Expert ID	From *PRSOLV* Class	Classified Into *PRSOLV* Class	
1	1	1	
2	-1	-1	
3	1	1	
4	-1	-1	
5	-1	1	***
6	-1	1	***
7	-1	-1	
8	1	-1	***

Discriminant analysis results where the asterisks indicate misclassification by *DEGREE*:

Expert Id	From *DEGREE* Class	Classified Into *DEGREE* Class	
1	1	1	
2	1	0	***
3	1	1	
4	0	1	***
5	1	1	
6	1	1	
7	0	0	
8	1	0	***

With so few experts (8), *any* misclassifications indicate poor discrimination by the ancillary variable. Thus neither of these variables is a good clustering mechanism for the answers.

■

The second use of discriminant analysis can be done in two ways. Ancillary variables can be analyzed with each other to see if they are mutual discriminators or their discriminating abilities can be compared by direct examination of the misclassification output. In example 14.3, the two ancillary variables did not have the same misclassification pattern. This difference in patterns indicates that these two variables had little in common. This result should be consistent with the results from the analysis of the ancillary variables done in chapter 13.

Using Analysis of Variance

Step 9: Examining between and within cluster variations.

Comparing between-expert and within-expert variations is the basic philosophy of analysis of variance procedures in statistics (see chapter 11). The formal analysis of between and within answer variation was suggested in chapter 13 as part of the understanding of the data-base structure. Interpreting the results from this analysis in the context of correlation is given in this step.

In the previous steps, analyses have concentrated on data clusterings. How do clusterings relate to correlation? One way of determining if data within a given cluster are correlated is by comparing the variance structures of between-expert variation to within-expert variation. Formal analysis methods such as discriminant analysis are based on complex, but similar, variation comparisons. If interexpert correlation is suspected, then within-expert variation can be used as a measuring standard. Within-expert variation is how closely each expert repeats himself on the answers to various, but similar, questions.

For example, if experts are asked multiple, but similar, questions, then the variation between experts can be compared to the variation within experts. If the two variations are the same, the interpretation would be that there is no difference in answers given by experts. If between-expert variation is much larger than within-expert variation, then this result indicates that the experts are acting more independently of each other. The variation comparison approach only works if multiple questions are asked of the experts, and these questions must be very similar in content, structure, and response mode.

If the multiple questions asked are very different in content or structure, then the within-expert variation comparison measures a combination of question variation and within-expert correlation. Even so, the between versus within variance comparison can still be useful. If each expert answers the various questions differently, indicating no consistent bias, such as always answering with low values, then the within-expert variation would be large. In this case, if the between variance is of the same size or larger than the within variance, correlation among the experts might not be suspected. If the between-expert variance is small relative to the within-expert variance, correlation might well be suspected.

Example 13.3 illustrates how these variations are calculated for the eight experts answering four very similar questions. Here the between-expert variation is 0.23 and the

within-expert variation is 0.02, significantly smaller by comparison using an F test. By examining the raw data, a systematic bias among experts 1, 2, 3, 4, 7, and 8 is suspected. Each of these experts consistently answers the four questions as low or high. Therefore, the within-expert variation is expected to be a small value and is a good measuring standard for the between-expert variation. In this case, the between variation is much larger indicating that the eight experts may not be correlated.

Using Simulation Techniques

A major concern with correlated data is when an aggregation or pooled estimate must be formed. Most rules for aggregation require the independence of the data (not just uncorrelated data). However, the chosen aggregation estimator can be used on the existing data to help determine if correlation is a problem. This determination is made by investigating the behavior of the chosen estimator comparing correlated and uncorrelated data sets. A convenient way of determining this behavior is by using simulation techniques.

Two such simulation techniques, Monte Carlo and bootstrap, are discussed in chapter 11. Using the median as an example of a chosen aggregation estimator, steps 10 through 12 illustrate how to use simulation for investigating correlation. Heavy use is made of the bootstrap technique because it does not require any assumed distributional forms for the data and relies solely on the information content of the raw data.

Step 10: Comparing different stratified bootstrap sample results.

The bootstrap simulation technique can be used to investigate correlation in conjunction with ancillary variables listed in steps 2 and 5. The resulting bootstrapped distributions of a chosen estimator (such as the median) are formed for each clustering using stratified sampling techniques. This sampling is different from the ordinary bootstrap as described in chapter 11 because the clusters of the data act as the strata. Each bootstrap sample is formed by randomly selecting one datum from each strata. The resulting distribution of the median from stratified bootstrapping for a cluster formation from one variable can then be compared to the results from other clustering variables.

Specifically, this comparison can be made by examining the dispersion of the resulting bootstrapped distribution of the medians. The dispersion of highly correlated data (r=0.9) is smaller than for uncorrelated data (r=0.0). Dispersions can be measured using the variances of the bootstrapped distributions, the ranges, or some central probability coverage interval such as the central 90% putative interval (the difference between the 95th and 5th percentiles).

If the central 90% putative intervals of the distribution of the median are used as the measure of dispersion, then these intervals are at least three times wider for uncorrelated data than for highly correlated data (Booker and Meyer 1988b). Thus by comparing the 90% putative intervals for the bootstrapped medians from different cluster formations, relative correlation structure can be determined. If one cluster formation (from an unsuspected clustering variable) results in an interval three or more times wider than from another (suspected or good) clustering variable, then the data in the first case is uncorrelated relative to the second case. Again the correlation investigation is geared to finding conditional variables in accordance with the definition of dependence.

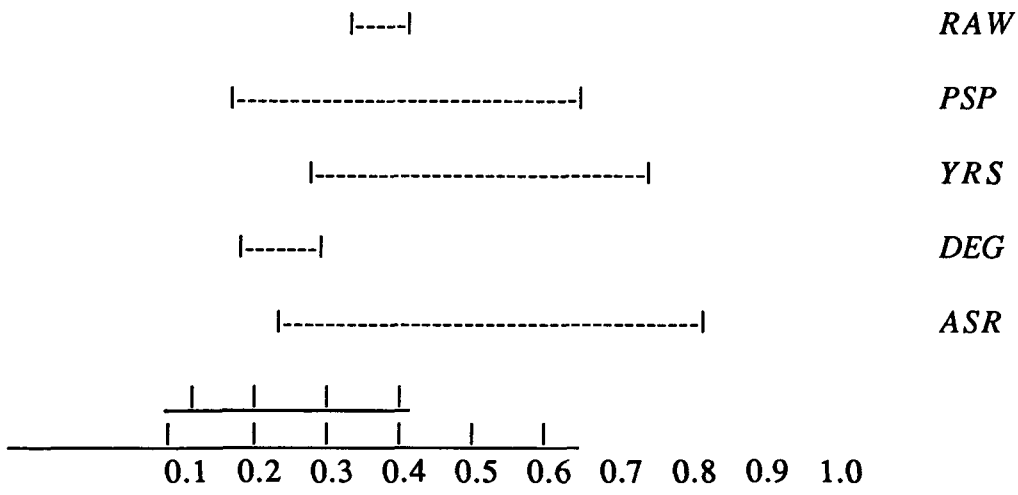
To form these bootstrapped distributions, a modification is needed in the bootstrap technique (chapter 11). The first step is to classify the data into clusters using the clustering variables from the lists in steps 2 and 5. Then form stratified bootstrap samples (e.g. 1000) by randomly sampling 1 datum from each of the m clusters. The median is calculated from the m values. For 1000 or so such samples, a distribution of the median is formed, and the corresponding percentiles (e.g., 5th and 95th) variance or range can be found, indicating dispersion.

Example 14.4 illustrates the results of stratified bootstrapped median distributions for 4 ancillary variables and for the raw data itself. Two of the ancillary variables were from the suspected list (step 2), and two were from the unsuspected list (step 5). The central 90% putative intervals plotted in the example indicate relative correlations due to the various clustering mechanisms. By far the raw cluster formation is the most narrow. *DEG* is the only other clustering variable indicating possible relative correlation. The other variables are not very promising as potential sources of correlation.

EXAMPLE 14.4: ***Ninety Percent Putative Intervals for Bootstrap Medians Using Different Variables as Strata***

Six clusters were formed for each of the five variables; *PSP* and *YRS* were from the suspected variable list, and *DEG* and *ASR* were from the not suspected list. Stratified bootstrap sampling was done forming 1000 samples. The medians of each sample were calculated and sorted. The 5th and 95th percentiles of these medians are marked by vertical lines (|---|). Each sample was of size six, one point was randomly selected from each of the six clusters (strata) in the manner of simple random stratified sampling (Cochran 1963). The variables chosen were as follows:

RAW - the raw data as it clustered numerically
PSP - a variable describing the problem-solving process of the experts
YRS - the number of years since the expert had worked on that type of problem
DEG - the discipline of the expert's highest degree
ASR - the number of years the expert worked as a code assessor



■

Step 11: Comparing bootstrap distributions using pairwise correlation.

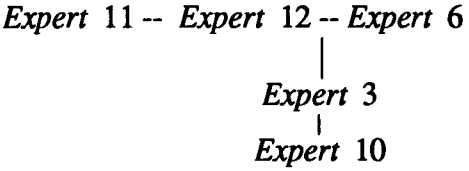
The bootstrap distribution of the chosen estimator (median) for the entire, unclustered, data set can be used in conjunction with the information from the pairwise correlation analysis of the answer data (from chapter 13). First, the entire data set is bootstrapped obtaining the distribution of the chosen estimator. Second, the experts' answers that were correlated in a pairwise manner from the correlation analysis are removed from the sample. Next, the bootstrap distribution of the estimator for this reduced, but perhaps more independent, data set is found. By comparing the dispersions of the two resulting distributions for the median, the effect of including those highly correlated experts can be readily seen.

Example 14.5 shows some results of this analysis for a data set of 31 experts (example 13.3) from the study by Meyer and Booker (1987b). In this example, the removal of the five correlated experts significantly changes the dispersion of the estimator, indicating that the inclusion of those correlated experts does affect the dispersion of the aggregation estimator (the median). The recommendation at this point would be to remove those five experts' answers from the data set and replace them with an average of their values. This average represents a single, but more independent, expert in conjunction with the remaining, uncorrelated experts.

EXAMPLE 14.5: ***Using the Bootstrap with Pairwise Correlation Results***

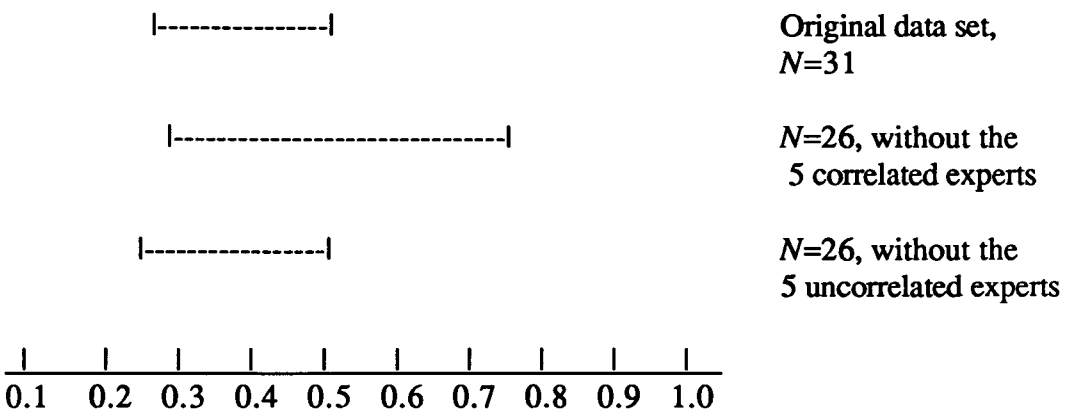
Thirty-one experts were asked a question with a response mode scaled from 0.0 to 1.1 (Meyer and Booker 1987b). The median was chosen as the estimator for aggregating the 31 answers. A bootstrap sampling procedure was done for 1000 randomly formed samples for this data.

In this set of 31 experts, a pairwise correlation of the experts over eight similar questions resulted in the following significant correlation structure:



A bootstrap distribution for the median was then calculated on the data with the above five experts' answers deleted from the data set. However, because dispersion is affected by sample size, the results for a sample of 31 might differ from a sample of 26. This effect would be more pronounced if the samples were 15 versus 10. In order to account for sample size changes in comparing these two bootstrapped distributions, a third bootstrap was done on the original data set deleting five experts' answers from five randomly selected experts that were not at all pairwise correlated.

The resulting 5th and 95th percentiles for the median from that bootstrapped distribution are plotted below.



The results of the three comparisons indicate that the sample size difference between 31 versus 26 has little effect on the median dispersions; however, removing the five correlated answers indicates a significant increase in the dispersion. Therefore, it is recommended that those 5 experts be replaced by a single average value for all 5, making a data set of 27 uncorrelated expert answers.

■

Step 12: Comparing simulation results of the data to known distributions.

Relative comparisons using bootstrap simulation do not give absolute information on the correlation structure of the data. To gain this information, the bootstrap results of the data can be compared to bootstrap results of data sets with known correlation structures. Comparisons can be done by forming random samples of data from distributions with known correlations and comparing these dispersions of the chosen estimator (e.g. median) to dispersions of the estimator for the data.

The major difficulty in this approach is deciding on which known distributions and correlation structures to compare with the raw data bootstrap results. For example, if a normal distribution is decided upon, then what values should be used for the parameters? If the mean and variance are estimated from the mean and variance of the raw data, then the resulting dispersion of the normal sample with zero correlation will be the same as the dispersion of the original data with unknown correlation. Such a comparison does not provide any information about the unknown correlation in the raw data. However, if many different correlation structures are bootstrapped, the effects in dispersion can be seen for changes in correlations. For example, the central 90% putative interval of the median estimator for a normal sample with correlation structure of 0.5 is about twice as much as the central 90% putative interval for a normal sample with correlation structure of 0.9.

Another way of comparing the data set with distributions of known correlation structures using the bootstrap technique is by using a mixture of distributions rather than a single distribution. This method has two advantages. First, the new data set can be modeled into clusters with means and variances reflecting the raw data set clusters. Second, the new data set can be mixed according to the results of the correlation analysis on the raw data set. An example of a mixture is given in example 14.6. Here the correlations and clustering structure of the 31 experts (Meyer and Booker 1987b) were used to form a three-mixture distribution using the normal family. The three mixtures are

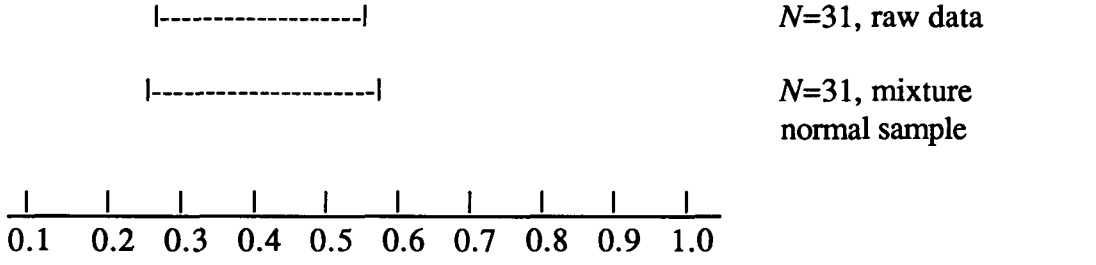
formed from the two distinctive raw data clusters and from the set of the five correlated experts (see example 14.5). The raw data bootstrap compares favorably to the bootstrap results for this three-normal mixture. Of course, other goodness of fit techniques (Conover 1971) could have been used in making this comparison, but the focus of this section is on the use of simulation techniques.

EXAMPLE 14.6: ***Using the Bootstrap for a Normal Mixture***

The following answers from 31 experts (Meyer and Booker 1987b) are cleanly grouped into two distinctive clusters. In addition, a correlation analysis indicates that experts 6, 3, 10, 11 and 12 are significantly correlated with correlations larger than 0.90. The data can be divided into three groups using the correlation and raw clustering information. These three groups are as follows:

Group Size	Mean	Variance	Correlation
17	0.24	0.03	0.00
5	0.18	0.01	0.99
9	0.84	0.02	0.00

Using the normal distribution with the above parameters for each group, a sample of 31 was formed by combining the three groups. This sample was bootstrapped and compared to the bootstrap distribution for the median of the raw data. The resulting 5th and 95th percentile values of the medians are given below. The close correspondence implies that the normal mixture is a reasonable fit for this data set.



■

Using Elicitation Methods

Depending upon the elicitation technique used, biases and correlation resulting from them can be somewhat controlled or monitored. Chapter 3 describes the biases and some methods for countering, reducing, or handling them. Bias consideration is an integral feature of designing the expert judgment elicitation as presented in chapter 8. Bias control and monitoring in conducting the elicitation is discussed in chapter 10.

Step 13: Examining the correlation or bias relative to the elicitation method.

Of the three basic elicitation situations, **interactive group, Delphi,** and **individual interview**, the first two are expected to induce correlation among experts. In these two situations, the analysis steps 1 through 12 above can still be performed, but the results should indicate much more correlation among the experts than the examples given in this chapter that were taken from individual interview situations.

In the group and Delphi situations, particular attention should be paid to the ancillary variables and their clustering or discriminating ability for the experts' answers. Strong clusterings should be expected. That is, experts should reach similar answers for similar reasons, especially for the group situation. In the Delphi situation, the analysis steps 1 through 12 can be performed at the various stages of the process. The results at each Delphi iteration will indicate how the correlation structure among the experts is becoming stronger. The answers of the first stage should be no more correlated than those from an individual interview situation.

In all three situations, the results of the correlation analyses steps can be summarized as indicated in the final section of this chapter. The conclusions reached from these steps will be used in the aggregation chapter (16).

Using Assumptions

Step 14: Making assumptions about the correlation structure based on the analysis or on any additional information.

An example of making such assumptions about the correlation structure was illustrated in the example 14.6. The results of the correlation and bootstrap analyses indicated a potential correlation structure for the data set. These results also indicated that there were 26 uncorrelated experts and 5 correlated experts. One possible conclusion based on this result would have been to form a 27th uncorrelated expert by using the average of the 5 correlated experts' answers.

The idea of combining the five correlated experts comes from the concept of forming an equivalent number of experts (Clemen and Winkler 1985). For k normally distributed experts with a common correlation, ρ, and a common variance, σ, the asymptotic equivalent number of independent experts is

$$n(\sigma^2,\rho) = k\,[1 + (k\text{-}1)\rho]^{-1} \quad .$$

The application of this formula can be impractical because $n(\sigma^2, \rho)$ is much smaller than k for large values of ρ, making the number of independent experts extremely small or equal to one. Another problem in using this concept is the assumption that the experts are normally distributed with a common correlation and a common variance.

Correlation structures can be assumed using a mathematically convenient model such as a multivariate normal model with an assumed value of a mutual correlation coefficient that provides the covariance structure for the data. This model is commonly assumed and can easily be used to handle the dependence problem. Because normality is assumed, the concepts of zero correlation and independence are interchangeable. Example

14.7 gives an example of how such an assumed model and correlation structure can be used (Winkler 1981). The major problem of this approach is the process of making the required assumptions. Since very little experimental evidence is available on distributions and correlations of expert judgment data, many analysts consider these assumptions too unrealistic to make.

Example 14.7: ***Dependent Experts With Assumed Normal Distribution***

Three experts ($k = 3$) are asked to estimate the average cost and standard deviation (in 1000's of dollars) for a newly designed earthquake-proof valve. If each expert's estimate is normally distributed with means and standard deviations as $m_1 = 60$, $s_1 = 6$; $m_2 = 62$, $s_2 = 5$; $m_3 = 70$, $s_3 = 7$, then the distribution for all three experts is a multivariate normal with mean vector, $m = (60, 62, 70)^t$ and covariance matrix, Σ. The elements of Σ are $s_{ij} = \rho_{ij} s_i s_j$, where ρ_{ij} is the correlation coefficient between the ith and jth experts. In this example the values for these correlations were found from previous estimates of these experts to be $\rho_{12} = 0.6$, $\rho_{13} = 0.5$, and $\rho_{23} = 0.6$. The resulting matrix Σ is

$$\Sigma = \begin{bmatrix} 36 & 18 & 21 \\ 18 & 25 & 21 \\ 21 & 21 & 49 \end{bmatrix} .$$

One way of formulating a single normal distribution from this multivariate (three-variate) normal distribution is to use a Bayesian approach. Chapter 16 describes this approach to aggregation in more detail. The way to reduce this multivariate normal to a single normal is by combining the multivariate normal with a prior for the parameter of interest. In this case the parameter of interest is the aggregation estimator for the combined mean responses. A prior that has little influence on the data would be an improper, diffuse prior that does not involve Σ. Combining the above multivariate normal with such a prior gives a posterior distribution (the aggregation distribution) that is normal with a mean, m^*, and a variance, s^{*2}, where

$$m^* = e^t \Sigma^{-1} m / e^t \Sigma^{-1} e \quad ,$$

and

$$s^{*2} = 1 / e^t \Sigma^{-1} e \quad ,$$

and where $e=(1,1,1)^t$ for three experts. This formulation gives weights to the three experts according to the following:

$$w_i = \sum_{j=1}^{3} s_{ij} \Big/ \sum_{j=1}^{3} \sum_{m=1}^{3} s_{mj} \quad .$$

The resulting values for this data are

$$m^* = 62.0$$

and

$$s^{*2} = 22.8 \quad ,$$

with the experts' weights being $w_1 = 0.26$, $w_2 = 0.67$, and $w_3 = 0.07$.

If the experts were considered independent, then the values for ρ_{ij} would be 0.0. The resulting values for the mean and variance would be

$$m^* = 63.2$$

and

$$s^{*2} = 11.3 \quad ,$$

with the experts' weights being $w_1 = 0.32$, $w_2 = 0.45$, and $w_3 = 0.23$.

By ignoring the dependence, a slightly larger mean results and a much smaller variance results.

■

The mathematical formulations in the above example are not trivial. Yet, the multivariate normal model is the most simple and convenient model that allows closed form calculations. If the data does not indicate that a multivariate normal distribution is appropriate, or if the correlation structures cannot be assumed as known, then even these formulas are not useful. In those cases, the results from the other steps that rely on the evidence from the raw data itself must be used to draw conclusions.

Analysis Summary and Conclusions

Investigating the possible existence of correlated data is an important step prior to aggregation analysis (chapter 16). Many aggregation methods assume that the experts' answers are independent. The steps (10-12) involving the bootstrap simulation as a tool for investigating correlation demonstrated the use of aggregation and its relationship with correlated answers. If dependencies can be identified or controlled, then the aggregation schemes can be used with the assurance that the independence assumption is not being violated. Identifying and controlling dependencies was the reason for investigating ancillary variables that might be affecting the answers. If such conditions were found, the experts might be conditionally independent. Under conditional independence, aggregation on a conditional basis can be done without violating the independence assumption.

Conclusions from the analyses done in the above steps (1-14) can be drawn by using the results collectively and relying on results that are consistently indicated by several

steps. After performing the above steps, a summary of the results from each step is useful in determining consistency. In most cases, the same results are indicated by several of the steps making conclusions obvious.

For instance, example 14.8 shows the summary of the results for a set of 31 experts answering a question (Meyer and Booker 1987b). Most of the results for the steps can be found in the examples in this chapter. The conclusion from all the steps is that five of the experts are correlated numerically but the source of this correlation is not known. The numerical correlation can easily be handled by averaging the answers of those five and using that average as a 27th expert. The normality fit in step 13 indicates that the data can be considered a mixed normal distribution. For a single normal distribution, zero correlation and independence are identical. This equivalence supports the conclusion that the data could be considered a set of 27 independent experts for the purposes of aggregation (chapter 16).

EXAMPLE 14.8: ***Summary of the Correlation Detection Steps***

The following summary is for the series of 14 steps outlined in this chapter for the detection of correlation and bias. The data set used in these steps is from example 13.3 and consists of 31 experts' answers to eight technical questions. The ancillary information gathered was reduced to 17 variables relating to the experts' background and problem-solving processes (Meyer and Booker 1987b).

Step	Step Summary
1	Granularity was chosen at the detail level of the analyzed ancillary variables. This level was more general than the original data gathered because a problem-solving score variable was formed for each (8) answer variable from combinations of the original problem-solving characteristics gathered. ***Results or Conclusions:*** *Results will be interpreted and valid at this level of generality.*
2	A list was formed of nine variables suspected as sources of correlation. ***Results or Conclusions:*** *These variables represented the experts' recent background and problem-solving processes.*
3	A cluster analysis was done on the data for each answer. **Results or Conclusions:** *The data formed two or three major clusters.*
4	The data was clustered using the suspected ancillary variables. ***Results or Conclusions:*** *The problem-solving scores were the only variables clustering the answer variables in a similar way to step 3.*
5	A list was formed of 11 variables not suspected as sources. ***Results or Conclusions:*** *These variables represented the experts' earlier history.*

6 The data for each answer variable was clustered using the not-suspected ancillary variables.

Results or Conclusions: *None of these ancillary variables clustered the answer data in a similar way to step 3.*

7 Pairwise correlation matrices were calculated for the 31 experts and the eight questions.

Results or Conclusions: *The eight questions were all highly correlated. Of the 31 experts, 5 were highly, mutually correlated.*

8 Discriminant analysis was tried on three of the suspected ancillary variables and on three of the not-suspected ancillary variables.

Results or Conclusions: *Only the problem-solving score variables discriminated between the answers for some of the eight questions.*

9 Analysis of variance was done on the eight answers for the 31 experts.

Results or Conclusions: *The within-expert variation was much smaller than the between-expert variation, indicating interexpert consistency and possibly bias, but also indicating less correlation among experts.*

10 Stratified bootstrap simulations for the median were done on the six variables from step 8 for each of the eight answer variables. The descriptions and results for the remaining steps are only listed for the one answer variable that indicated potential expert correlation. For this one answer variable, each ancillary variable was used to cluster the answer data into six clusters of nearly equal cluster sizes.

Results or Conclusions: *Some of the results are in figure 14.4. Of the variables analyzed, only one of the variables for one of the answer variables (the one for DEGREE) indicated any potential as a source of correlation. This result may have been because one cluster of the DEGREE variable was of size one, inducing a bias into the stratified sampling.*

11 The raw data for the answer variable in step 10 was bootstrapped with the five correlated experts deleted, five uncorrelated experts deleted, and none of the experts deleted.

Results or Conclusions: *Removing the five correlated experts resulted in a significant increase in dispersion compared to the bootstrap results with no experts removed (example 14.5). Removing five uncorrelated experts produced the same results as the case with no experts removed. Therefore, the five correlated experts are affecting the variance of the median, and their estimates should be combined into a single value to represent a single expert that is uncorrelated to the others.*

12 The raw data bootstrap results on the answer variable in step 10 for the median were compared to a three-normal distribution mixture using different means, variances, and correlations.

Results or Conclusions: *The bootstrap results for this mixture modeled the data well (example 14.6). The correlation structure of the mixture could represent the raw data. If the data can be modeled using normals, then the concepts of zero correlation and independence can be interchanged.*

13 The **ethnographic method** with **verbal protocol** methods were used in an individual interview situation.
Results or Conclusions: *Biases and correlations during the interview process were controlled and minimized. No other biases were suspected to affect the results of the data analysis.*

14 No additional assumptions about the correlation structure were made.
Results or Conclusions: *An analysis of a mixture of three normals could have been developed similar to the one in example 14.7. However, such development would be useful only as a special case for this problem. The information already gained in the previous steps is enough to draw some conclusions about correlation for this data set.*

Using the results from steps 2-6 ,8 ,10, and 13, two possible conditional variables were indicated. Steps 6 and 8 indicated a rather strong conditioning effect from the problem-solving score. However, there was a good reason to doubt the effect of the *DEGREE* variable from step 10. Step 9 also indicates that there is not much numerical indication of interexpert correlation. Some numerical correlation was indicated from the results in steps 7, 11, and 12, the equivalent number of independent experts for the five correlated experts would be one expert. This expert would be added to the remaining 26 for a set of 27 pseudoindependent experts. Therefore, the final conclusion would be to use the 27 experts and consider them conditionally independent upon the problem-solving score variable which was formed at a more general level of detail than the raw information gathered.

■

15

Model Formation

In chapters 12 through 14, the emphasis was on investigation and preliminary analyses beginning with gaining familiarity with the information gathered from the elicitation. This chapter and chapter 16 focus on what might be termed final analysis procedures that have the goal of establishing interpretable conclusions. This chapter concentrates on forming models whose results provide inferences. Chapter 16 concentrates on forming aggregations or combinations of the experts' judgments also for inference purposes.

In this chapter, modeling techniques and suggestions for describing the experts' answers in terms of other variables are presented. These models are chosen based upon information already gained from the exploratory analyses done on the data base. Some modeling techniques use standard statistical procedures, such as least squares regression, as their foundation. These techniques fall under the heading of general linear models (GLMs). The multivariate structure of the data base lends itself to being modeled using multivariate methods such as factor, discriminant, and cluster analyses. However, as in previous chapters, the use of these techniques for final conclusions is not recommended because of the assumptions required in using them. Other, more applicable modeling techniques are based on decision analysis methods and can be described as conditional models.

In general, the model is a functional relationship between the answer variables, y, and the ancillary variables, x, and is described as

$$y = f(x) \quad .$$

General linear models define the expected value of the answer variable as a function of the ancillary variables:

$$E(y) = f(x) \quad .$$

GLMs also define the observed values of y as the function, f, of the observed values of the x's plus some error residual, ε,:

$$y = f(x) + \varepsilon \quad .$$

The above model is also applicable in discriminant analysis where f is the linear discriminant function.

Model formation based on conditional relationships between the answer variable and the ancillary variables can be written as

$$f(y/x) \quad .$$

Cluster analysis can be written in terms of conditional relationships between any and all variables of x and y:

$$f(y/x), \; f(x/x'), \; f(y/y') \quad .$$

Factor analysis can also be modeled in terms of defining new variables, z, from x and y:

$$z = f(x,y) \quad .$$

Special problem areas, such as granularity and the structure of the available data base, arise in model formation. Models can be formed at general or specific levels, depending upon the chosen granularity and the granularity inherent in the variables. Model selections are also integrally linked to the elicitation technique used in gathering the information. As with any analysis technique, the assumptions required for the selected modeling procedure such as cluster analysis must be examined and/or tested prior to application.

General Linear Models

General linear models (GLMs) refers to the models formed using the statistical method of least squares. The least squares method is so named because the model coefficients (b's in the equation below) are determined such that the squared distances between the values of y and the $E(y)$ are minimized. The commonly used techniques that use the least squares method are regression, analysis of variance, and their multivariate counterparts. Analysis of variance refers to the cases where the y's are continuous, numerical variables and the x's are categorical or rank variables. Analysis of variance is usually not applicable for models. Its uses are discussed in more detail in chapter 11. Regression usually refers to the cases where both y's and x's are continuous, numerical variables. Because the x's may be dummy variables, the term GLM better describes the models discussed in this section.

Full-Scale General Linear Models

In a full-scale GLM, the experts' answers are modeled as functions of the ancillary variables. The form of the GLM is

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m + \varepsilon \quad ,$$

where

y is the dependent variable and is one of the answer variables,
x_i (i=1,m) are the independent variables and include many, if not all, ancillary variables,
b_i (i=0, m) are the coefficients estimated in the GLM analysis that are used to determine the significance of the x_i variables, and
ε is the random error in the model, the residual that cannot be modeled by the other variables.

In a full-scale GLM, all the independent variables (the x's) are included in the model in the same form as they appear in the data base. There are procedures for screening out redundant variables and insignificant variables to form a streamlined model. Redundant variables are variables that contain overlapping information so that substituting one for another does not change the model. Insignificant variables are variables that do not predict the values of y so that their b values are not statistically different from zero.

The screening procedures for both types of variables are called stepwise regression and are available on most statistical packages that have GLMs (Snedecor and Cochran 1978: chapter 13). In a step-up procedure, a model is formulated by adding x's one at a time. The x's are added according to which ones best model or predict the dependent variable. In a step-down procedure, a model is formulated by starting with the full set of x's and eliminating, one at a time, those x's that contribute least to the model. In a stepwise procedure, both the step-up and step-down techniques are done simultaneously to select the best sets of x's that model the y.

Stepwise procedures are a convenient way to find the best model for y for any selected number of x's. However, there are cautions and assumptions necessary for using this procedure. Some cautions are discussed in the last section of this chapter. The assumptions are the same as for any GLM:

1. The x's (all the x's) are independent variables with no measurement errors.
2. The ε's are distributed as normal random variables with a mean of 0 and a variance of σ^2.

Violation of the first assumption occurs when one or more of the answer variables are included in the model as x's because the answers are considered as variables measured with error. There may be other x's that cannot be considered as measured without error. One way to avoid violating the first assumption is to include as x's only the variables that are measured without error. This, of course, limits model formation possibilities, which limits discovering some conditional relationships among the variables. The effect of including x's with errors is to underestimate the variance of the dependent variable. This underestimate has the most impact upon the variance associated with predictions made by the model. Approximations for the prediction variances are available (Booker 1978). However, predictions using GLMs in expert judgment applications are not usually necessary nor are they recommended.

Testing for compliance with the second assumption is relatively easy. A quick test can be done by plotting the residuals, ε's, on normal probability paper. If they plot as a

straight line, then the normality assumption holds. Many statistical packages have tests for normality, such as the *W* test. These can also be used to provide easy verification. However, the violation of this assumption is not as important as the violation of the first assumption.

A caution that needs to be mentioned in full-scale modeling is that poor models are possible when many or all the variables have missing values. If there are many missing values (from the experts) on many of the variables used in the model, an adequate model fit is not possible using GLM procedures. Appropriate models can be formed from the original data by collapsing, combining, or redefining variables. In changing the original variables to more general variables, the granularity of the model changes to a more general level. The next topic discusses such a model formulation.

Combination Models

In chapter 13, several analysis techniques were suggested for analyzing the ancillary data. From the knowledge of the relationships among these variables gained in those analyses, combinations of the ancillary variables are possible. Combinations of information from many ancillary variables can form new variables that represent scores or indices.

For example, information (many variables) is elicited on the experts' problem-solving processes. These variables may be in the form of rules, assumptions, heuristics, and problem-solving steps used by each expert. Each expert solves the same problem differently; therefore, many variables are gathered that have missing values for many experts (as shown in example 15.1). A full-scale GLM using all these different variables would not be possible because these variables would form a sparse matrix of information. If the variables could be combined to form scores with no missing values, then a GLM analysis would be possible, because each expert would have a value for this score. The score would be a new variable for the GLM.

EXAMPLE 15.1: ***Scoring Using the Anchoring and Adjustment Model***

Seven experts were interviewed in a face-to-face interview. The information gathered on their problem-solving processes indicated the use of four different assumptions. The matrix below gives the usage of these assumptions by the experts:

	Assumption No.			
Expert No.	1	2	3	4
1	x		x	
2		x		
3		x		
4			x	
5		x		
6				x
7	x			x

Not all the experts used the same assumptions. The assumption matrix has many holes and a GLM for these 7 experts and 4 assumptions would not be recommended.

The experts all began solving the problem with some initial impressions. These impressions could be considered anchors and the assumptions used could be considered as adjustments made on the impressions forming an anchoring and adjustment model (see chapter 3, *The Four Cognitive Tasks*). The following information on the experts' initial impressions of the problem is used to score the anchoring:

Expert No.	Initial Impression	Value Assigned to Anchor
1	Highly possible	2
2	Possible	1
3	Not likely	-1
4	Not sure about this	0
5	Can never happen	-2
6	Don't believe this	-2
7	Could be true	1

The following evaluation of the assumptions is used to score the assumptions as adjustments from the anchors:

Assumption No.	Evaluation	Value
1	Assuming this gives a pessimistic view	-1
2	Assuming this has no effect on the problem	0
3	Assuming this gives an optimistic view	1
4	Assuming this gives an optimistic view	1

To produce the final score, the matrix entries of the original assumptions are replaced by the assumption evaluations, the initial impression values are added, and the score is formed as follows:

	Assumption Value					
Expert No.	1	2	3	4	Anchor	Score
1	-1		1		2	-1+1+2 = 2
2		0			1	0+1 = 1
3		0			-1	0+-1 = -1
4			1		0	1+0 = 1
5		0			-2	0+-2 = 2
6				1	-2	1+-2 = -1
7	-1			1	1	-1+1+1 = 1

■

Because score or index variables are combinations of several variables, the granularity of the model with these new variables is more general than the original variables gathered. Thus, the results and their interpretations must be done at this more general

level. As in example 15.1, if the problem-solving score variable is found to be significant in modeling the answers from a GLM analysis, then the interpretation is that the answers are conditioned on this general score variable, not on the specific problem-solving features (assumptions and initial impressions) of the experts.

Combination variables can be formed in many ways. Some of the most commonly used ones are described below.

Anchoring and Adjustment Scores

This variable combination scheme uses the cognitive theory of problem solving which states that the expert anchors to an initial value or idea and then proceeds with a series of adjustments from that value (anchoring and adjustment heuristic). To model the expert's problem solving using this theory, the expert's initial impression (*good, bad, indifferent*) of the problem (as in example 15.1) is formulated into a new variable, and then the expert's adjustments (e.g., *up, down, neutral)* are formulated into one or more new variables. Adjustment variables can be formed from any of the relevant problem-solving information. The one illustrated in example 15.1 relates to assumptions made by the experts in their problem solving. The new anchor and adjustment variables can be easily quantified into ranks (1, -1, and 0). A final score or index for each expert is then found by summing up the new anchoring variable and the new adjustment variables.

Cumulative Scores

Cumulative scoring is a very general variable combination scheme that produces a final score variable at a general granularity. An example of a cumulative score is the counting up of all the problem-solving features used by the experts. This accumulation produces a score that reflects how much effort and thought each expert used in solving the problem. Of course, there is freedom to determine which problem-solving features are counted, and there is also the flexibility of weighting the various features to provide a weighted sum as a score. Example 15.2 uses the information from example 15.1 to illustrate the formation of a cumulative score.

EXAMPLE 15.2: ***Scoring Using Cumulative Scores***

Using the matrix of assumptions from the experts in example 15.1, a cumulative score is formed by counting up the number of assumptions used by each expert. No evaluation of the assumptions is done to establish different weights.

	Assumption No.				
Expert No.	1	2	3	4	Cumulative Score
1	x		x		2
2		x			1
3		x			1
4			x		1
5		x			1
6				x	1
7	x			x	2

The interpretation or meaning of the cumulative score can be vague. In this example, the meaning refers only to the number of assumptions used by the experts. This cumulative score variable may not be important for modeling of the answers because it provides very little and very general information on why the answers were chosen by the experts.

■

Collapsing Variables

Variables can be combined or collapsed together to form new variables with different, usually more general, interpretations. Example 15.3 illustrates how information on the expert's background relating to his education degrees and disciplines can be combined to form a more general variable reflecting all the information. There are two items worth mentioning in this process. First, the analyst must be careful not to impose his own views or interpretations on the original data in order to successfully collapse variables. Second, the granularity can change through several levels producing an extremely general variable that may be of little use in conjunction with other variables (at a different granularity) or be of little use in the interpretation of the results.

Example 15.3: ***Scoring by Collapsing Variables***

Background information on the education of seven experts is listed below. A new variable is formed by collapsing this information into a single composite variable representing the experts' education.

	Degree Code*		
Expert No.	BS	MS	PhD
1	1	3	
2	4	4	4
3	2		
4	6	4	
5	5	5	
6	3	3	3
7	5	4	4

**1=physics*
2=mathematics
3=nuclear engineering
4=mechanical engineering
5=civil engineering
6=electrical engineering

By assigning weights to the degrees (BS, MS, and PhD) and to the degree disciplines (the 6 codes*), the above information can be collapsed into an overall education score. The weights are calculated from ranks of importance for the degrees; i.e., a BS degree is not very desirable; an MS degree is; a PhD is only slightly better than an MS. The ranks are normalized so that they sum to 1.0:

Degree	Rank	Normalized Weight, w_d
BS	1	0.1
MS	4	0.4
PhD	5	0.5
	10	1.0

The same ranking, weight calculation process is done for the degree code (the discipline area). Here the engineering degrees were valued as more relevant.

Degree (Code)	Rank	Normalized Weight, w_c
Physics (1)	1	0.05
Mathematics (2)	2	0.09
Electrical engineering (3)	4	0.17
Civil engineering (4)	4	0.17
Mechanical engineering (5)	6	0.26
Nuclear engineering (6)	6	0.26
	23	1.00

The collapsed variable is then the combination of the degree weights and the degree code weights as follows:

$$\text{Collapsed variable for the } i\text{-th expert, } V_i = \Sigma w_c w_d \quad .$$

The values for the collapsed variable are given below:

	BS			MS			PhD				
Expert	w_c	w_d	+	w_c	w_d	+	w_c	w_d	=	V_i	V_i *new*
1	0.05	0.10		0.26	0.40		0	0.50	=	0.11	0.42
2	0.26	0.10		0.26	0.40		0.26	0.50	=	0.26	1.00
3	0.09	0.10		0	0.40		0	0.50	=	0.01	0.03
4	0.17	0.10		0.26	0.40		0	0.50	=	0.12	0.47
5	0.17	0.10		0 17	0.40		0	0.50	=	0.09	0.33
6	0.26	0.10		0.26	0.40		0.26	0.50	=	0.26	1.00
7	0.17	0.10		0.26	0.40		0.26	0.50	=	0.25	0.97

Because the maximum possible score for this weighting scheme is only 0.26 and the variation among the experts is small, the V_i's can be transformed according to the highest score as follows: V_i *new* $= V_i / 0.26$.

■

Multivariate Models

In chapter 14, some multivariate analysis procedures were used to aid in the detection of correlation and bias. These procedures are briefly described in chapter 11. They can also be used here in the model formation process. However, the assumptions and circumstances for applying these techniques makes their use limited. Specifically, factors from a successful factor analysis can be used as new variables in a GLM. Discriminant analysis models can be used for describing relationships between answer variables that are categorical in structure and ancillary variables. Cluster analysis can be used on the variables (rather than on the values of a variable) to model variate relationships.

Factors From Factor Analysis

A successful factor analysis on the set of ancillary variables will produce a set of new variables, called factors, that are combinations of the original variables from a shared information analysis. These factors could be used as new variables in a GLM. The key word here is *successful.* Success implies that the factor analysis produces factors from clear-cut subsets of the original variables and that the factors have an interpretation or meaning that is consistent with the variables that went into the factors' formation. Example 15.4 illustrates these concepts.

Example 15.4: ***Using Factor Analysis to Form New Variables***

There are 12 numeric, ancillary variables gathered from an elicitation of 15 experts. These 12 are all problem-solving variables: variables H_2, H_3, H_5, and H_{11} are variables describing heuristics used; variables A_7, A_8 and A_{12} describe assumptions used; variable R_1 is a rule of thumb used; and variables C_4, C_6, C_9, and C_{10} describe cues used from the problem. The data is as follows:

	Variable											
Expert	H_2	H_3	H_5	H_{11}	A_7	A_8	A_{12}	R_1	C_4	C_6	C_9	C_{10}
1	-1	0	1	0	-1	-2	-1	0	-1	0	-2	-1
2	0	0	0	0	-2	-3	-1	0	-2	-3	-2	-3
3	1	-1	-1	1	-3	-3	-1	0	-1	-1	-3	-3
4	2	1	1	2	-1	-1	-1	1	-1	-2	-3	-2
5	2	2	1	1	-1	-1	0	1	-2	-1	-1	-3
6	1	2	2	2	0	0	-1	1	-1	-1	0	0
7	-2	-1	-1	-2	0	0	-1	0	0	0	0	0
8	-1	0	0	-1	0	0	0	0	0	-1	0	0
9	-1	0	-1	0	1	0	0	0	0	-1	1	0
10	2	1	2	2	-1	0	0	1	-1	1	1	0
11	2	2	2	2	2	2	3	1	1	3	1	2
12	1	1	2	2	1	1	1	1	2	2	2	2
13	0	0	0	-1	2	2	3	0	3	2	3	2
14	0	-2	-2	0	2	2	3	0	3	2	3	2
15	0	-1	0	0	3	3	3	0	3	3	3	2

A factor analysis on the twelve variables using the principle components method resulted in the following factor loadings on 2 new factors:

Ancillary Variable	Factor 1	Factor 2
H_2	-0.23	0.85
H_3	-0.35	0.83
H_5	-0.21	0.87
H_{11}	-0.27	0.86
A_7	0.94	0.15
A_8	0.93	0.27
A_{12}	0.90	0.20
R_1	-0.23	0.93
C_4	0.96	-0.03
C_6	0.87	0.30
C_9	0.93	0.12
C_{10}	0.93	0.20

Factor 1 is a new variable that combines the information from the rule of thumb variable (R_1) and the 4 heuristic variables (H_2, H_3, H_5, and H_{11}). Factor 2 is a new variable that combines the information from the 3 assumption variables (A_7, A_8, and A_{12},) and the 4 cue variables (C_4,C_6,C_9, andC_{10}). Therefore, the original 12 variables can be reduced to only 2, factor 1 and factor 2. The values, or scores, for each expert for factor 1 and factor 2 follow:

Expert	Factor 1	Factor 2
1	-0.64	-0.64
2	-1.27	-0.87
3	-1.19	-0.84
4	-1.12	0.71
5	-0.99	0.79
6	-0.56	1.07
7	0.10	-1.47
8	-0.02	-0.83
9	0.13	-0.80
10	-0.32	1.17
11	0.78	1.69
12	0.60	1.21
13	1.39	-0.26
14	1.49	-0.76
15	1.64	-0.17

The original twelve variables are now represented by two variables. The meaning or interpretation of these 2 new variables are factor 1 represents the assumptions and

problem cues used indicating the problem background items; factor 2 represents the rules used in solving the problem.

Some information is lost by using the two new factors in place of the 12 original variables. A measure of how much is lost can be determined from most factor analysis procedures by examining the percentage of the total variance (from the twelve variables) explained by the 2 factors. In this example, that percentage is 87%. Therefore, 13% of the variation is lost using the 2 factors as new variables.

This loss of variance or information can also be interpreted as a change in the granularity represented by the change from 12 variables to 2. If these 2 new variables are used in a GLM, the results from that model would have a more general interpretation than a GLM using the original 12 variables.

■

Discriminant Analysis

Discriminant analysis determines a linear discriminant function, f, which determines how well the ancillary variables map or classify the categories or groups defined by the answer variable. The assumptions for using discriminant analysis are (1) the variables must follow a multivariate normal distribution, and (2) the answer (dependent) variable must have a structure of groups or classes, e.g., multiple choice responses, qualitative responses, or naturally occurring numerical groupings. The first assumption is very restrictive and would not be expected to hold true for expert judgment data. The second assumption is often very applicable to expert answers. However, the answers should be distinctively and cleanly clustered and the reasons for this clustering should be evident from the exploratory analysis results. Example 15.5 illustrates model formation using this technique under these conditions.

Example 15.5: *Using Discriminant Analysis in Model Formation*

Eleven experts are asked to estimate the likelihood of a specific event under certain conditions in a nuclear reactor. The response or answer mode given to them was the Sherman-Kent scale (in chapter 7), and their percentage answers follow:

(10, 10, 90, 80, 99, 20, 10, 80, 99, 90) .

Each expert provided the type of work environment (listed 1-3 for variable W), used certain cues (listed as 1-5 for variable C), used certain formula calculations (listed 1-3 for variable F), used some rules of thumb (listed 1-5 for variable R), provided information on background (listed 1-5 for variable B), provided the highest educational degree (listed 1-3 for variable D), provided information on work experience (listed 1-5 for variable E), and used certain assumptions (listed 1-3 for variable A). The data are as follows:

Expert No.	*ANSWER*	*W*	*C*	*F*	*R*	*B*	*D*	*E*	*A*
1	10	2	3	3	1	1	1	5	1
2	10	2	2	1	3	2	2	5	1
3	90	1	5	3	5	3	3	5	1
4	80	3	3	3	3	4	3	4	2
5	99	2	5	2	3	5	2	5	3
6	20	2	2	3	2	5	2	4	1
7	10	2	1	1	3	4	2	3	3
8	80	3	2	3	3	3	2	3	3
9	90	2	4	2	3	2	2	2	2
10	90	1	5	3	4	1	1	2	3
11	20	2	1	3	2	1	1	1	1

The goal is to determine which variables are good discriminators for the five different answer categories (10, 20, 80, 90, 99). If a discriminant analysis is run on this data set, most packages will either give a warning, an error message, or will not complete the calculation for this data because there is a singularity present in the variable set. The presence of the singularity means that not all the eight variables can be used to model the answer because one or more of the variables are exact functions of one or more other variables. In this case variables eliminating variables *E* and *W* would allow a solution. This problem is common in data sets with either large numbers of variables and/or small numbers of experts.

Two solutions are possible: (1) run a cluster analysis or correlation analysis on the variables (as done in chapters 13 and 14) to learn more about the variable structure or to find the singularity; or (2) use available information to make some discriminant analysis runs on subsets of the variables. In this example, the cluster analysis in example 15.6 will indicate a solution, but there is already a clue on how to categorize the eight variables from their definitions. Four of them (*A*, *C*, *F*, and *R*) are items that the expert used to solve the problem, and four of them (*B*, *D*, *E*, and *W*) are features of the expert's background and work environment. Two discriminant analysis runs were made using the two sets of variables.

I. Variables A, C, F, and R as discriminators for the answers

Most discriminant analysis programs supply a table of how well this discrimination was done and of any observations that were misclassified. The table for this analysis follows:

Expert No.	From *ANSWER*	Classified Into *ANSWER*
1	10	10
2	10	10
3	90	90
4	80	80
5	99	99
6	20	20
7	10	10
8	80	80
9	99	99
10	90	90
11	20	20

In other words, no misclassifications implies that the four variables did a good job of discriminating the five answer classes.

The model between the four variables, x_i, and the categories of the answer variable, y_j, is very similar to the GLM:

$$y_j = l_{0j} + l_{1j}\, x_1 + l_{2j}\, x_2 + l_{3j}\, x_3 + l_{4j}\, x_4 \ ,$$

where l_{ij} are table of coefficients below:

Linear Discriminant Function Coefficients

	ANSWER				
Term in Model	10	20	80	90	99
Constant	-241.9	-337.4	-538.2	-833.0	-438.0
$A = x_1$	26.1	29.1	38.7	47.7	36.0
$C = x_2$	15.3	12.3	20.1	32.1	27.0
$F = x_3$	105.5	130.3	160.4	191.9	134.2
$R = x_4$	100.2	118.2	149.4	185.4	132.0

This table specifies the model of the experts' answers as classified by these four variables. This model could be used to predict the answer (10, 20, 80, 90, or 99) of a twelfth expert given his values for the four variables. However, this prediction capability is not necessary nor important here. The goal is to determine which variables are influential in determining the answers. To attempt inferences beyond this goal would be stretching the limit of the information contained in the data gathered.

II. Discriminant analysis using *W*, *E*, *B*, and *D*.

The following classification table resulted from this analysis:

Expert No.	From *ANSWER*	Classified Into *ANSWER*
1	10	10
2	10	10
3	90	90
4	80	80
5	99	99
6	20	20
7	10	20*
8	80	80
9	99	90*
10	90	90
11	20	20

The * on experts 7 and 9 indicates that if these four variables were good discriminators, then expert 7 should have answered with a 2 and expert 9 should have answered with a 9. While some may argue that this is not too bad a misclassification of the answers, the four variables in the first run of this example did a better job with no misclassifications. Also, with so few experts, even two misclassifications are not considered a good result.

Therefore, the results of these discriminant analyses indicate that the first set of variables were good discriminators of the answer variable.

■

Because the function determined in discriminant analysis is a linear function, a model from this technique should be similar to a GLM. In other words, if a particular set of ancillary variables is found to be significant in affecting the answer variable in a GLM, that same set should also be significant linear discriminators for the answer variable in a discriminant analysis. If the results do not match, then perhaps assumptions required for one or both procedures were violated. As a general rule of thumb, the GLM is a more forgiving and more widely applicable procedure. It is recommended over discriminant or even cluster analysis.

Cluster Analysis

The primary use of cluster analysis in chapter 13 was as an exploratory analysis tool where determinations were made about how the data formed various clusters or groups. Using cluster analysis as a modeling tool is done by determining variable relationships according to how the variables are clustered or grouped. Cluster analysis can also be used as a premodeling tool for discriminant analysis or for GLM. In the discriminant analysis case, a cluster analysis is done on the answer data to determine if the data forms clean, definite clusters that form the categories for the discriminant analysis like the one in example 15.5. The results from a cluster analysis on the variables can then be used to determine which variables would be good discriminators, as illustrated in example 15.6. The results from variable clustering can also be used to set up a GLM.

Example 15.6: ***Using Cluster Analysis in Model Formation***

Using the responses and ancillary information from the 11 experts in example 15.5, a cluster analysis was performed on all nine variables using the centroid method to determine clusterings of the variables. The results of the first cluster formation indicated that the two clusters should contain the following variables:

Cluster 1	Cluster 2
$(A, F, R, C, ANSWER)$	(B, D, E, W)
Interpretation	Interpretation
The problem-solving variables and answers	The background and experience variables

This result indicates a strong relationship between the answer and the problem-solving variables. The implication is that these four ancillary variables would be good discriminators for the answers (as used in example 15.5) and would be good independent variables for GLM with *ANSWER* as the dependent variable. The four variables in cluster 2 are not good candidates for discriminator or independent variables for *ANSWER*.

The cluster analysis done indicates other cluster formations for three to nine clusters. Deciding on which formation to use is done by using different criteria. One such criterion is the measure of the proportion of the original variance as explained by each cluster formation. For this data, 72% for the variation is explained by the following four-cluster formation:

Cluster	Variables
1	$(ANSWER, A, C, R)$
2	(B, D, E)
3	(F)
4	(W)

Another criterion is common sense or logical interpretation of the cluster formation. Although the above four-cluster formation indicates a sizeable percentage of the variation, it makes little sense to have clusters 3 and 4 with only one variable each. A better clustering is the two-cluster formation where the variables in each cluster have common definition, even though only 50% of the variation is explained by the two clusters.

■

Using cluster analysis as a stand-alone modeling tool provides only a conditional model. The model is represented as a cluster formation like the two-cluster formation in example 15.6. This model can be described using a general conditional form of the answer variables, y, and the ancillary variables, x and x', as

$f(y/x)$ for cluster 1

and

$f(x/x')$ for cluster 2 .

Conditional models such as these are described further in the next section.

Conditional Models

The functional relationship between the answer variables and the ancillary variables can be modeled using techniques from the decision science community. These modeling techniques also serve as the response modes for the experts' answers. In other words, the model selection dictates how the questions are asked of the experts. Therefore, the model selection process becomes an integral part of the elicitation process (chapter 7) and must be determined prior to gathering the data. The analysis philosophy of part III of this book has been to explore the gathered data with the freedom of letting the data and its information content direct the analyses. Formulating a model prior to the gathering of the data has the disadvantage of losing some of this freedom and runs contrary to the analysis philosophy.

Two of the more popular decision analytic methods are described below. Their popularity stems from several advantages: (1) ease of implementation, (2) wide applicability, and (3) track record of success. The first is based on Saaty's pairwise comparison technique, and the second is a general technique based on decomposition of the problem using decision trees or diagrams.

Saaty

The Saaty pairwise comparison method yields a set of relative weights comparing the items from a set of competing alternatives. This method is, therefore, a decision-making tool. The resulting comparison weights are not probability or likelihood estimates. However, they could represent relative likelihood comparisons. A more detailed description, including advantages and disadvantages of Saaty's method, is given in chapter 11.

The Saaty method acts as both a model formation tool and as a response mode. The model formed is a very qualitative one structured as a hierarchy of descriptive conditions under which the alternatives are compared. The pairwise comparisons are made at all levels of the hierarchy. Therefore, conditional comparisons are made among all the conditions in all the levels; also comparisons are made of the alternatives, given the various conditions.

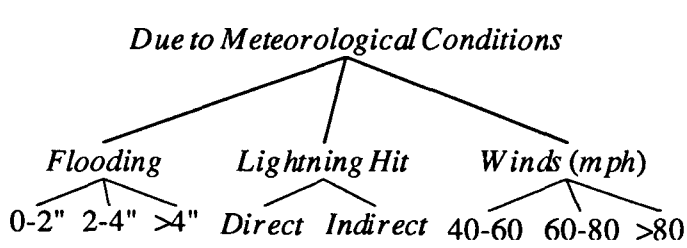
The analysis of the hierarchy to determine the weights is a straightforward set of calculations based on the eigenvalues of the matrix of comparisons at each level. For each level, the weights for the items being compared are calculated from the eigenvectors of the principle eigenvalue. Weights are propagated through the levels of the hierarchy by multiplication so that the final weights for the bottom level of competing alternatives can be determined.

The weights at any level are relative comparisons only, and have no numerical interpretation. For example, if item A has a weight of 0.25 and item B has a weight of 0.50, then A is not half as important (or half as likely or good) as B. The only interpretation is a qualitative one: item A is more important or likely than item B.

Example 15.7 illustrates how the conditioning and calculations are done for a simple, two-level problem. Most problems have a much more complex structure than only two levels. The more levels and items per level in a problem, the more relative comparisons are necessary, and the more time required for evaluation. For each level of m items, $m(m-1)/2$ comparisons are required.

Example 15.7: ***Using Saaty's Pairwise Comparison Technique for Model Formation***

In reactor safety analysis, the loss of off-site power (LOSP) can lead to other consequences and important events. Therefore, it is important to investigate how LOSP can occur. One set of possible events responsible for LOSP is meteorological conditions--floods, lightning, and high winds. Each of these conditions can occur with varying degrees of impact on the likelihood of LOSP. Questions such as how much flooding? where does the lightning hit? how high are the winds? are important for evaluating the impacts. To answer these questions, a hierarchy of the impacts on LOSP is represented as follows:



The comparisons of the bottom eight specific meteorological conditions are made conditional on the levels above them. Although a numerical Saaty scale given in chapter 11 is used to make the comparisons on a pairwise basis and the resulting weights are numerical, the interpretation of the weights is qualitatively done. This interpretation preserves both the qualitative structure of the input information and its granularity.

The pairwise comparisons are done as follows:

LEVEL 1: LOSP due to meteorological conditions.

LEVEL 2: Which of the 3 general conditions is most likely to cause LOSP?

Flooding vs lightning?	1/4	(meaning lightning is more likely than flooding using a 4 on the Saaty scale)
Flooding vs winds?	1/3	
Lightning vs winds?	2	

These three comparison form an upper triangular matrix that is filled in by placing 1's on the diagonal and reciprocal answers on the lower triangular portion.

	F	*L*	*W*
Flooding	1	1/4	1/3
Lightning	4	1	2
Winds	3	1/2	1

The weights for F, L, and W are found by normalizing the eigenvector of the principle eigenvalue of this matrix and are (0.12, 0.56, 0.32), respectively.

LEVEL 3: **Which of the specific conditions is most likely to occur given the general condition of flooding occurs?**

0-2 inches vs 2-4 inches given a flood?	2
0-2 inches vs more than 4 inches?	6
2-4 inches vs more than 4 inches?	3

The weights for the flooding matrix are (0.67, 0.22, 0.11).

Which of the specific conditions is most likely to occur given the general condition of lightning occurs?

Direct hit vs indirect hit given lightning hit? 1/4

The weights for the lightning matrix are (0.20, 0.80).

Which of the specific conditions is most likely to occur given the general condition of winds occurs?

40-60 mph given high wind?	2
60-80 mph?	8
Greater than 80 mph?	5

The weights for the winds matrix are (0.75, 0.16, 0.09).

FINAL WEIGHTS: By multiplying the weights of levels 2 and 3, the final (unconditional) weights for the eight specific conditions are obtained as follows:

(0.08, 0.03, 0.01, 0.11, 0.45, 0.24, 0.05, 0.03) .

The results indicate that the most likely cause of LOSP is an indirect lightning hit; the second most likely is winds 20 to 40 mph. These results then would be the important conditions and causes of concern for LOSP.

It is *not* correct to make statements like the indirect lightning hit is almost twice as likely as is the 20-to-40-mph winds. Such numerically based conclusions are not valid with a relative (nonnumeric) comparison method.

■

One application of this technique is to use it as a weighting scheme for aggregating multiple experts (chapter 16). A hierarchy can be done representing a conditional model of information of each expert. Multiple experts can be aggregated by combining their hierarchies. This combination is done by making a new level for the experts and placing their hierarchies beneath it.

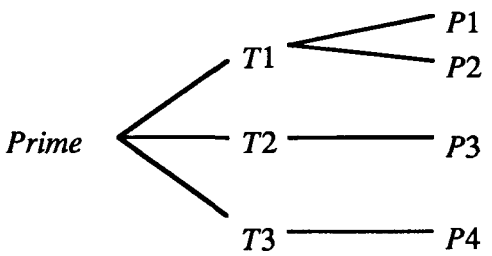
Decomposition Diagrams

The hierarchical structure of the Saaty method described in the previous section is not the only way to diagram conditions. A more established way is to form an event tree or decision tree structure with branches and nodes (Raiffa 1970). The tree usually begins as a qualitative description of alternatives or events connected with branches. The tree can be quantified into a decision diagram, like an event tree, by assigning values or distributions to the branches. If values or distributions are propagated through the branch pathway, then the values or distributions at the end of the tree will have a numerical interpretation. If only qualitative information is propagated through the tree branches, then the results have a qualitative interpretation such as the results from Saaty's method. In either case the diagram serves as a model describing the conditions relevant to the final results.

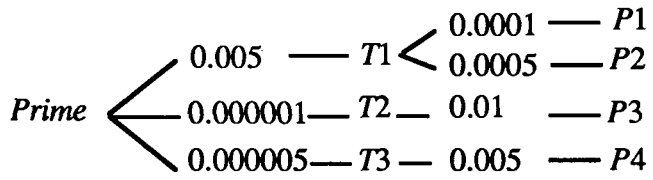
Using the diagramming approach as a decomposition tool helps the expert evaluate a complex or large problem in smaller pieces, one piece at a time (U.S. NRC 1989). The advantages of using the decomposition principle are cited in chapter 5 and in Kahneman and Tversky (1982). An example of the use of the diagram method is given in example 15.8.

EXAMPLE 15.8: ***Using Decomposition Diagrams for Conditional Modeling***

An expert is asked to determine the probability of failure for an important, but never observed, event *prime*. The failure of *prime* depends on the temperature and the pressure of the system. The expert is asked to provide a set of potentially fatal temperature values (T_1, T_2, and T_3) and a set of potentially fatal pressure values (P_1, P_2, P_3, and P_4). The expert then diagrams the relationships between these temperatures and pressures as they affect *prime*. The diagram is as follows:

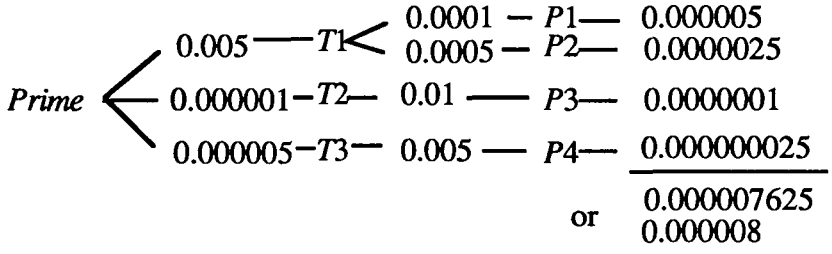


This diagram is a qualitative model of *prime*. To quantify it, the expert provides the likelihoods or probabilities of the temperatures and the pressures.



The values listed for the pressures could be independent of the temperatures (if pressure is independent of temperature) or the pressure values could be conditional on the given temperature (e.g., the chance or P_1 given T_1 is 0.0001). In the above diagram, the pressures are found to be conditional on temperature because the P_2 values change depending on whether T_1 or T_2 is used.

The resulting probability for *prime* is determined by multiplying the values across pathways:



■

The example 15.8 illustrates a conditional model construction for a single expert providing single estimates for the branches. In chapter 16, a similar example illustrates how to combine multiple experts using the diagram approach. In chapter 17 a similar example illustrates how to propagate probability distributions on the branches.

Model Selection Suggestions and Cautions

This chapter has described some of the ways of formulating functional relationships between the answers and the ancillary or conditioning variables. In the examples provided, some of the pitfalls and shortcomings have been illustrated. This section summarizes these cautions and makes some suggestions for model selection.

1. A fully descriptive model of the answer in terms of the ancillary variables (such as GLM) is useful for determining and defining which variables are influential. Such models provide better understanding of the answers given. However it is not recommended that these models be used for prediction purposes because the independent variables may not be measured without error. The variable relationships established from these models should be consistent with the results from the exploratory data analysis (chapters 12 and 13).
2. The steps necessary to formulate some models may require interpretations or quantification of variables by the analyst. Interpretations should be avoided,

and quantification should be held to a minimum and be done carefully. If the information content or granularity of the elicited or gathered data is insufficient for forming the desired models, then the model selected is inappropriate. If the data is all highly qualitative or uncertain in nature, perhaps the model selection should not be done.

3. If more than one modeling procedure is done, consistency of the modeling procedures chosen is important. The examples listed in this chapter indicate how the multivariate procedures are compatible with GLM models. Likewise, the conditional models are compatible among themselves. Only in a very general way are the statistical and conditional types of models compatible.
4. Granularity is important when collapsing or reformulating variables for modeling. As noted in several examples, the new variables formed are more general in information content, and the conclusions resulting from the analyses must also be stated in the more general terms. Granularity is also important when making relative comparisons, such as with the Saaty method. Even though numbers are used in the analysis, the results can only be interpreted in relative comparison terms.
5. When using GLMs or multivariate procedures, it may be necessary to consult a statistician. The regression and multivariate methods results are sometimes difficult to interpret. The procedures themselves require prepackaged programs that may be difficult to run, and there may be several different methods available for the analysis. For example, there are many different ways to do a cluster, factor, and discriminant analyses, and results do depend on the method chosen.
6. The conditional models serve as both elicitation and analysis techniques. They are somewhat difficult to use because they are time-consuming and may require training of the experts.
7. The philosophy of this handbook is that models should not be selected prior to gathering the data, nor should they be selected based on popularity or convenience of the analyst. The elicitation is complicated and compromised by many choices, and the experts must be motivated to participate in its implementation. It is easier on the experts and provides more flexibility for the analyst if the model choice is based upon and directed by the data and the elicitation.
8. The primary purpose of model formation is to describe the answers in terms of the other variables (information). The results of the modeling effort provide the appropriate conditional interpretations for the answers. Enough other information should be available from the elicitation and from the data base so that these relationships can be understood and so that they can be used for aggregating experts (chapter 16), handling the inevitable uncertainties (chapter 17), and making the final inferences (chapter 18).

16

Combining Responses--Aggregation

This chapter is divided into three major sections. First in *Choosing the Aggregation Scheme* we present several recommended and widely accepted aggregation estimators and also methods for forming aggregation distributions. Next in *Application Environments* is presented aggregation using the above methods in various problem settings and environments that involve multiple experts, decision makers, and analysts. Last, the similarities of solving the aggregation problem and of solving the problem of characterizing uncertainties as discussed in chapter 17 is presented.

Choosing the Aggregation Scheme

One of the most difficult analytical problems in expert judgment is how to combine the experts' answers into either a single value (estimator) or a single distribution of values. This aggregation is a **mathematical aggregation**. There is no shortage of techniques available for mathematical aggregation; however, many of these techniques impose restrictions on the data, the experts, the analyst, and on the interpretations of results. The techniques presented in this chapter for both estimators and distributions reflect the general philosophy of the book: from the elicitation side, the aggregation should not require the experts to be force-fitted into unknown or uncomfortable modes of providing data; from the analysis side, the aggregation should not be so complex that a doctorate in mathematics is required to understand and use it. To achieve both goals, not all available techniques are represented in this chapter. However, most of the general types of estimators are given in some form. For a complete report on all the different types of mathematical aggregation schemes, see a survey paper by Genest and Zidek (1986)

Aggregation Estimators

The most commonly used method of combining a set of answers is to calculate a single value from a formula called an estimator using all the values in the data set. The most popular estimators are the mean, median, and geometric mean. As discussed below, each estimator has its own properties, making it appropriate for different applications.

Regardless of the estimator chosen, it should be accompanied by an estimate of the variance of that estimator. For example, both the mean and the variance of the mean are estimated.

The mean, or arithmetic average of the values, has the advantage of an easily calculated variance of the mean. For any mean value from a large sample (e.g., 30), the variance for that mean is estimated by the variance of the data divided by the square root of the sample size, n.

$$var\ (mean) = \sigma 2(mean) = \left(\sum_{i=1}^{n} (x_i - mean)2 \right) / n(n-1) \ .$$

This variance formula is from the central limit theorem which claims that the formula is valid for any sample of sufficient size. However, if the sample size, n, is small, as it is in most expert judgment applications, or if the data set is not unimodal and symmetric, the central limit theorem does not accurately determine the variance of the mean. For samples of size less than 10 (even for symmetric, unimodal data), the theorem does not work well. Therefore, in these cases it is recommended that an alternative estimate for the variance of the mean be used. One such alternative is to calculate the variance of the mean from a simulation such as the bootstrap.

A second noteworthy property of the mean is that it gives equal weight to each datum. This equal weighting implies that if one expert gives an answer that is far away in value from the rest and if there are only a few experts providing estimates, then the mean value will be greatly influenced by that extreme value. This result may not be a desirable especially if that extreme value is questionable or seems unreasonable.

To overcome the influence of extreme values in forming an aggregation estimate, the median or geometric mean can be used. Both of these estimators are influenced by the central values of the data set and are not so influenced by the extreme values.

The median is the 50th percentile value. It is defined as the middle of the data set such that half of the data is larger than the median and half of the data is smaller than the median. If the data set is of odd sample size (n is odd), then the median is calculated by finding the central value of the ordered data points. If the data set is of even sample size, then the median is the average or halfway between the two center values. There is no general or convenient formulation of the variance for the median. As suggested in earlier chapters, this variance can be found using simulation techniques such as the bootstrap or Monte Carlo methods.

Another interesting reason for using the median in expert judgment applications can be found in the studies of Kahneman and Tvsersky (1982). They have shown that when experts are providing numerical answers, they are really estimating the median value rather than the mean. If all the values for the answers given by the experts form a distribution that is symmetric in shape (cutting the distribution in half results in one half being the mirror image of the other), then the mean and median are the same. However, most expert judgment data distributions are not symmetric (they are skewed with the mode shifted to one side of the center and the other side having a long tail); consequently, the mean and median cannot be considered the same. Therefore it is a common (and recommended) practice to consider the answers given by experts as median values.

The geometric mean is an average of the data values based on a logarithmic scale whereas the simple mean is based on a linear scale. The geometric mean is formed by the

product of all the n values raised to the $1/n$th power. The variance for the geometric mean is also not readily available and can be determined using simulation methods. The geometric mean is popular for use in expert judgment applications because of its log-based nature. Many estimates elicited from experts are small values or probabilities that fit better in a log scale than in a linear scale.

Example 16.1 shows how the above three different aggregate estimators give different results for the data set of seven experts estimating a very low probability value of an event. The variance estimates are from a bootstrap sampling done on the data. The FORTRAN computer program for this bootstrap simulation is given in Appendix D.

EXAMPLE 16.1: ***Comparison of Three Aggregation Estimators***

Seven experts provided the following probability estimates for a rare event.

Expert	Estimate
1	0.00001
2	0.00010
3	0.00010
4	0.10000
5	0.00001
6	0.00001
7	0.00005

Expert 4's estimate is several orders of magnitude larger than the others. The three aggregation estimators--the mean, the median, and the geometric mean--produce the following results.

Mean	=	0.014
Median	=	0.000050
Geometric mean	=	0.000091

The influence of expert 4's large value is quite noticeable in the large mean value. The median and geometric mean are much closer and do not reflect the influence of expert 4.

Using the central limit theorem (CLT), the variance of the mean for this data set is

$$var(mean)_{CLT} = var(data)/7 = 0.0014/7 = 0.000054 \quad .$$

The variances for the geometric mean and the median can be found easily by using the bootstrap simulation (BS). (The code for this simulation is in Appendix D.) Here 1000 samples of size 7 were formed by sampling with replacement from the original data. The mean, median, and geometric means were calculated for each of the 1000 samples. The calculations for the variances from the 1000 means, the 1000 medians, and the 1000 geometric means are as follows:

$$
\begin{aligned}
var(mean)_{BS} &= 0.00016 \\
var(median)_{BS} &= 0.000060 \\
var(geometric\ mean)_{BS} &= 0.000000095
\end{aligned}
$$

A large discrepancy is evident in comparing the bootstrap variance for the mean with the variance of the mean from the central limit theorem. This data is a small data set and is highly skewed (with the extreme value from expert 4). It is not expected that the central limit theorem would give accurate results for such a data set. It is interesting to note that in this case the variance for the mean from the CLT is the same as the bootstrap variance for the median.

In comparing the bootstrap variances for the three estimators, the mean has the largest, the median is smaller, and the geometric mean has the smallest.

In conclusion, the small sample size and extreme value of expert 4 makes the mean estimator and its theoretical variance inappropriate. Either the median or geometric mean estimates are fine for this data. In either case, a variance estimate for that estimator must come from a simulation such as the bootstrap.

■

Because the mean estimator weights each datum equally and the median does not give weight to the extreme values, some analysts prefer to use a weighted average or mean. Each datum (expert answer) is given its own individual weight, and the mean is calculated as

$$weighted\ mean = \sum_{i=1}^{n} x_i w_i / \sum_{i=1}^{n} w_i \quad .$$

The advantage of this estimator is that the analyst can control which values (or experts) influence the estimator. The variance for the weighted mean is also available from theory; however, due to the small sample sizes and potential skewness of most expert judgment data sets, simulation is again advised for determining the variance. The biggest disadvantage is that the weights must be determined for each expert.

Determining Weights

Determining weights is not an easy process. It often requires information about the experts and how they arrived at their answers. It can lead to the dangerous situation where the analyst imparts his knowledge and influence (perhaps erroneously) to the results.

However, there are ways of determining weights based on the data itself, on qualitative comparisons of the experts and ancillary data, and on information from model formations. There is also a simple rule of thumb for determining weights and that is to use equal weights.

Data-based determinations

Weights can be determined from the data itself using no other information. For example, if one expert gives an extreme value relative to the others, that expert can be assigned a lower weight than the others. To illustrate this use of the data to determine

weights, example 16.2 shows the effects of different weights on a weighted mean estimator. In this example, expert 4 must be given a weight of 1/1000 of the other six experts in order to produce a weighted mean value comparable to the median or geometric mean. The purpose of this example is not to imply that the goal is to achieve the same value for the weighted mean as for the median or geometric mean. This example merely indicates how the three estimators are weighting the seven responses.

EXAMPLE 16.2: ***Using the Weighted Mean Estimator***

Using the data from example 16.1, different sets of weights are proposed for the seven experts. In each case expert 4 is given a smaller weight relative to the other six. The other six are all given the same larger weight, and expert 4 is given a weight of 1. The effect on the weighted mean estimates is shown below.

$W_{1,2,3,5,6,7}$	W_4	Weighted Mean
1	1	0.014
2	1	0.0039
10	1	0.00086
50	1	0.00021
100	1	0.00013
1000	1	0.000055
1	0	0.000047

For cases where the other six are given weights 1000 times that of expert 4, the weighted mean is comparable to the median and geometric mean values. The mean with expert 4 eliminated ($w_4 = 0$) is also comparable with the median and geometric mean.

■

If the analyst truly felt that a weight as low as weight 1/1000th or even 1/50th that of the others was warranted, why would that person be considered an expert. There is probably some underlying reason for the extreme value given by that expert. Rather than eliminating him from the sample, it is better to discover why that expert gave such an extreme answer rather than resolving the issue analytically with outrageous weights. The reason for the extreme value can be made by reviewing the rationale recorded by the experts. The reason may also be found by reviewing the preliminary data analysis results as prescribed in chapters 12 through 15. These results should contain information about this expert and his answer relative to the others.

If no evidence can be found that this expert used different assumptions, cues, problem-solving methods, or any information different from the others, then there is no reason for giving him a low weight or eliminating him from the sample. His answer is just as valid and reasonable as the others. In this case, it would be better to use an estimator that has a wide variance to reflect the wide range of data values. The bootstrap results from example 16.1 indicate that the variances are wide for the mean and median.

Using only the data to determine weights can lead to the dangerous situation of the analyst trying different weights to achieve some goal such as the elimination of a particular

expert. Therefore, other methods of determining weights are recommended. These methods rely more on finding and using the reasons why experts' estimates differ rather than on focusing on mere numerical differences.

Saaty weight determinations

Instead of using the gathered data to help determine the weights for the experts, it is sometimes desirable to determine the weights before the answers are given. The Saaty pairwise comparison method is helpful in this determination and can be used by the decision maker, the analyst, or even the experts themselves to determine expert weights.

The determinations are based on other information about the experts. The major advantage of the Saaty method is that this information can be qualitative in nature, but the results are a set of numerical weights. However, care must be taken in the use and interpretation of these resulting weights. In forming the weights, quantification of the original information has changed the granularity from general to specific. The resulting weights should not be used as numerical values. They are only relative comparisons in numerical form. However, relative comparisons can be used to help determine other, numerical, weights for a weighted mean. If the weighted mean value is calculated using the Saaty weights, it should be accompanied by a caveat such as *this mean is the result of relative weights that are values from qualitative comparisons and are not exact numerical values*. Because even exact numerical weights are highly uncertain in value or fuzzy in nature anyway, such a caveat would not be unusual.

A FORTRAN program in Appendix A for a single level (evaluation) of the Saaty method can be used to determine the weights of the experts. The user (decision maker, knowledge engineer, analyst, or expert) supplies the comparisons of the experts on a pairwise basis. The resulting relative weights are normalized to sum to 1.0. Example 16.3 illustrates how this can be done.

Example 16.3 Using Saaty's Method to Determine Weights

The decision maker wishes to determine weights for the seven experts in example 16.1. He is familiar with their qualifications and will make the comparison according to those criteria before he sees the answers that were elicited. Based on his knowledge of the seven, he compares them as follows:

1 versus 2 ----Better
1 versus 3 ----Same
1 versus 4 ----Same
1 versus 5 ----Better
1 versus 6 ----Better
1 versus 7 ----Better

2 versus 3 ----Worse
2 versus 4 ----Worse
2 versus 5 ----Worse
2 versus 6 ----Worse
2 versus 7 ----Worse

3 versus 4 ----Same
3 versus 5 ----Better
3 versus 6 ----Better
3 versus 7 ----Same

4 versus 5 ----Same
4 versus 6 ----Better
4 versus 7 ----Same

5 versus 6 ----Same
5 versus 7 ----Worse

6 versus 7 ----Worse

Using the Saaty method (and code in Appendix A), these verbal comparisons are translated using the following recommended numerical scale:

$$\begin{aligned} \text{better} &= 2.72, \\ \text{same} &= 1.00, \\ \text{worse} &= 0.37, \end{aligned}$$

and the resulting relative weights for experts 1-7 are

$$(0.23,\ 0.06,\ 0.19,\ 0.17,\ 0.10,\ 0.08,\ 0.17) \quad .$$

The consistency ratio for this matrix is 0.06, which is less than the critical value of 0.10 and indicates good consistency among the comparisons.

If the seven expert answers are combined using these weights, the result is

$$\sum_{i=1}^{n} x_i w_i / \sum_{i=1}^{n} w_i = 0.017 \quad .$$

Instead, the weights can be used for determining other numerical weights. The weights of experts 1, 3, 4 (the extreme-valued expert), and 7 are high relative to those of experts 2, 5, and 6. A 2-to-1 weighting of the high group over the low group is suggested. With this weighting, then

$$\sum_{i=1}^{n} x_i w_i / \sum_{i=1}^{n} w_i = 0.018 \quad .$$

This value is not much different from the original mean value of 0.014 or the weighted mean using the Saaty weights (0.017). It is interesting that expert 4 is in the high-weight group, making the weighted mean estimate even more influenced by his answer.

■

The results from example 16.3 may appear to be disappointing. The decision maker's information about the experts only added to the problem of over-influence of the extreme value given by one of the experts. Thus the decision maker did not have any information about this expert that would suggest that the experts solved a different problem. In fact, the information used by the decision maker tended to reinforce the idea that expert 4 should be included just as any other expert.

Model-based determinations

In supplying the weights in example 16.3, the decision maker compared the experts using a single, cumulative criterion representing the knowledge that he had about the experts. These comparisons are therefore conditions on the experts' estimates. Other conditions such as the rationale information recorded in the elicitation session may be important. This information was formulated into a data base in chapters 12 and 13 and was

also used to formulate models in chapter 15. Weights can be found by using the conditional variables found to be significant in model formation.

CONDITIONAL MODELING. The Saaty method is a conditional type of model for use in formulating relative weights for the experts. Any variables found significant in other models can also be used as conditional variables and put into a hierarchical structure for analysis in the Saaty method. If only a single conditional variable is found important, the hierarchy is a single level and the Saaty method is convenient to use. If a hierarchical structure of many variables is too complex, becoming not practical, then other weight-determining methods in this chapter should be used. Hierarchies of a few levels with a few items in each level are manageable by the Saaty method.

The method has the advantage of making pairwise comparisons among experts either by a simple *better, same,* or *worse* comparison or by a convenient numerical scale such as the Saaty scale (given in chapter 11) or Sherman-Kent scale (given in chapter 7). Therefore, quantitative or qualitative conditional variables can be used in this procedure. Example 16.4 illustrates how a significant variable from a general linear model is used to obtain relative weights for the experts using a simple scale of *better, same,* or *worse.* Here the analyst uses the conditional variable to determine the weights rather than using his own knowledge or judgment about the experts.

EXAMPLE 16.4: ***Using Conditional Variables and Saaty's Method to Determine Weights***

Eight experts provided answers to a likelihood comparison question using a linear scale from 0.0 to 1.1. A problem-solving variable, ps_1, was created as a cumulative score of several heuristic and cue variables (example 15.1). A general linear model analysis indicated that ps_1 was the best variable (and the most significant) in predicting the answers, v_1. The second best variable for predicting the answers was an experience variable, *yrt*, specifying the number of years that the expert had worked in the particular field of the problem. The data follow.

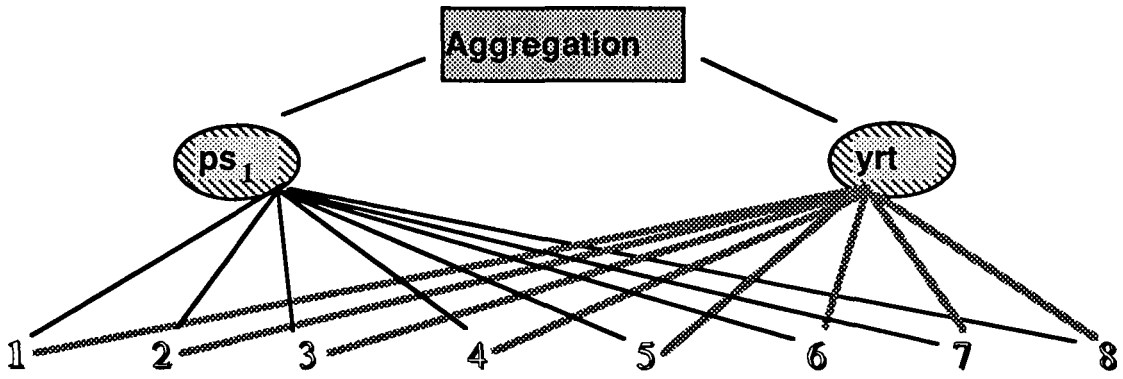
v_1	ps_1	*yrt*
0.20	-2	5.49
1.00	2	6.00
0.04	-3	2.00
0.00	-2	2.00
1.10	6	1.99
0.45	0	4.50
0.00	-3	3.00
0.75	2	4.90

Using the values of ps_1 and *yrt*, pairwise comparisons follow where

> ----------is greater (better) than,
= ----------is equal, and
< ----------is less (worse) than.

ps_1	*yrt*	ps_1	*yrt*
1 < 2	1 < 2	3 < 6	3 < 6
1 > 3	1 > 3	3 = 7	3 < 7
1 = 4	1 > 4	3 < 8	3 < 8
1 < 5	1 > 5		
1 < 6	1 > 6	4 < 5	4 > 5
1 > 7	1 > 7	4 < 6	4 < 6
1 < 8	1 > 8	4 > 7	4 < 7
		4 < 8	4 < 8
2 > 3	2 > 3		
2 > 4	2 > 4	5 < 6	5 < 6
2 < 5	2 > 5	5 < 7	5 < 7
2 > 6	2 > 6	5 < 8	5 < 8
2 > 7	2 > 7		
2 = 8	2 > 8	6 > 7	6 > 7
		6 < 8	6 < 8
3 < 4	3 = 4		
3 < 5	3 > 5	7 < 8	7 < 8

These two variables can be used as two items in one level of a hierarchy:



The hierarchical structure is used to combine the influence of the two variables into a single aggregation measure for weight determination. The above pairwise comparisons of the eight experts for each conditional variable form the entries of the two matrices (one for each variable) needed for the analysis. Each matrix results in a set of weights for the experts. These two sets are combined using weights assigned for the two variables. From the GLM done on these two variables, ps_1 accounted for most of the model variation, and *yrt* added very little. A weighting for the two variables might be 0.9 for ps_1 and 0.1 for *yrt*.

Using the Saaty single matrix code in Appendix A, the weights for the eight experts for ps_1 and *yrt* are given below. These two sets of weights are combined using the 0.9, 0.1 split on the two variables.

ps_1 Weights	*yrt* Weights	Combination $= 0.9 \cdot ps_1 + 0.1 \cdot yrt$
0.08	0.20	0.09
0.18	0.26	0.19
0.05	0.06	0.05
0.08	0.06	0.08
0.15	0.05	0.14
0.16	0.12	0.16
0.08	0.09	0.08
0.22	0.16	0.21

Following is the weighted mean value using the combination weights.

v_1	Combination	$v_1 \cdot$Combination
0.20	0.09	0.02
1.00	0.19	0.19
0.04	0.05	0.00
0.00	0.08	0.00
1.10	0.14	0.15
0.45	0.16	0.07
0.00	0.08	0.00
0.75	0.21	0.16
		0.59

This mean value is larger than the original mean of 0.44. The reason for this discrepancy is simply due to the choices of the variables for weight determinations. Both variables have large values for large answers, v_1; therefore, the weights will be large for large values of v_1. The final mean value is inflated due to this effect. This may seem like a weighting scheme with a built-in bias. The scheme is based on conditioning which can give a bias. The validity of the scheme lies in using conditioning variables that are important in determining the answers.

■

Example 16.4 brought up one problem inherent in weighting schemes. Sometimes the method for determining the weights induces a bias into the results. The weights in that example biased the resulting mean on the high side simply by the method used to determine the weights. The analyst or decision maker determining weights can also induce biases. Other methods and examples follow that tend to minimize these biases.

GLM MODELING. For example, problem solving variables have been found to be important in determining answers (Booker and Meyer 1988a and Meyer and Booker, 1987b). Example 16.5 illustrates how one such problem-solving variable can be used to form weights for the experts based upon the residuals from a linear model regression analysis done on that variable and the answer variable. The residuals measure how far away each answer is from the model (line) predicted by the problem-solving variable.

Therefore, one possible set of weights is the inverse of these residuals. This weighting scheme gives larger weights to values with smaller residuals.

The advantage in using residuals to determine weights is that the residuals can come from a model of more than one variable. Therefore, the impact of several conditional variables can be simultaneously incorporated into the weight determination. However, as is generally true of multivariate models, the influence of the most important one or two variables has the greatest impact in the model that impacts the residuals and which ultimately impacts the weights.

EXAMPLE 16.5: ***Using Residuals to Determine Weights***

Weights for the eight experts in example 16.4 can be determined from a model of their answers. A general linear model analysis indicated that the problem-solving variable, ps_1, was very significant in predicting the answers, v_1. The residuals from a regression of v_1 on ps_1 follow.

v_1	ps_1	Residual	Weight=1/Residual
0.20	-2	-0.095	10
1.00	2	0.137	7
0.04	-3	0.115	9
0.00	-2	-0.058	17
1.10	6	-0.023	43
0.45	0	-0.045	22
0.00	-3	0.007	141
0.75	2	-0.038	26
			275

Weight	Weight•v_1
10	2.0
7	7.0
9	0.4
17	0.0
43	47.3
22	9.9
141	0.0
26	19.5
	80.06/275 = 0.31 = weighted mean

The mean of the original data is 0.44, and the median is 0.45. Therefore, the residual weighting scheme did change the mean value, but not significantly, from 0.44.

■

DIRECT ESTIMATION. Other conditional variables that were found to be important in model formation can be used to determine weights. Specifically, these

variables may be qualitative- or classification-type variables that were analyzed in cluster analyses and found to be important in cluster formations.

Weights can be determined from these model variables by direct estimation. This requires knowledge of how to use qualitative or class variables to form quantitative weights. As seen in similar quantification problems, the Saaty method is useful for this type of quantification. Also as previously noted with this type of quantification, the granularity can change . It is easy to transform the words into numbers and then to use those numbers for interpreting the results, forgetting that the original data is nonnumeric and has no numerical meaning. Therefore, any weights formulated from qualitative or class information should be considered relative weights, as in example 16.3.

With direct estimation, the decision maker or analyst can also form relative weights without the formal Saaty technique. The criteria (variables) chosen for the comparison can be evaluated according to how they cluster the data, as in example 16.6. Here the classification variable is important because it discriminates between the experts according to a critical assumption made in solving the problem. Such a clear-cut distinction would provide a valid reason for eliminating expert 4 from the data set. Based on expert 4's use of this assumption, he was actually solving a different problem from the others. He was also solving a different problem from the one being asked.

EXAMPLE 16.6: ***Using Direct Estimation from Cluster Model Variables to Determine Weights***

One of the critical assumptions made in determining the probability of a rare event was not discovered until the exploratory data analysis and data-base formation was done. The seven experts made explicit assumptions regarding the constancy of temperatures. Some assumed that the temperatures would be constant (assumption = 1), some assumed it would vary negligibly (assumption = 2), and some assumed it would vary greatly (assumption = 3). Originally, the decision maker assumed that the temperature variation would not be great and wanted the problem solved without large temperature variation. He did not realize that this assumption might be important to the problem. However, a cluster analysis revealed that the seven answers were clustered according to this assumption.

Estimate	Assumption /Cluster
0.00001	1
0.00010	2
0.00010	2
0.10000	3
0.00001	1
0.00001	1
0.00005	2

Only expert 4 made the assumption of the large temperature variability. Therefore, expert 4 was actually solving a different problem from the other experts and from the one intended. A weight determination was made according to this assumption as follows.

Estimate	Weight	Estimate•Weight
0.00001	2	0.00002
0.00010	1	0.00010
0.00010	1	0.00010
0.10000	0	0.00000
0.00001	2	0.00002
0.00002	2	0.00002
0.00005	1	0.00005
	9	0.00031/9 = 0.000034

This weighted mean value is not much different from the unweighted mean value with expert 4 deleted (example 16.2). Therefore, the weights for experts 1through 3 and 5 through 7 made little difference.

■

The basic idea is to make weight determinations using variables or conditions that influence the experts' answers as determined from models. As with any weight determinations, the methods presented here must to be done carefully. Because these variables are known to be important in the experts determination of their answers, they can induce a bias into the weights that will result in a bias in the weighted mean value. Also, care must be taken in using qualitative or more general leveled information to formulate weights. Such formulation can cross levels of detail (granularity is not constant) and make accurate interpretation of the results tricky at best, or incorrect at worst.

One final word is needed on weight determinations using the methods presented here or elsewhere. As seen in many of the examples in this chapter, regardless of how good the choices for the weights are, in most cases the weights make little difference in the value of the weighted means. The question then becomes, why bother with weights? The answer follows.

Equal weights

The best recommendation to date on the weight determination problem is to use equal weights. This idea is not new (Seaver 1978); however, several studies have indicated that it is still the best idea (Genest and Zidek 1986).

Equal weights are best for combining individual responses from experts. Also, as demonstrated in the remaining chapter sections, equal weights are best for combining distributions.

Unless some unusual circumstances arise that clearly indicate why and how some different weights should be used, equal weights should be used. Example 16.7 below summarizes the lessons from the first sections of this chapter. These lessons enforce the idea of equal weights for the two applications used as examples in this chapter.

EXAMPLE 16.7: ***Summary of Weight Determinations***

In examples 16.1 through 16.3, the seven experts were asked to estimate the probability of a rare event. Expert number 4 gave an answer that seemed to be different (much larger) than the other six. In those examples, the mean (equal weights), median, and geometric mean were calculated. Also, several different weighting schemes were calculated using various weights and the Saaty method to determine weights. Those results are summarized below.

Estimator	Weighting Scheme	Estimate
Mean	Equal	0.014
Median	Equal	0.000050
Geometric mean	Equal	0.000091
Weighted mean	w_4=2	0.0039
Weighted mean	w_4=10	0.00086
Weighted mean	w_4=50	0.00021
Weighted mean	w_4=100	0.00013
Weighted mean	w_4=1000	0.000055
Weighted mean	w_4=0	0.000047
Weighted mean	Saaty - decision maker	0.017
Weighted mean	Saaty - model	0.000034

Even with all these methods, there are still only two basic results:

1. Estimates of the order of 0.01 resulting from equal weighting of expert 4 with the rest.
2. Estimates of the order of 0.00001 resulting from deleting expert 4 from the set with equal weighting for the rest.

The results from simply increasing the weight of expert 4, w_4, are of little use except to demonstrate that severe weights are needed to lessen the impact of expert 4. The clear decision in this example is either to include or exclude expert 4 entirely.

From the knowledge about the assumptions used in solving the problem in example 16.6, it was discovered that the other six experts made acceptable assumptions for the particular problem but that expert 4 made an unacceptable assumption. This is valid reason for deleting expert 4.

It should be noted that the results could have been completely reversed. Expert 4 may have been the only expert making an acceptable assumption and that was why his answer differed from the rest. In evaluating the results of an elicitation session, the information on the conditions under which all experts answered the question must be tested and evaluated. Hopefully, some problem areas such as the expert using questionable assumptions, drifting from the question, or solving a different problem than asked, can be corrected during the elicitation. It is vital that all the experts answer the same question (and therefore, estimate the same quantity).

In examples 16.4 through 16.5, eight experts estimated the likelihood of an event. Here there are no experts that appear to give responses different from the rest, but there are

two distinctive clusters of answers from the experts (experts 1, 3, 4, 6 and experts 2, 5, 8). In this case, any centroid type of estimator (mean, median) will give a result that is between the two clusters of data (where no actual data exists). A weighting scheme in this case is useful to pull the aggregation estimator toward one cluster or the other. However, the same cautions about choosing and using weights in this case still apply. Following are the results for this data set and the chosen estimators.

Estimator	Weight Scheme	Estimate
Mean	Equal	0.44
Median	Equal	0.45
Weighted mean	Saaty, 2-variate	0.59
Weighted mean	GLM residuals	0.31

All of these estimates are close in value. The residuals tend to lower the original mean value because the regression model is more influenced by the larger sized cluster, which is the lower valued cluster. There is a bias built into the Saaty weightings that favors the larger values. Neither of these weighting schemes are based on conditions that definitively favor one cluster over the other.

In examining the original data base, no other conditions are apparent for making such a distinction. In other words, there is little explanation why the two clusters emerged and no reason to attempt to eliminate one of them from influencing the final estimate.

In conclusion, the equal weighting (simple mean) is appropriate here. The uncertainty in the data is such that the experts spanned the entire available range of values (response mode) for their answers. The mean estimator will give a value in the center of the range and between the two clusters.

■

It is hoped by examining the two problems in 16.7 that the steps are evident for determining weights for the experts. The choices in both these cases came down to a simple, equal-weighting mean estimator. By using the data from the data base, many possible explanations were examined for apparent differences in the answers. In the first problem, a valid difference was found. In the second, no difference was found. The basic rule for searching and determining such differences is that *all* experts must be providing estimates (answers) to the *same* problem.

Aggregation Distributions

Rather than restricting the expert to providing a single, best estimate for an answer, many analysts prefer the experts provide either a range of values or an entire distribution of values (a set of possible values with corresponding probabilities of their occurrence). Many experts prefer to give multiple values. These multiple values do not imply that the experts provide values for different conditions. Giving multiple values for different conditions would be the case of providing single estimates for multiple problems.

If the problems in handling single-valued estimates from several experts appeared difficult, the problems are equally difficult in handling distributions or multiple values from

many experts. In this section the emphasis is placed on the case where distributions (probability distributions) are available from the experts. In the next section on dealing with uncertainties, the topic of handling ranges is also included (see also chapter 17).

The aggregation methods presented here are in keeping with the philosophy in this book. In particular, the methods should be comfortable and easy for the experts to use. The analyst does not have the luxury of deciding upon his favorite method and then force-fitting the experts into its use. The methods should also be easy for the analyst to use. These are the methods illustrated in the examples. Genest and Zidek (1986) reference others.

Using Bayesian methods

The philosophy of using Bayesian methods incorporates the idea that all information can and should be used to form final estimations (chapter 11). Bayes Theorem provides the mechanics for the combination of information from various sources. The information can be in several or differing forms, such as (1) information from measured data that is combined with expert judgments, (2) information from the present that is combined to update previous information, and (3) information from experts that is combined with information from the decision maker (DM) or analyst.

The first case is the more classical reliability problem that is common in PRA applications (Martz and Waller 1982). The second case is modified to include combinations of several information sources (experts) without regard to a time line and is examined below. The last case is examined in more detail in the *Application Environments* section of this chapter.

The major difficulty in applying Bayesian methods is that the information (all information sources) must be quantified into probability distribution forms. Common forms such as the beta distribution are easy to use because (1) they require only two parameters to estimate the entire distribution, (2) they range from 0.0 to 1.0 in values, which is the range of probabilities, and (3) they combine in a mathematically tractable form with other common distributions such as the binomial to form the resulting or posterior distribution.

Example 16.8 illustrates how two estimates provided by each expert can be combined into a single beta distribution that represents the probability of a single failure, p, of an event.

EXAMPLE 16.8: ***Bayesian-Based Aggregation of Distributions***

Six experts provided estimates of 5th and 95th percentile values for the distribution of a probability, p, for an event. Based on previous studies (Kahneman and Tversky 1982), these percentiles really only represent a fraction of the true uncertainty. Therefore, the given 5th percentile was chosen to represent the 30th percentile and the given 95th percentile was chosen to represent the 70th percentile.

Assuming that each experts' percentiles come from a beta distribution with parameters x_0 and n_0, the values of those parameters are given below:

Percentiles		Beta Parameters	
30th	70th	x_0	n_0
0.0010	0.004	0.82	242.59
0.0001	0.010	0.19	11.04
0.0010	0.100	0.18	1.52
0.0001	0.010	0.41	41.80
0.0050	0.050	0.41	8.64
0.0100	0.050	0.65	15.30

The values of the beta parameters can be obtained from the beta subroutines in the beta Monte Carlo code in Appendix B. These parameters have specific meaning in a binomial process: x_0 represents the number of failures in n_0 trials, each with a probability p of occurrence.

To combine the six different beta probability distribution functions into a single beta distribution function, $f(p:x,n)$, where

$$f(p:x,n) = \sum_{i=1}^{6} w_i \; f_i(p_i:x_{0i},n_{0i}) \quad ,$$

with equal weights for w_i, parameter sets are combined as follows (Winkler 1968):

$$x = \sum_{i=1}^{6} x_{0i} = 2.66$$

and

$$n = \sum_{i=1}^{6} n_{0i} = 320.89 \quad .$$

There is also a convenient Bayesian method for using this new beta as a prior distribution to combine with a distribution of a DM to make a resulting posterior. This combination is illustrated in example 16.10 below.

This new resulting beta distribution is the aggregation distribution for the experts. It has a mean and variance of

$$x/n = 0.008$$

and

$$x(n\text{-}x)/(n^2(n+1)) = 0.00003 \quad ,$$

respectively.

■

Using assumed distributions

There are other analytical methods for combining assumed distribtuion functions. An aggregation function is based on the formation of a joint distribution of all the experts. Using this joint or composite distribution as a prior distribution, the decision maker can combine the joint distribution of all the experts with his distribution, $f_0(x:\theta)$ in a Bayesian manner as follows:

$$f(x:\theta) = c(x:\theta) \cdot f_0(x:\theta) \cdot f_c(x:\theta) \quad ,$$

where $f_c(x:\theta)$ is the joint distribution function with parameter θ for the random variable x of all k experts and where $c(x:\theta)$ is a joint calibration function for all k experts. Another way of expressing that equation if all the experts are independent is by

$$f(x:\theta) = c(x:\theta) \cdot f_0(x:\theta) \cdot f_1(x:\theta) \cdot f_2(x:\theta) \cdot \ldots \cdot f_k(x:\theta) \quad ,$$

where the individual distribution functions of the experts are factored out. Another way of saying this is that the joint distribution of all the experts is expressed as the products of the individual expert distributions. If the expert's distributions are not statistically independent, then this factorization is not possible.

Joint distribution functions are usually assumed, easy-to-work-with distributions. In example 16.8, the expert information was characterized using a beta distribution for each expert. Other distributional forms produce mathematically tractable combinations and can be assumed for the experts. One common choice is the normal distribution (Winkler 1981). If very few estimates are provided by the experts, or if the experts provide a simple range of values, then the uniform distribution is appropriate. In this case, the posterior or resulting distribution will also be uniform .

If the information provided by each expert can be assumed to follow a normal distribution, then a multivariate normal distribution is the combination distribution. Using a multivariate normal has the advantage of allowing for a specified correlation structure among the experts. The disadvantage is that this correlation structure must be known.

Example 16.9 illustrates the information requirements necessary from the experts to use the multivariate normal. As seen in this example, it may be difficult and uncomfortable for experts to provide estimates for the required parameters of the normal, the mean, and the variance. It is also difficult to obtain a correlation structure for the experts. Usually such estimates are assumed by the DM or analyst. In using the normal, it is assumed that the quantity of interest (quantity being estimated) follows a normal distribution. Most such quantities, such as failure rates and probabilities do not tend to be symmetric in shape nor unimodal (Meyer and Booker 1987b) and cannot be considered normally distributed.

EXAMPLE 16.9: ***Multivariate Normal Distribution for Aggregation***

Three experts are asked to estimate a temperature range that is critical for a component failure (Winkler 1981). The experts agree that this temperature range should be distributed as a normal distribution, and they understand how to estimate the mean of this

normal. The variances of the temperature ranges are known and are not estimated by the experts. Also the correlation structure of these three experts is known to be as follows:

$$\text{Correlation of experts 1 and 2} = \rho_{12} = 0.60$$
$$\text{Correlation of experts 1 and 3} = \rho_{13} = 0.50$$
$$\text{Correlation of experts 2 and 3} = \rho_{23} = 0.60$$

The experts' mean values, μ, and corresponding known variances, σ^2, follow:

$$\mu_1 = 60 \qquad \sigma^2_1 = 36$$
$$\mu_2 = 62 \qquad \sigma^2_2 = 25$$
$$\mu_3 = 70 \qquad \sigma^2_3 = 49$$

The multivariate normal mean vector, μ, and variance-covariance matrix, Σ, are

$$\mu = (60, 62, 70)$$

and

$$\Sigma = \begin{vmatrix} 36 & 18 & 21 \\ 18 & 25 & 21 \\ 21 & 21 & 40 \end{vmatrix} ,$$

where the diagonal elements of Σ are the variances and the other elements are $\sigma_i \sigma_j \rho_{ij}$.

In order to combine these estimates, a Bayesian technique is used by assuming a diffuse (imparts little added information) prior to combine with the multivariate normal (the joint distribution of the experts) to produce a posterior distribution for the temperature range. With e being the unit vector of 1s, the posterior mean and variance are

$$\mu_{post} = e^t \Sigma^{-1} \mu / e^t \Sigma^{-1} e = 62.02$$

and

$$\Sigma_{post} = e^t \Sigma^t e = 22.83 \quad .$$

The μ_{post} is actually a weighted mean value with weights corresponding to the following:

$$w_i = \sum_j a_{ij} / \sum_m \sum_j a_{mj} \quad ,$$

where a_{ij} are the ijth elements of Σ^{-1}. In this case the weights are

$$w_1 = 0.263, \; w_2 = 0.669, \text{ and } w_3 = 0.068 \quad .$$

The third expert has a small weight due to his large variance and the high correlations with the other two.

It is difficult to imagine a situation where the variances are known. If the variances for the experts are estimated, then the posterior variance calculation is much more complicated (Winkler 1981).

■

The multivariate normal distribution requires estimates for its parameters that may be difficult to obtain or may rely on assumptions. However, the normal is a convenient distributional form for a combined distribution. Certainly other distributions could be used to make multivariate combinations; however, these would also require assumptions and estimates of their parameters. Other multivariate distributions are not as easy to form or to work with as the normal, and there is no more precedence for using them than for the normal. If multivariate forms become too complex or difficult to use, then empirical (data-based) distribution forms are suggested in conjunction with simulation techniques.

Using empirical distributions

If experts prefer to provide several values or are comfortable giving percentiles without specifying distribution forms, then empirical or step distribution functions can be constructed from these estimates. Sometimes interpretations or rules of thumb are needed to form the distributions. The rules below are based on several studies (Kahneman and Tversky 1982).

1. When experts provide 5th and 95th percentiles, they really are only giving 30-40th and 70-60th percentiles.
2. When experts provide maximum and minimum values they really are only giving 5-10th and 95-90th percentiles.
3. When experts provide their best central estimate, they really are giving a value that corresponds to a median (50th percentile) rather than a mean.
4. When experts provide a variance, they really are only representing less than half of the variance.

Empirical distributions are formed from percentiles (i.e., those percentiles using the rules above rather than what the expert provides directly). The percentiles provide the points for a step cumulative distribution function. The empirical probability distribtuion function, $f(x)$, is also formed from these percentiles in the shape of a histogram. For example, an expert provides

Best estimate ----------0.3
5th percentile----------0.2
95th percentile ----------0.4
Minimum value ----------0.1
Maximum value----------0.5

for a random variable, x, with a possible ranges of values from 0.0 to 1.0. The distribution of x is then a histogram composed of a series of rectangles with starting values at x_l, ending values at x_u, and having heights of f:

x_l	x_u	f
0.0	0.1	0.5
0.1	0.2	2.5
0.2	0.3	2.0
0.3	0.4	2.0
0.4	0.5	2.5
0.5	1.0	0.1

These values for f are calculated by

$$f_i = (\text{percentile level of } x_u)/(x_u - x_l) \quad .$$

The areas under these rectangles ($f_i \cdot (x_u - x_l)$) sum to 1.0. There is a program in Appendix C that forms these empirical distributions for each expert and combines them by a user-specified aggregation function using Monte Carlo simulation.

The aggregation function is usually of the form that combines the expert's individual empirical distribution functions, $f_i(x)$ and the decision maker's empirical function, $f_0(x)$, by the following weighting scheme (Winkler 1968):

$$f(x) = w_0 f_0(x) + \sum_{i=1}^{k} w_i f_i(x) \quad .$$

Determining these weights has the same difficulties mentioned in the first section of this chapter. The rule of thumb for these weights is also the same: equal weights are best (Seaver 1978). Distributions can be empirically combined using this equation without specifying distributional forms (such as normals, beta, or uniforms) or parameters. Example 16.10 illustrates weighting methods for empirical distributions from two experts.

EXAMPLE 16.10: ***Empirical Distribution Aggregation***

Two experts provide the following estimates for a random variable x, without specifying any distributional form such as the normal or beta. The given values are interpreted (used) according to the rules of thumb listed above.

Given As	Used As	Expert 1	Expert 2
Best estimate	Median	0.30	0.10
5th percentile	30th	0.20	0.05
95th percentile	70th	0.40	0.15
Minimum value	5th	0.10	0.01
Maximum value	95th	0.50	0.20

Here the absolute possible range of values for x is from 0.0 to 1.0. The resulting empirical distributions from the code in Appendix C follow.

Expert 1		
x_l	x_u	f
0.0	0.1	0.5
0.1	0.2	2.5
0.2	0.3	2.0
0.3	0.4	2.0
0.4	0.5	2.5
0.5	1.0	0.1

Expert 2		
x_l	x_u	f
0.0	0.01	5.00
0.1	0.05	6.25
0.2	0.10	4.00
0.3	0.15	4.00
0.4	0.20	5.00
0.5	1.00	0.06

The following pooled distribution results from using a weighted sum (with equal weights of 0.5) in a Monte Carlo simulation of the two empirical distributions.

Percentile Level	Value
1st	0.042
5th	0.079
10th	0.099
20th	0.13
30th	0.15
40th	0.18
50th	0.20
60th	0.23
70th	0.25
80th	0.28
90th	0.32
95th	0.43
99th	0.58

Mean = 0.22
Variance = 0.011
Standard deviation = 0.11

■

If a pooled (weighted) distribution of just the experts is to be found and used alone in a non-Bayesian context such as in example 16.10, then the weighting scheme used is very important. Even if a DM is added to the set, he becomes like another expert in terms

of influence on the final results. Of course the choice of weights can shift influence slightly toward or away from the DM or any other expert. However, as seen before, results will not differ by much unless drastically different weights are used, and such weights require justification. Lacking good reasons for drastically unequal weights, equal weight assignments are recommended.

Using Monte Carlo simulation

The use of Monte Carlo simulation has already been demonstrated in the previous section for aggregation by using empirical distributions from experts. There are other applications where Monte Carlo simulation is useful for aggregation.

One commonly used method of eliciting and modeling expert information is through a series of distributions, each representing a phase or characteristic of the problem. The idea behind this method is to decompose the problem into simpler parts. Information is then gathered on each of the parts. The information can be in the form of distributions of variables of interest in the various parts. The difficulty becomes how to recombine the information from all the parts for each expert and how to combine all the information from the experts.

In many applications the parts are structured by a tree diagram. This type of diagram is commonly used in decision analysis applications. Some diagrams could be hierarchical in structure while others could be quite complex with feedback loops and ill-defined connections. Whatever the structure, logic must be used in order to determine how the parts should fit together. It can be very difficult to accurately diagram a complex problem, but, like any model formation process, it is very important to do it correctly.

In combining different structures for several experts, another difficulty arises. Each expert will have a structure that is uniquely his own. These structures can be viewed as conditional models, $f(x|c)$, for the final result or answer, x. Because the structures are complex, one way of combining them to determine $f(x|c)$ is through simulation. Usually the information provided for the various parts of the structure is characterized by assume [1] or empirical distributions. The connections between the parts (however complex) can be characterized by arithmetic expressions similar to Boolean expressions for the failure or reliability of a system fault tree. For each expert, simulation is then done by sampling from each of the distributions in his structure and combining the sampled values using the expression to form a result. A final or resulting distribution of the whole structure is found by performing the sampling and calculations many times. The resulting distributions for each expert can then be combined using any of the above distribution aggregation methods including Monte Carlo simulation. Example 16.11 illustrates how this is done for two experts solving a problem by decomposition.

EXAMPLE 16.11: ***Decomposition and Aggregation by Simulation***

Two experts are asked to evaluate the probability of an event E. The question posed to them provides a limited set of conditions for determining the probability. These conditions involve establishing specific values: temperature (T), pressure (P), and flow rate (F). Both experts agree that these are the three most important conditions; however, each differs in his assessment of how the three conditions affect the event and what probabilities are associated with them.

The probability of the event $Pr(E)$ is found by summing the products of the conditional probabilities times the probabilities of those conditions:

$$Pr(E) = Pr(E|T)\cdot Pr(T) + Pr(E|P)\cdot Pr(P) + Pr(E|F)\cdot Pr(F) \quad .$$

Expert 1 claims that the probability of T, P, and F are all uniformly distributed with a value of 0.01. Expert 2 claims that all three follow a unimodal distribution between the values of 0.0 and 1.0, with a reasonable range of 0.1 to 0.9. This information can be characterized as a beta distribution with 40th and 60th percentiles of 0.1 and 0.9, respectively, using the rules of thumb for ranges and percentiles.

The experts are asked to estimate the probability of the event given these values of T, P, and F. The probabilities associated with the ranges of the three conditions given are as follows:

	Expert 1	Expert 2
$Pr(E\|T_{hi})$	0.1	0.01
$Pr(E\|T_{lo})$	0.01	0.001
$Pr(E\|P_{hi})$	0.01	0.001
$Pr(E\|P_{lo})$	0.001	0.0001
$Pr(E\|F_{hi})$	0.01	0.01
$Pr(E\|F_{lo})$	0.001	0.001

Because a range was given for each condition, this range can be used to determine the 40th and 60th percentiles of some probability distribution. Again a useful form for distributions of probabilities is the beta. If these values are used to form betas, the following beta distributions result.

Expert 1

	Beta Parameters		
	x_0	n_0	Mean
$E\|T$	0.17	0.83	0.21
$E\|P$	0.18	4.32	0.04
$E\|F$	0.18	4.32	0.04

Expert 2

	Beta Parameters		
	x_0	n_0	Mean
$E\|T$	0.17	0.83	0.21
$E\|P$	0.18	4.32	0.04
$E\|F$	0.18	4.32	0.04
T,P,F	0.11	0.22	0.50

Using Monte Carlo simulation of beta distributions for the above equation for $Pr(E)$, a distribution for the probability of E is found for each expert. For expert 1, $Pr(T)$,

$Pr(P)$, and $Pr(F)$ are constant values (0.01), and the conditional probabilities of E given T, P, and F are the beta distributions listed in the above table. For the expert 2, $Pr(E)$ is found as the sum and products of the beta distributions listed in the above table. Those empirical distributions have the following characteristics.

	Expert 1	Expert 2
Mean	0.0028	0.045
Standard deviation	0.0031	0.094
1st percentile	0.00001	0.00016
5th percentile	0.00006	0.00080
10th percentile	0.00013	0.0016
20th percentile	0.00025	0.0032
25th percentile	0.00034	0.0040
50th percentile	0.0016	0.0080
75th percentile	0.0042	0.043
80th percentile	0.0053	0.064
90th percentile	0.0082	0.14
95th percentile	0.0097	0.24
99th percentile	0.011	0.47

Using the empirical aggregation program in Appendix C, the two expert's distributions can then be combined. Equal weights were used for the pooling of the two distributions and give the following results:

Mean	0.030
Standard deviation	0.060
1st percentile	0.00037
5th percentile	0.0010
10th percentile	0.0017
20th percentile	0.0028
30th percentile	0.0038
40th percentile	0.0049
50th percentile	0.0073
60th percentile	0.013
70th percentile	0.020
80th percentile	0.036
90th percentile	0.078
95th percentile	0.14
99th percentile	0.33

■

As indicated in example 16.11, the decomposition, aggregation problem is a multi-step, complex procedure. First, much information is needed from the experts. Then, that

information is formulated into distributions, requiring some assumptions. Next, the distributions must be combined for each expert. Finally, the aggregation of all the experts' combined distributions is done. All these steps involve assumptions, information transfer, and quantification, and all the problems associated with these such as imposing new information not originally present and changing granularity.

Granularity can change several times in this series of steps, making the interpretation of the final results meaningless. At best in the above example, the mean value and its standard deviation might be useful and meaningful for interpretation. However, this example is a relatively simple one. In practice, there are usually more than three simple conditions and more than two experts. Recombining such decomposed information requires more and more assumptions and makes final inference more and more difficult.

In spite of these obstacles, information is gained that can be used in the inference process. For instance, in Example 16.11, one interesting feature to note is that the experts agreed on the basic conditions affecting the event. This agreement indicates that they are solving the same problem in a similar way. They disagreed as to how these conditions affected the event on a more detailed level, and that disagreement translated into their different estimates. Such numerical differences can be interpreted by considering the wide ranges of values as representing the true uncertainty in estimating the event.

Application Environments

Thinking in terms of inference, it is important to understand who is making the inferences--the experts, the decision maker, the analyst, or all three? To answer this question, the problem or application environment becomes important. Choosing an aggregation method involves considering the application environment. Examples given below demonstrate the differences in methods for some application environments. The specific cases examined follow:

1. One expert and one decision maker (DM)
2. DM and several (n) experts
3. One analyst and n experts

When a DM is involved, it is because he has some information relevant to the problem, just as an expert would. The DM may be an expert of equal expertise, or he may have erroneous or dated information. The influence that his information has in the aggregation is up to him. However, his views are subject to change after seeing the information from the other experts. Because of this influence, it seems a logical assumption that the DM is not independent of the experts (Morris 1977, Genest and Zidek 1986).

On the other hand, the analyst is supposed to be independent of the experts acting as a neutral party. He is never to impart his own information or biases into any assumptions, definitions, or information transformations. This ideal is unrealistic; however, the steps and recommendations in this handbook are designed to minimize problem areas and approach the ideal.

Decision Maker and One Expert

In the situation where the DM has a problem for solving and has his own information (DM's prior), he consults a chosen expert (usually one he feels has more information than he does). That expert gives his information (expert data) to the DM. The DM then combines the information forming a posterior (Morris 1977). This type of aggregation problem is characterized by a Bayesian philosophy. The DM is imparting information into the final posterior in several ways:

1. The DM's prior
2. The choice of the expert
3. The aggregation of the information
4. The inference

As previously noted, when using Bayesian methods, the way of combining the information sources to form the posterior (in this case the DM's prior and the expert's data) can result in one or the other sources being emphasized. The DM has the power to determine which source is emphasized in the aggregation (item 3 above). If the DM feels uncomfortable with his prior, he can reduce its influence even to the point of using a noninformative prior that imparts little (but some) information into the aggregation process.

Another effect results from item 3 above. After seeing the expert's prior, the DM may revise his own prior to either match or differ from the expert's. This is also a part of the aggregation process because the DM has decided on how his information is to be combined with the expert's information. A simple example, 16.12, illustrates these effects.

EXAMPLE 16.12: ***Decision Maker and One Expert***

The Decision Maker (DM) has knowledge about an event. He has never seen nor heard of a particular component failing in 10 plants in a combined 60 years of operations. He asks his favorite plant operator to estimate a failure rate for this component. The expert estimates that there should be a possibility of one failure in 60 operation-years or one failure in 120 operation-years. Thus the expert gave a range of values from 1/60 -1/120 failures/operation-years. Using a Bayesian context for this problem, the DM can aggregate in many different ways.

Part 1: The DM considers his information as valid data and combines it with the expert's information using that as his prior. He uses the binomial process to characterize his information (0 failures in 60 years), and he uses the expert's prior information as a beta distribution with 40th and 60th percentiles of 0.008 and 0.017, respectively.

The DM finds that the mean of the expert's beta is 0.0205 with parameters x_0 = 0.677 and n_0 = 33.096. The posterior distribtuion resulting from the Bayesian combination of the DM's binomial and expert's beta prior is also a beta with mean equal to

$$x_0/n_0 = (0.677 + 0)/(33.096 + 60) = 0.677/93.096 = 0.0073$$

or a mean value of 1 failure in 137 operation-years and a variance equal to

$$x_0(n_0-x_0)/[n_0^2(n_0+1)] = 0.000077 \quad .$$

Part 2: The DM believes his information is incorrect and uses a noninformative prior for himself. The expert's information can be characterized as the data from a binomial process with 0.677 failures in 33.096 trials. With a noninformative prior, the resulting posterior distribution is a beta distribution with mean equal to

$$x_0/n_0 = (0.677 + 0.5)/(33.096 + 1) = 1.177/34.096 = 0.035$$

or a mean value of 1 failure in 29 operation-years and a variance equal to

$$x_0(n_0-x_0)/[n_0^2(n_0+1)] = 0.00095 \quad .$$

Part 3: The DM believes the expert is too cautious and decides to use a noninformative prior for the expert, relying more on the DM's information. The DM claimed 0 failures in 60 trials. With a noninformative prior on the expert, the resulting posterior is a beta with mean equal to

$$x_0/n_0 = (0.0 + 0.5)/(60 + 1.0) = 0.5/61.0 = 0.0082$$

or a mean value of 1 failure in 122 operation-years and a variance equal to

$$x_o(n_o-x_o)/[n_o^2(n_o+1)] = 0.00013 \quad .$$

Part 4: After seeing the expert give a range of values for an estimate, the DM realizes that he too would be more comfortable giving a range. He likes the expert's evaluation of 1 failure in 60 years and decides to use that for an upper bound. The DM now has a beta distribution with 40th and 60th percentiles as 0.0 and 0.017, respectively. Because the beta distribution has a minimum value of 0.0, for calculation purposes it may be necessary to make the 40th percentile an extremely small value (relative to the other estimates) such as 0.0000000001 instead of exactly 0.0. In doing so the expert has a beta and the DM has a beta to be pooled into a final distribution that has the following mean:

$$x_0/n_0 = (0.021+0.677)/(0.062+33.096) = 0.698/33.158 = 0.021$$

or a mean value of 1 failure in 48 operation-years and a variance equal to

$$x_0(n_0-x_0)/[n_0^2(n_0+1)] = 0.00060 \quad .$$

In the four different parts the DM uses the expert's information in a variety of ways. The different ways produce very different final mean and variance estimates.

DM binomial and expert beta prior	0.0073	0.00008
DM noninformative and expert beta prior	0.035	0.00095
DM binomial and expert noninformative	0.0082	0.00013
DM beta and expert beta	0.021	0.00060

It is interesting to note that the lower estimates occur when the DM uses a single estimate. Higher estimates occur when the DM's information is formulated into a distribution, either noninformative or beta. In this case the DM's choice on how to handle his own data drives the final results.

■

The DM should play an important role in the inference process regardless of how many experts provide information. In the example 16.12 above, the illustrations concentrated on the answers only. The DM has access and needs to use all the ancillary information about the expert gathered at the elicitation.

1. Is the expert solving the correct problem?
2. Are the assumptions, cues, definitions and problem-solving methods used by the expert reasonable and in agreement with the DM?
3. Can the DM spot any relevant effects that this ancillary information might have on the expert's given answer?

Because the DM is also a knowledgeable party, he can answer these questions. This simple exercise will ensure that conditionality is monitored. The DM can then use what he has learned by answering these questions to make any adjustments or different ways of combining his information with that of the expert.

Decision Maker and *n* Experts

When a DM is faced with combining his information with that of several other experts (more than 1), the aggregation becomes more complicated. The DM is immediately faced with a choice.

Choice 1: He can decide to aggregate all the experts into a single distribution or estimate (accompanied by a variance or uncertainty estimate) and then combine that result with his information. This is represented by the following equation where the DM's distribution is f_0 and the combined distribution of the experts is f_c.

$$f_{\text{final}}(x) = K \cdot f_0(x) \cdot f_0(x) \quad , \qquad \text{(Morris 1977)}$$

where K is a normalizing constant.

Choice 2: He can decide to aggregate his information with the experts as if he were just another expert. This is represented by the following equation where the expert distributions are f_i:

$$f_{\text{final}}(x) = K \cdot f_0(x) \cdot f_1(x) \cdot \ldots \cdot f_n(x) \quad , \qquad \text{(Morris 1977)}$$

or

$$f_{final}(x) = K \cdot w_0 \cdot f_0(x) + \sum_{i=1}^{n} w_i \, f_i(x) \quad . \qquad \text{(Winkler 1968)}$$

In the choice 2 case, the aggregations posed in the beginning of the chapter are already applicable. In the choice 1 case, some modifications of these techniques are necessary. Specifically, the process is to first determine a composite distribution of the experts and then combine that distribution with the DM.

In terms of the DM's behavior, there may be distinctive differences in the one expert case versus n experts. It is more likely that the DM will change his own views (information) if he has information from several experts than if he just has information from one expert. The exception to this would be if the DM is extremely dogmatic in personality. Then the DM will not reduce the influence of his information or change his information in view of the other expert or experts .

The methods for applying the above combinations are modifications of the methods already mentioned in the *Aggregation Distributions* section. These methods are illustrated in the examples below. The examples include using Bayesian methods (example 16.13), assumed distributions such as the normal (example 16.14), and empirical distributions with simulation (example 16.15). Weight determinations for combining distributions use the same methods as discussed in the weight determination section above. These determinations also suffer from the same problems as mentioned there, and the resulting conclusion for using equal weights is also applicable (Seaver 1978). Example 16.15 also illustrates weight determinations and results.

EXAMPLE 16.13: ***Decision Maker and n Experts: Bayesian Aggregation***

A DM has the information elicited from the six experts from example 16.8. The information was formulated into six beta distributions, one for each expert. The parameters of these distributions are as follows (Winkler 1968):

	x_0	n_0
$f_1(x)$	0.82	242.59
$f_2(x)$	0.19	11.04
$f_3(x)$	0.18	1.52
$f_4(x)$	0.41	41.80
$f_5(x)$	0.41	8.64
$f_6(x)$	0.65	15.30

The combination distribution for the six betas is also a beta with parameters equal to x' and n' where

$$x' = \sum_{i=1}^{n} w_i x_i$$

and

$$n' = \sum_{i=1}^{n} w_i n_i \quad .$$

The DM has reviewed the ancillary information and analysis of this data and has agreed with the analyst that there are no special conditions or circumstances that make any expert's estimate different from any other. All experts are using reasonable assumptions, definitions, cues, and problem-solving processes to solve the same problem (the one at hand). There is no good reason for unequal weights. With equal weights, the values for x' and n' are

$$x' = 2.66$$

and

$$n' = 320.89 \quad .$$

The DM then has to combine his information with the other experts. He estimates the occurrence of the event as 1 in 100 trials. Because he agrees with their problem-solving methods, he simply adds his estimates into theirs to make a posterior beta with the following parameters:

$$x'' = x' + x_{DM} = 2.66 + 1 = 3.66 \quad ,$$

and

$$n'' = \text{n}' + n_{DM} = 320.89 + 100 = 420.89 \quad .$$

Therefore, the aggregation result is a *beta* distribution with a mean and variance of

$$3.66/420.89 = 0.0087$$

and

$$[3.66(420.89\text{-}3.66)]/[420.89^2(421.89)] = 0.000020 \quad .$$

■

EXAMPLE 16.14: ***Decision Maker and n Experts: Normal Aggregation***

Using the three experts from example 16.9, the DM wants to combine his estimates of a mean value of 65 and a variance of 25 into the resulting posterior normal distribution formed from the experts. The experts' distribution, $f_c(x)$, is a normal with mean, μ_{post} = 62.02, and variance, σ^2_{post} = 22.83 (Winkler 1981). Adding the DMs normal distribution as the prior results in the following *normal* posterior with mean and variance parameters

$$(\mu_{DM}/\sigma^2_{DM} + \mu_{post}/\sigma^2_{post})/(1/\sigma^2_{DM} + 1/\sigma^2_{post})$$

$$= (65/25.00 + 62.02/22.83)/(1/25 + 1/22.83) = 63.44$$

and

$$1/(1/\sigma^2_{DM} + 1/\sigma^2_{post}) = (1/25 + 1/22.83) = 11.93 \quad .$$

■

EXAMPLE 16.15: *Decision Maker and n Experts: Empirical Aggregation with Saaty-Based Weights*

A DM is given the empirical distributions functions from the two experts in example 16.10. However, he is told by the analyst that there are certain conditions that were highly significant in determining the answers given by these experts. These conditions were a set of five cues used by the experts. The cues came from a set of descriptions listed in the problem statement that the experts focused upon when solving the problem.

In order to use these conditions to formulate weights for the experts, the DM set up the cues into a hierarchical structure. He then evaluated each expert's usage of the cues using Saaty's pairwise comparison method. This method allowed the DM to formulate a set of relative weights for the two experts based on their usage of the cues.

First the DM evaluates the importance of the cues (C1-C5) relative to the problem being solved. His pairwise comparisons of the cues are indicated below.

C1 vs C2----------w
C1 vs C3----------s
C1 vs C4----------b
C1 vs C5----------b

C2 vs C3----------s
C2 vs C4----------b
C2 vs C5----------b

C3 vs C4----------b
C3 vs C5----------b

C4 vs C5----------s

The b = important (Saaty weight = 2.72, the natural log base, e); s = neutral (Saaty weight = 1.00); and w = detrimental (Saaty weight = 0.37, 1/e). The resulting relative weights for C1 through C5 are

$$(0.22, 0.33, 0.26, 0.09, 0.10) \quad .$$

Then the DM evaluates the experts given each of the five cues.

Expert 1 vs 2 given C1 ----------b
Expert 1 vs 2 given C2 ----------b

Expert 1 vs 2 given C3 ----------b
Expert 1 vs 2 given C4 ----------w
Expert 1 vs 2 given C5 ----------w

The following relative weights result for the experts given each cue.

(0.73, 0.27) for C1
(0.73, 0.27) for C2
(0.73, 0.27) for C3
(0.27, 0.73) for C4
(0.27, 0.73) for C5

To calculate the final relative weights for the two experts, each cue weight is multiplied by the expert weight for that cue; then these products are summed over all cues.

$$\text{Expert 1 weight} = 0.73(0.22) + 0.73(0.33) + 0.73(0.26) + 0.27(0.09) + 0.27(0.10) = 0.64 \quad ,$$

and

$$\text{Expert 2 weight} = 0.27(0.22) + 0.27(0.33) + 0.27(0.26) + 0.73(0.09) + 0.73(0.10) = 0.36 \quad .$$

Therefore the relative weights of the experts are 0.64 and 0.36. These are relative weights and should not be taken at their numerical values. (This is a granularity issue.) From this analysis the DM can see how the two experts rate relative to each other regarding the important cues in the problem solving. The DM can then assign numerical values based on this relative assessment. The DM decides to use weights of 0.7 and 0.3 for the two experts, indicating that he feels that the relative weights closely represent true weights for the experts.

The DM now wishes to combine the empirical distribution functions of the two experts using the following aggregation function:

$$f_c = 0.7f_1 + 0.3f_2 \quad .$$

The DM also wishes to combine his information (empirical distribution function) with the experts.

Median	0.25
30th	0.20
70th	0.30
5th	0.05
95th	0.45

The DM uses equal weights for his distribution, f_0 and f_c:

$$f = 0.5f_0 + 0.5f_c = 0.5f_0 + 0.35f_1 + 0.15f_2 \quad .$$

Using the empirical code in Appendix C, the following aggregation distribution, *f*, results.

Percentile Level	Value
1st	0.08
5th	0.12
10th	0.14
20th	0.18
30th	0.20
40th	0.22
50th	0.25
60th	0.27
70th	0.29
80th	0.32
90th	0.37
95th	0.43
99th	0.58

Mean = 0.26
Variance = 0.0097
Standard deviation = 0.098

This aggregation is not very different from the original two experts, equally weighted, in example 16.10. One reason for that is that the DM gave estimates similar to both experts. Another reason is that even though different weights were applied, the weights did not have a significant impact.

■

Analyst and *n* Experts

The reason this application environment is listed separately from the ones involving a DM stems from the anticipated uses of this book by analysts. In the application environments discussed above involving a DM, the text and examples mentioned that the analyst is the person supplying the DM with vital information and results in addition to the answers from the experts. It is also a part of the analyst's role to help the DM synthesize and assimilate this information. To do this, the analyst should make the DM aware of the concepts of granularity, conditionality, and quantification; and the analyst should guide the DM in the aggregation process.

The DM may view himself as just another expert and rely on the analyst to take the DM's information, do the analyses, and report back to the DM. There may also be cases where the DM is not an expert at all. In this case, the analyst is faced with a similar situation: do the analysis without the DM's added information and report back the findings.

In either case, the analyst is left alone with the information (data). The analyst does his job and reports his findings.

This application environment has advantages. First, the analyst is the only source of influence on the information elicited. He is familiar with the cautions and

recommendations presented in this book designed to minimize his influence on the data. Second, the analyst either elicited the data or worked closely with the elicitor. He is therefore familiar with the data and has helped with the development of the elicitation. Third, the analyst can be an objective bystander who is not concerned with the results and conclusions of the study. He can be free to *let the data speak,* and make inferences accordingly.

The analyst may find it extremely difficult in the reverse environment where the DM takes the analyzed results and changes them without the benefit of the analyst or the guidelines presented in this book. Careful analysis of the information can be quickly destroyed in such a case.

The focus of part III is from the analyst's viewpoint. Most of the examples given in this book reflect the analyst and *n* experts environment. However, it is also recommended that the analyst be an integral part of the elicitation. Facilitators or moderators of the elicitation and the analyzers of the data must work together. If they do not, then definitions, assumptions, problem solving, and conditions can change during the study. Also, granularity can change without notice, making interpretations meaningless. Common sense and simple consistency in designing and implementing the study from start to finish are the key to success.

Aggregation and Uncertainty Analysis

Uncertainty analysis is addressed in more detail in chapter 17. However, many of the problems associated with aggregation are related to uncertainty characterizations. In fact, many of the techniques are useful for both, such as simulation methods, Bayesian methods, tree diagrams for decomposition, and the use of distributions rather than single estimates.

As is illustrated in chapter 17, methods used to characterize uncertainties automatically aggregate estimates and distributions in the same way that was illustrated in the examples in this chapter. The difference is one of interpretation. In this chapter, the ranges and distributions provided by the experts were not specifically labeled as characterizing uncertainties; although they did represent uncertainty. The primary goal was to aggregate all the information into a convenient estimate plus variance or into a distribution. The interpretation of those final estimates and distributions do incorporate and reflect the uncertainties in the information given by the experts and also the uncertainties among the experts. The goal and emphasis in chapter 17 is to identify and characterize the uncertainty.

17

Characterizing Uncertainties

There are ways of characterizing and handling all uncertainties even in the restrictive environment of expert judgment applications. This chapter examines some of the easier ways. The concept and definition of uncertainty given here is, in a very broad sense, covering the four basic sources of uncertainty that prevail throughout the data and the analysis.

To control, or at least understand, this single uncertainty characterization, the following steps are suggested in this chapter. First, uncertainty measures for the answer data that are needed are discussed in *Obtaining Uncertainty Measures*. Second, uncertainties in the data that can be modeled either separately or as additional terms in the full data analysis models are discussed in *Modeling Uncertainties*. This chapter concludes in *Comparison of the Methods* with a comparison of various methods for handling uncertainties. Later in chapter 18 on making inferences, the relationship that uncertainty has to the inference process is discussed.

Living with Uncertainties

There are different kinds of uncertainties that become important in any sampling or experimental data-gathering process. Some uncertainties can be controlled and, therefore, reduced to an acceptable noise level of influence simply by taking larger samples or by doing careful experimental design and measurements. Other uncertainties cannot be controlled or reduced by any practical means. These are the uncertainties with which we all must live. The most that can be done is to understand their importance and effects.

Uncertainties can stem from different sources such as (1) definitions, (2) sampling errors, (3) nonsampling errors such as missing data, and (4) scientific or modeling techniques (Stoto 1988). In expert judgment applications, uncertainties come from all four sources. Uncertainty from definitions can be reduced by careful elicitation as proposed in Part II of this book. Sampling error uncertainty can be reduced by taking large sample sizes; however in expert judgment applications, this may not always be practical. Nonsampling uncertainty cannot be reduced by any simple or practical means. Modeling

uncertainty can be reduced by proper experimental design (where possible) and even by the use of cross validation in the analyses.

Obtaining Uncertainty Measures

Uncertainty values for the experts' answers can be obtained directly from the experts during the elicitation, or they can be estimated indirectly from the post-elicitation data. In either case, the expert or the analyst may be required to make additional assumptions or to provide estimates that stretch the limits of their current knowledge. However, the purposes of estimating uncertainties are (1) to represent the possible inaccurate estimation of the variables of interest by the experts, and (2) to increase the chance of estimating (covering) the true answer by allowing for a range of possible values rather than relying on a single value.

Using Elicitation

Asking experts to estimate uncertainty measures for an answer is, on the one hand, like asking them to estimate a variance or distribution of values. As noted in chapter 7, these estimations are difficult and may be highly inaccurate for most experts. On the other hand, asking for a simple range of possible values from an expert is not much different from asking for the original answer. Eliciting a range of values requires the same care as eliciting a single answer. One advantage of eliciting a range of values is that many experts are comfortable with providing uncertainties in this form, realizing their existence and importance. In fact, many experts prefer to give a range of possible values, being uncomfortable with the pinpoint accuracy implied for a single value estimate.

In chapter 7, several dispersion measures were offered for selection. These included error bars, variances, percentiles, and ranges. All of these can be used to characterize uncertainties in the answers and can be elicited from the experts along with the answers. The advantages and disadvantages for each are given below.

Error bars

Most engineers use the term error bars to connote some measure of uncertainty; however, there is little agreement on a firm definition of how much uncertainty is characterized by error bars. Therefore, error bars should be elicited using some sort of definition. Most such definitions will overlap with the other uncertainty measures. For example, error bars could be defined as plus or minus one standard deviation from the best estimate (a standard deviation definition), or error bars could be defined as the middle 90% (percentiles definition) of the distribution.

The philosophy of this handbook is to keep restrictions on the expert to a minimum. In keeping with that, the expert would be asked to provide error bar values and then provide his definition of what those values represent. This way is the recommended use for error bars.

Variances or standard deviations

Most engineers have heard and used the terms variance or standard deviation but may not have a good understanding of them. It is also not recommended that the experts be asked if they are comfortable with these concepts. In general most people do not like to admit a lack of knowledge or understanding of any concept. The experts can be trained in these concepts during the elicitation. However, studies have indicated (Martz, Bryson, Waller 1985) that even experts with expertise on these concepts are not very good at estimating variances. In general variances are underestimated in value, sometimes by a factor of two or more. Therefore, using variances and standard deviations will result in large underestimates of uncertainty.

Because of the unfamiliarity of these concepts by many experts, it is not recommended that they be used as uncertainty measures.

Percentiles

Percentile estimation involves the concept of a probability distribution of values. The 95th percentile is the value such that 5% of the distribution is larger than that value and 95% is smaller . As with variances, it has been demonstrated that even statisticians have difficulty in estimating percentiles. People will also tend to underestimate the uncertainty in the form of percentiles. When asked to estimate 95th percentile values, people only estimate about the 60-70th percentile values, and 5th percentile values are really only about the 30-40th percentile values. There is another problem inherent in this process. Even if the expert is comfortable with the concept of a distribution, he will tend to think in terms of a symmetric, bell-shaped distribution. Such an assumed distribution may be totally inappropriate for the problem. The result is a distortion of the values and percentiles that the expert is trying to estimate.

Again, because of the difficulty in defining and using the concept of percentiles, they are not recommended for use as uncertainty measures.

Ranges

As mentioned above, many experts will prefer to give a range of possible values instead of a single point estimate. This preference reflects their uncertainty in providing a single value. Many experts will give a range of values whether a range is elicited or not. There is a problem with interpreting a given range. Usually experts are unwilling (or unable) to provide a definition of what the range represents. These definitions might also involve other uncertainty characterizations such as variance that should be avoided.

It is recommended that definitions of ranges not be provided in terms of the other uncertainty measures. It is also recommended that some rationale be gathered concerning what the expert has in mind about the ranges: Are they equally possible values? Do they represent extreme values? It is difficult for the analyst to use ranges that have no meaning attached; however, as demonstrated in the next section, undefined ranges can be a source of information for analysis.

Using Post-Elicited Data

This section provides some suggestions on obtaining uncertainty measures from the data that are not elicited from the experts and on using such information. Uncertainty characterizations can come from several sources: (1) directly from the experts, (2) indirectly from the experts, (3) assumed by the analyst from the data, or (4) any combination of these three. In any of these four sources, assumptions are being made by the experts, the analyst, or both. Therefore, following the suggestions given in this chapter should be viewed with extreme caution because of these assumptions. Different interpretations of the uncertainties should be tried and compared. In *Modeling Uncertainties* it it is indicated how these uncertainty measures can be analyzed and then a final comparison section is given.

1. As mentioned above, many experts will volunteer ranges of values when answering a question. It is important to query them about the meaning or interpretation that they attach to these ranges. Otherwise the analyst is forced to assume some interpretation such as the 40th percentile and 60th percentile values. Volunteered range values can be used in uncertainty analyses calculations as described below. They can be used as repeated measures for determining variations or parameters for assumed distributions. They can also be used as repeated measures for supplementing a sparse or small data set. For example, if five experts give five best guesses and 5 minimum values with 5 maximum values, then there are 15 values for the data set.
2. Experts may also supply ranges indirectly. This can be done in several ways.

   First, the expert may give his best estimate and then, later on in the session, revise that estimate or give another possible estimate. If the rationale is recorded, many such references to estimate changes in values and assumptions (if changed) can be noted and recorded. If an expert changes his estimate, it is vital to find out why. In some cases, he may be updating his thought processes; in other cases, he may be changing the estimate because the problem has changed. If he is changing the problem, then the new estimate is not useful for a range value.

   Second, the expert may have recorded his response in such a way as to indicate a range of values. For example, if a continuous number scale is the response mode, an expert may use a wide mark or smear his response along the line indicating a spread of values. This may only indicate a narrow range, but it is a useful range nonetheless. This kind of volunteered range value is best considered as a very narrow uncertainty measure and can be used as repeated measurements to increase the sample size.
3. The analyst can make many assumptions about either the direct or indirect ranges volunteered by the expert. Percentile values, fractions of standard deviations, or multiples of standard deviations definitions can be assumed by the analyst to apply. A common rule of thumb (Kahneman, Slovic, Tversky 1982) is to take the experts' uncertainty range and double or quadruple it to make a 90% coverage interval (the difference between the 95th and 5th percentiles). In determining the validity of such a rule, the effects of different uncertainty measures can be compared or can be studied by using simulation.

Another way that analysts can use uncertainty measures is by using the direct or indirect ranges as additional estimates to increase the sample size. This method is especially useful for applications which have only a few available experts. Again, a simulation technique such as the bootstrap is useful for using the range values to increase sample size.

The analyst can also make assumptions even when no ranges are directly supplied by the experts. Such indirect assumptions can be made by examining the variation in the given answers. One such method would be to induce a range of values by doubling the variance of the original set of single estimates and assume a distributional form for the estimates with this doubled variance and original mean. A similar uncertainty measure could be determined by doubling the original range (maximum value - minimum value) of the data. To assist the analyst in making such determinations, the experts may have provided some verbal clues about the uncertainties in their estimates. The analyst will be required to transform any such qualitative statements into quantitative values of uncertainties. Therefore, any and all such information provided by the experts should be viewed as part of the uncertainty characterization.

Modeling Uncertainties

After determining some uncertainty measures for the experts' estimates by the above methods, modeling or using uncertainties can be done by many different methods. A few more commonly used techniques are described below, and a comparison of these techniques is given at the end of the section for the problem described in the examples.

Bayesian Methods

Using one prior

In many risk and reliability problems, expert judgment is used to supplement existing (but usually small amounts of) data. The expert judgment is used to formulate a prior distribution that is combined with the data and its distribution using Bayes Theorem to form a posterior distribution that reflects a composite of all the available information. In this type of application, the expert judgment information serves two purposes (1) it augments the data, and (2) it serves to characterize the uncertainty in the data.

The major disadvantage to this approach is that probability distribution forms are needed for both the expert judgment data (as the prior) and for the existing data. Then the process of combining these distributions can be difficult from a mathematical viewpoint. In some difficult cases, numerical analysis techniques are required to find a solution. In other difficult cases, simulation may be used to find a solution. To avoid mathematical difficulties in calculating a posterior result, many analysts assume commonly used distributions that have convenient mathematical properties that provide easy forms for the posterior distribution. One such commonly used combination of the prior information with the data is the binomial-beta combination. Here the existing data are assumed to follow a binomial process--in n independent trials, x failures are observed, with each trial having a

probability of failure equal to p. The prior information (usually from the experts) is also in the form of x_0 failures in n_0 trials. However, here the prior distribution is a beta distribution. This beta distribution is the distributional form for the binomial parameter p. The beta is convenient for three reasons: (1) the beta distribution can have many different shapes, (2) the beta ranges from 0 to 1 just as p does, and (3) the beta parameters are interpreted as x_0 failures in n_0 trials. The resulting posterior of this binomial-beta combination is also a beta distribution. Its mean and variance are easily calculated functions of the n, x, n_0, and x_0 values as follows:

$$posterior\ mean = \frac{x + x_0}{n + n_0} ,$$

$$posterior\ variance = \frac{(x + x_0)(n + n_0 - x - x_0)}{(n + n_0)^2(n + n_0 + 1)} .$$

Two other prior distributional forms combine easily with the binomial process and are commonly used in situations where little or no prior information is available. These are the uniform distribution on the (0,1) interval and the noninformative prior $k(p\text{-}p^2)^{-1/2}$, where k is any constant. By choosing k in terms of the gamma functions (Martz and Waller 1982), this noninformative prior becomes a beta distribution with $x_0 = 0.5$ and $n_0 = 1.0$. The posterior distribution is then also a beta distribution function with the above mean and variance formulae. The uniform distribution is a special case of the beta distribution with parameters $x_0 = 1.0$ and $n_0 = 2.0$. Again the resulting posterior is a beta distribution with mean and variance formulae given above. Thus for either the uniform or noninformative priors the above formulae can be used to determine the posterior mean and variance. One word of caution is necessary here in using either of these priors. The uniform prior is used when no prior information is available and is sometimes called the ignorance prior. This prior spreads the information evenly across the entire range of values, from 0.0 to 1.0. It has almost a negligible impact on the posterior. On the other hand, the noninformative prior is really a misnomer. This prior is informative and does have an impact on the posterior. Its name should be the prior of little information.

Choosing the distributional form and corresponding parameters for the prior distribution can make a significant difference in the posterior. Using the binomial-beta combination to form the posterior, example 17.1 shows the influence that the prior parameters can have on the final results.

EXAMPLE 17.1: ***Using Bayesian Methods for Uncertainty--Forming a Single Prior***

Ten experts have provided estimates for the probability of failure per year of a subsystem in a reactor as follows:

Expert	Estimate of Failure Rate
1	0.00250
2	0.00100

3---------- 0.05000
4---------- 0.00500
5---------- 0.01000
6---------- 0.02500
7---------- 0.00100
8---------- 0.00250
9---------- 0.00010
10---------- 0.00005

It is also known that this subsystem has been operational for 12 years without any failures. This data follows a binomial process with $x = 0$ failures in $n = 12$ years. The 10 experts' estimates can be used to form a prior distribution that is combined with this binomial data. The major advantage here is that the data alone do not contain enough information to form a failure-rate estimate. Just using the data would give a failure-rate value of 0/12 = 0. By combining the data with the experts information, a more reasonable estimate is possible.

The variation in the 10 experts' estimates represents the uncertainty in the value of the failure rate. By forming a distribution out of these 10 values, that distribution will represent the uncertainty. The mean of the 10 values is 0.010; the standard deviation is 0.016. A beta distribution with that mean and standard deviation has parameters $x_0 = 0.40$ and $n_0 = 39.5$. These parameters represent an average failure rate of $x_0/n_0 = 0.010$ or 1 failure in 98 years. A beta distribution with the above parameters, x_0 and n_0, has the following characteristics:

Mean	0.0102
Variance	0.00025
Minimum	0.00000023
5th percentile	0.000012
50th percentile	0.0038
95th percentile	0.042
Maximum	0.074

Combining the expert information (beta prior distribution) with the data (binomial process) gives the mean and variance of the posterior beta distribution as

$$\frac{1+0}{98+12} = 0.0091$$

and

$$\frac{1(110\text{-}1)}{110^2(110+1)} = 0.000081 \quad .$$

The influence of the data and the prior information on the final estimate is demonstrated by examining changes in the data. Suppose there was 1 failure in 12 years,

then the final mean estimate becomes 0.018, or twice the original value, with a variance of 0.00016. Here the data is more dominant, increasing the mean estimate. Suppose the data was 0 failures in 144 years. Then the final estimates of the mean and variance are 0.0041 and 0.000017, respectively. Here the final mean is dominated by the experts.

Of course, changes in the experts' estimates affect the posterior in a similar fashion. The purpose of this exercise is to emphasize that the prior and the data both are influential and care must be taken to represent each appropriately.

■

Assuming distribution forms and determining posteriors can be arbitrary and difficult. It is therefore recommended that Bayesian methods be used only in the simple binomial-beta cases such as in example 17.1. For other cases, a statistician should be consulted.

Using multiple priors

To characterize the uncertainty in each estimate (answer) given by the experts, Bayesian methods can be used to establish a prior distribution for each expert. Each expert provides an estimate or best guess and a corresponding range of values for that estimate. These ranges represent the uncertainties that the experts have about the accuracy of their single-point estimates. Distributions representing the estimates and their uncertainties can be formed for each expert using the ranges and the estimates. Example 17.2 discusses some of the ways of forming these distributions and shows one method in detail.

EXAMPLE 17.2: ***Using Bayesian Methods for Uncertainty -- Forming Multiple Priors***

The 10 experts in example 17.1 provided uncertainty ranges with their estimates as follows:

Expert	Estimate of Failure	Rate Ranges
1	0.00250	0.00100 - 0.0040
2	0.00100	0.00010 - 0.0100
3	0.05000	0.00100 - 0.1000
4	0.00500	0.00100 - 0.0100
5	0.01000	0.00500 - 0.0500
6	0.02500	0.01000 - 0.0500
7	0.00100	0.00500 - 0.0025
8	0.00250	0.00100 - 0.0050
9	0.00010	0.00010 - 0.0100
10	0.00005	0.00005 - 0.0005

Given this information, there are several ways of formulating distributions for each expert. The type of distribution is the first choice to be made. For convenience, the beta distribution is chosen. Following are some ways of forming the beta distributions:

1. The single estimates represent the mean of the beta. The lower range values represent specified percentiles. For expert 1, his resulting beta distribution would have parameters $x_0 = 1.90$ and $n_0 = 761.9$ if his lower range was the 20th percentile. This beta distribution has a mean of 0.0025 and a variance of 0.0000033.
2. The single estimates represent the mean of the beta. The upper range values represent specified percentiles. For expert 1, his resulting beta distribution would have parameters $x_0 = 1.08$ and $n_0 = 430.8$ if his upper range was the 80th percentile. This beta distribution has a mean of 0.0025 and a variance of 0.0000058.
3. The single estimates represent the median of the beta. The lower range values represent specified percentiles. For expert 1, his resulting beta distribution would have parameters $x_0 = 0.64$ and $n_0 = 142.5$ if his lower range was the 30th percentile. This beta distribution has a mean of 0.0045 and a variance of 0.000031.
4. The single estimates represent the median of the beta. The upper range values represent specified percentiles. For expert 1, his resulting beta distribution would have parameters $x_0 = 1.32$ and $n_0 = 401.8$ if his upper range was the 70th percentile. This beta distribution has a mean of 0.0033 and a variance of 0.0000081.
5. The range values represent specified percentiles. The single estimate is not used. For expert 1, his resulting beta distribution would have parameters $x_0 = 0.31$ and $n_0 = 38.6$ if his lower range was the 40th percentile and his upper range was the 60th percentile. This beta distribution has a mean of 0.0080 and a variance of 0.00020.
6. The range values represent specified percentiles such as the 40th and 60th percentiles. Beta distributions are formed from these. These betas act as priors to be combined with the information in the single estimate where the single estimates represent 1 in p failures such that $x = 1$ and $n = 1/p$. For expert 1, the range values form a beta prior with parameters $x_0 = 0.31$ and $n_0 = 38.64$. The single estimate represents a binomial process with $x = 1$ and $n = 400$. Therefore, a resulting beta for expert 1 has parameters $x_0 + x = 1.31$ and $n_0 + n = 438.64$. The resulting beta has a mean of 0.0030 and a variance of 0.00000014.

As seen in the example of using expert 1, the beta parameters and the variances can change quite significantly depending on the interpretation of the uncertainty range values. However, the means remain fairly similar to the original single estimated value (except in item 5, where the mean is not used).

Using the method in item 5., the lower ranges are the 40th percentiles and the upper ranges are the 60th percentiles. With these percentiles, the following are the beta parameters for each expert:

Expert	Beta Distribution Parameters x_0	n_0
1	0.31	38.64
2	0.09	0.60
3	0.09	0.30
4	0.18	4.32
5	0.18	1.25
6	0.26	2.60
7	0.26	42.76
8	0.26	21.64
9	0.09	0.60
10	0.18	76.19

These 10 beta distributions can be combined according to the distribution aggregation methods from chapter 16, such as using Monte Carlo simulation. The resulting combination distribution could then be used as a prior to combine with the data or with a decision maker's distribution to form a posterior. The aggregation by simulation is performed in example 17.3 in the next section on Monte Carlo simulation.

■

Simulation Methods

In chapter 11, the bootstrap and Monte Carlo simulation techniques were introduced. These techniques were also used in chapter 14 for exploring correlation among experts, and in chapter 16 for aggregation of expert estimates. Both techniques are useful for characterizing uncertainties.

Monte Carlo simulation

One of the easiest and most effective (Martz et al. 1983) ways of propagating uncertainties through a model is to use Monte Carlo simulation. Simulation can be used to combine various types of distribution functions without relying on difficult or complex mathematical formulations of the combinations. The mathematical difficulties in aggregating expert estimates was seen in chapter 16. Similar problems can also arise in characterizing uncertainties because uncertainties are commonly represented by distributions on estimates. These uncertainty distributions must usually be combined in some manner with original estimates or with other uncertainty distributions to obtain an overall effect of all the uncertainties.

Example 17.3 illustrates how uncertainty characterizations for the 10 experts' estimates from example 17.2 can be combined to form a single distribution that represents the uncertainties in all 10 distributions. In example 17.2, each expert's range was used to formulate a beta distribution as the uncertainty distribution for his estimate. Using the ranges and the original estimates, several ways of formulating such beta distributions were given in example 17.2. In example 17.3, the effect of choosing different distributional forms for the uncertainty distribution is illustrated. First, the ranges are used to establish

the upper and lower limits of a uniform distribution. Second, the ranges are used to determine the parameters of a beta distribution. Then, in both cases, the 10 distributions are aggregated to determine the distribution of the median (50th percentile value) of the 10 experts' estimates using Monte Carlo simulation. The simulation is done by forming 1000 different samples by randomly choosing one value from each of the 10 expert distributions. For each sample the median of the 10 values is calculated. The end result is a distribution of 1000 medians. The variance, percentiles, and mean of this final distribution provides the estimates of the variance, percentiles, and mean for the median of the 10 experts.

EXAMPLE: 17.3: ***Uncertainty Characterization Using Monte Carlo Simulation***

Using the 10 experts' ranges given in example 17.2, individual uncertainty distributions can be determined for each expert. These distributions can then be combined to form a distribution for an overall estimate of the 10 experts. According to chapter 16 on aggregating expert judgment, the median of the experts is a commonly used aggregation estimator. Monte Carlo simulation allows the analyst to determine the distribution of the median.

In this example, the ranges given by the experts are used to form two different distributions: (1) the lower and upper range values form the 40th and 60th percentiles of uniform distributions, and (2) the lower and upper range values form the 40th and 60th percentile values for beta distributions as done in example 17.2.

Final distribution for the median of the 10 experts is formed from 1000 median values calculated from 1000 samples of size 10. Each sample is formed by randomly selecting a value from each expert's distribution.

Two simulations are done. The first uses uniform distributions to represent the experts' uncertainties in their estimates, and the second uses beta distributions. The two resulting median distributions have the following characteristics:

Uniform Uncertainty Distributions for the Experts

Mean	0.0047
Variance	0.0000072
Minimum	0.0000013
5th percentile	0.00030
50th percentile	0.0038
95th percentile	0.012
Maximum	0.026

Beta Uncertainty Distributions for the Experts

Mean	0.0091
Variance	0.00024

Minimum	0.00000033
5th percentile	0.00038
50th percentile	0.0041
95th percentile	0.034
Maximum	0.14

The variance for the medians from the beta uncertainty distributions is much larger than from the uniforms. The maximum for the beta is a little larger, and the minimum for the beta is much larger. However, the medians of both are the same, and the means are also quite close. Thus the central measures of the final distributions for the median are not affected by the choice (beta or uniform) of the uncertainty distribution form used for the experts. The greatest effect from distributional choice is in the variances and the shapes of the tails of the resulting distributions.

■

In examples 17.2 and 17.3, two problems were identified in translating the experts' range values into uncertainty distributions. The first problem (example 17.3) is deciding upon a form for the distributions. The second problem (example 17.2) is deciding how to use the ranges to form the parameters of the chosen distributions.

In example 17.3, the effect of the first problem was seen. The choice of the uncertainty distributions (beta or uniform) made a difference in the variances and tails of the final distributions for the median. One way of handling this problem is to run several simulations, each using a different, but reasonable, distribution choice. Results from different choices should be consistent with each other, or inconsistencies should be resolved based on the assumptions that were made. For example, in 17.3 the wider variance in the beta case is a result of the interpretation made about the experts' range values (the second problem). If the ranges had been used to represent the 5th and 95th percentiles, then the resulting variance from the beta case would have corresponded better to the uniform case.

There are some logical choices for distributions that can be tried and compared for various types of estimates. For probability estimates, the uniform and beta distributions are logical choices. The normal distribution is also commonly used; however, care must taken not use a normal distribution that gives values for probabilities which are negative or greater than 1.0. The normal, lognormal, and gamma distributions are often used to represent estimates of physical quantities, such a temperatures or of failure rates.

Another way of handling the first problem is to do the simulation using a technique such as the bootstrap that does not require a distributional assumption. This technique uses the data itself to form an empirical distribution for the simulation. Details on this technique are given below.

In handling the second problem, understanding that experts underestimate uncertainty is useful. Because experts do underestimate uncertainty, it is recommended that either the range be doubled for use as tail percentile values (e.g., 5th and 95th) or that the range values represent inner percentiles, such as the 30th to 40th and 60th to 70th percentiles. As in example 17.2, different interpretations should be tried. Usually the different interpretations will tend to affect the variances of the uncertainty distributions rather than the centers of the distributions. The basic idea is to represent the uncertainties in

the quantity being estimated. If the experts feel that the uncertainty is great, then larger variances are to be expected, and these variances may even be larger than the experts expect.

Bootstrap simulation

Distributions can be formed by sampling and resampling from the original data set. This sampling and resampling process is referred to as sampling with replacement. To form a single sample of size n, n values are randomly chosen from the original data set. If a particular value is chosen once, it can be chosen again in the same sample. The simulation is done by forming N such samples (N = 1000). As in the Monte Carlo simulation, a calculation or model is formed for each sample. The N results of this calculation or model are collected to form a final distribution.

The main advantage with bootstrap sampling is that no distributional assumptions are required on the data. Its main disadvantage is that the sampling/resampling procedure produces a final distribution with a small variance. In other words, the variation of the final distribution is limited to the variation of the original data set.

One way to overcome the restricted variance problem is to induce more variation (more uncertainty) into the original data set. The ranges of values that represent the uncertainties in the values serve to expand the variation of the original data. Therefore, if the ranges are included in the original data set, the results from a bootstrap simulation will have a wider variation than the original sample without ranges. Example 17.4 illustrates the difference in the bootstrap final results when calculating the median of the 10 experts' estimates for cases *without* the range values and *with* the range values

---

EXAMPLE 17.4: ***Uncertainty Characterization Using the Bootstrap***

Using the data from the 10 experts from example 17.1, a bootstrap sampling procedure of those 10 estimates represents the uncertainty in the estimates. Samples are formed in a similar manner to the Monte Carlo simulation except random selections are taken from the original data set and not some specified distribution. For each sample, a single datum can be chosen either once, more than once, or not at all.

By forming 1000 such random samples from the original 10 estimates and calculating the median value of each sample, a distribution of 1000 medians is formed with the following characteristics:

Mean	0.0044
Variance	0.0000085
Minimum	0.0010
5th percentile	0.0018
50th percentile	0.0037
95th percentile	0.010
Maximum	0.027

The central values (mean and median) are the same as the ones in example 17.3. However, here the variance is between the two results in 17.3. It was noted in that

example that the smaller variance probably reflected the tendency of the experts to underestimate uncertainty. It is also known that the bootstrap simulation produces a small variance--a variance limited by the variation in the original data set. Therefore, the resulting variance of the median may also be too small to adequately represent the uncertainty .

One solution to this underestimation of uncertainty is to expand the variance of the original data set by including the range values given by the experts. One way of doing this is to supplement the data set by adding these upper and lower range values as if they were additional estimates. Now the original data set is increased to 30 values. However, when performing the bootstrap simulation, the sample size of 10 is recommended so that false benefits from an increased sample size are not induced into the simulation.

Results for the bootstrapped median with range values follow:

Mean	0.00083
Variance	0.00000089
Minimum	0.0000010
5th percentile	0.0000010
50th percentile	0.00055
95th percentile	0.0030
Maximum	0.0063

The final distribution for the median is stretched over a wider range of values than the final distribution without the range values. However, because the range values are biased toward the lower values, this final distribution is shifted in that direction. This shift may not be a desirable result. One way to avoid this shift would be to add range values that are symmetric about the original estimates.

■

In conclusion, both simulation techniques (Monte Carlo and bootstrap) have their advantages and disadvantages. Care must be taken to decide which advantage is most desirable and which disadvantage is most harmful. In the case of uncertainty characterization, the Monte Carlo distribution assumption disadvantage is less harmful than the bootstrap restricted variance disadvantage because uncertainties tend to be underestimated and more diversity is generally needed to adequately represent the true uncertainty. This is a conservative approach to uncertainty. However, unless more information about the size and effect of uncertainties is known, the conservative approach is the approach accepted by the risk and reliability community.

Decision Analytic Methods

The decision analysis community has adopted and developed many ways of characterizing uncertainties including the Bayesian and simulation methods mentioned above (Booker and Bryson 1985). Many of these techniques rely on strict model formulations. Some also require that the data be distributed with a multivariate normal distribution. Uncertainties are also assumed to follow distributions that combine with the data in a mathematically convenient fashion. It is not in keeping with the basic philosophy

in this book to impose such restrictions on the data or to require experts to give their estimates in forms that are merely convenient for the analyst. Whereas some distribution assumptions may be necessary, in this book we advocate their usage be kept to a minimum or that they be used in cross-validation with other techniques.

There is one decision analytic technique whose usage is in keeping with the philosophy expressed in this book. It is known as the maximum entropy technique and can be easily implemented using the following description and example:

The idea behind maximum entropy is to formulate a distribution for the data such that the distribution maximizes the uncertainty in the data. To determine this distribution, several values are required, and a choice for the prior distribution on the variable of interest (e.g., probability of an event, p) is also required. At least two percentile values are needed, and the value for the absolute maximum of p and minimum of p is needed. Two commonly and easily used prior distributions are the uniform and the noninformative prior (Cook and Unwin 1986).

Using a uniform prior on p from the absolute minimum value a to the absolute maximum value b and two percentiles x_l and x_u, the maximum entropy distribution for p is

$$p(x) = \begin{cases} L(x_l - a) & a \le x < x_l \\ (U - L)/(x_u - x_l) & x_l \le x < x_u \\ (1 - U)/(b - x_u) & x_u < x \le b \end{cases},$$

where U and L are the percentage values (e.g. 0.95 and 0.05) for the x_l and x_u percentiles.

Using a noninformative prior on p with two percentiles a and b, the maximum entropy distribution for p is

$$p(\log(x)) = \begin{cases} L/\log(x_l/a) & \log(a) \le \log(x) < \log(x_l) \\ (U - L)/\log(x_u/x_l) & \log(x_l) \le \log(x) < \log(x_u) \\ (1 - U)/\log(b/x_u) & \log(x_u) < \log(x) \le \log(b) \end{cases},$$

where log is the base 10 logarithm.

These two distributions will be shaped as three blocks or steps. If other percentiles are easily estimated, the above formulae can be expanded and the distribution will have more steps:

$$p(x) = \begin{cases} L_1/(x_1 - a) & a \le x < x_1 \\ (L_2 - L_1)/(x_2 - x_1) & x_1 \le x < x_2 \\ (L_3 - L_2)/(x_3 - x_2) & x_2 < x \le x_3 \\ \quad\vdots & \\ (1 - L_n)/(b - x_n) & x_n < x \le b \end{cases},$$

for n percentile estimates $(x_1, x_2, ..., x_n)$.

Example 17.5 illustrates the use of these formulas for the 10 experts' estimates from example 17.1. Prior distributions other than the uniform and noninformative can be

used; however, forming the maximum entropy distribution for those can be mathematically difficult. For such complex cases, simulation can be used to find a solution.

EXAMPLE 17.5: ***Forming a Maximum Entropy Distribution***

For the 10 experts in the previous examples, the smallest range value is 0.00005, and the largest is 0.1. These numbers for the 5th and 95th percentile values in the maximum entropy formulae form the uniform prior distribution. The values for the absolute minimum and maximum are also needed. These values are listed below with the percentiles:

$$\begin{aligned} \text{Absolute minimum value, } a &= 0.000001 \\ \text{Absolute maximum value, } b &= 0.20 \\ \text{Number of specified percentiles} &= 2 \\ \text{Percentile levels, } U &= 0.95 \text{ and } L = 0.05 \\ \text{Percentile values, } x_u &= 0.10 \text{ and } x_l = 0.00005 \end{aligned}$$

The following is the distribution of the failure rate, $p(x)$.

$$\begin{aligned} p(x) = \; & 1020.41 && 0.000001 \leq x < 0.00005 \\ & 9.00 && 0.00005 \leq x \leq 0.10 \\ & 0.50 && 0.10 < x \leq 0.20 \end{aligned}$$

This distribution has the following characteristics:

Mean	0.00056
Variance	0.000041
Minimum	0.0000010
5th percentile	0.000050
50th percentile	0.025
95th percentile	0.10
Maximum	0.20

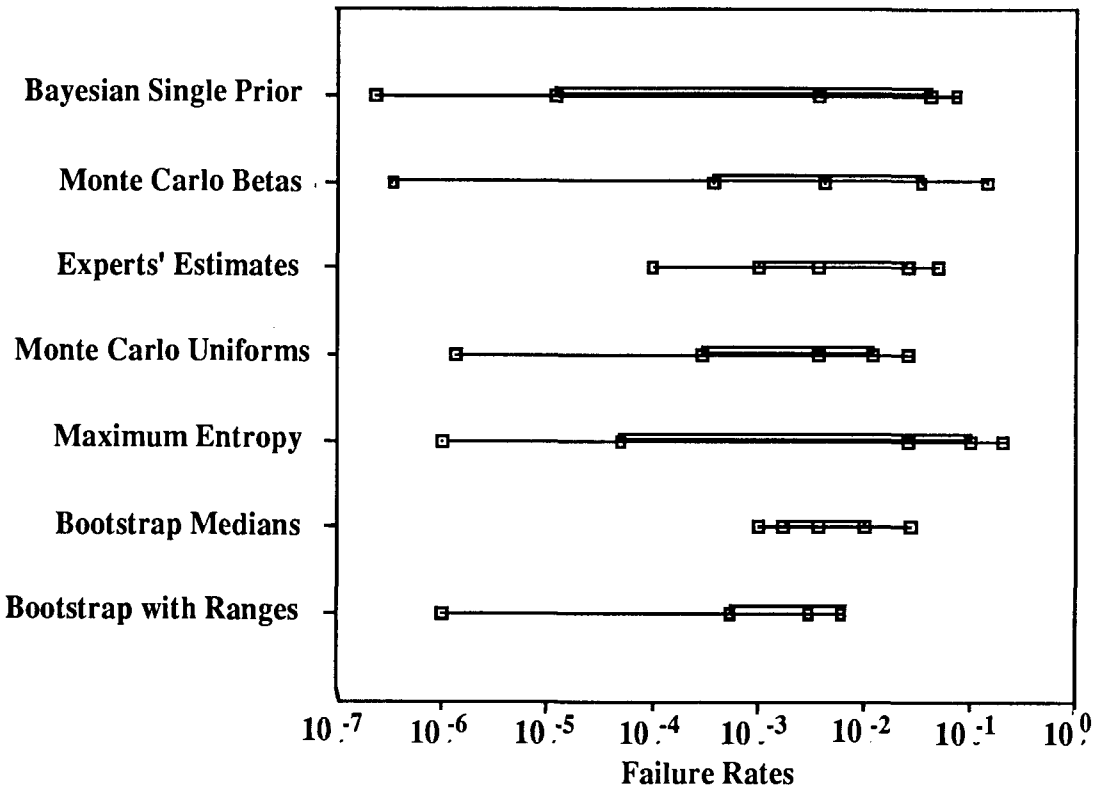
■

Comparison of the Methods

The methods for characterizing uncertainty presented in this chapter are chosen for their ease of implementation and for the minimal amount of assumptions required. Example 17.6 shows how these methods for characterizing uncertainty compare using the 10 expert estimates and their range values.

Example 17.6: ***Comparison of uncertainty characterizations***

On this graph the maximum and minimum values for the final distribution are plotted at the ends of the straight line. The narrow boxes connect the 5th and 95th percentile values for the median. The 50th percentile is the point inside the box. The final distributions for the bootstrap and Monte Carlo plots are the distributions for the median. The final distributions for the experts' estimates, Bayesian single prior, and maximum entropy plots are the distributions for the failure rate (not the median of the failure rates).



■

Some interesting results are evident from this graph. First, the bootstrap without ranges is very narrow, indicating little variation in the values. It is even narrower than the raw data and is not unexpected. Small variations are indicative of the bootstrap (Efron 1979), and the variation of the median is expected to be smaller than the variation in the raw data. The bootstrap for the median with ranges has a wider spread than the bootstrap from just the raw data because there is increased variability of the data set from inclusion of the range values. Second, most plots indicate a skew of the values with more of the distributions shifted to the right or to higher failure rates. This is not surprising because the raw data and range values are shifted to the higher failure rate values. Third, the Monte

Carlo with beta priors, the maximum entropy, and the single beta prior indicate the most spread and are very similar to each other. The maximum entropy and single prior also have similar variability in the center (the boxes). The Monte Carlo betas has narrower center because it represents the median and not the raw estimates. Finally, there is increased variability over the raw estimates indicated in the plots of the methods using the ranges to represent the uncertainty.

The differences in the methods result from the different assumptions made about the distributional forms (beta, uniform, and empirical) and the interpretations of the range values (where used). Complete consistency should not be expected. However, inconsistencies should be explained. In this comparison, the plots for all methods are quite similar with the exception of the raw data and the bootstrap median of the raw data. It is expected that these plots should show very little variation because the range values were not used to represent uncertainties in the raw estimates. As in other chapters, it is suggested that several different uncertainty characterizations be tried and compared. However, if only one method is chosen, the Monte Carlo with uniform priors would be a reasonable choice (Martz et al. 1983).

18

Making Inferences

The purpose of eliciting and analyzing expert judgment has always been to use the information gained either as data where none existed or as supplemental data where sparse data existed. The goal is then to take this information from the experts and draw conclusions from it. This process is referred to as making inferences.

In this chapter the possible inferences that can be made are examined taking into account the design features of the problem and the use of analysis methods. Finally, some comments about inference relating to expert judgment applications are presented.

What Inferences Can Be Made

Besides inference referring to information gained either as data where none existed or as supplemental data where sparse data existed, inference also refers to drawing conclusions that apply on a more universal scale. These inferences are based on statistical principles of sampling. For example, results from a sample that is representative of a larger population are used to make statements regarding that population. In expert judgment applications, however, such extended inference is usually not possible. For one reason, the information from the experts is not a random sample (or representative sample) of the true state of knowledge. In most cases it is not even possible to form a random sample of the experts used for the elicitation. Therefore, making inferences from expert information about the true state of the universe is not a good idea.

In most expert judgment applications, the experts' knowledge represents the state of the existing or available knowledge. In that sense, inference can be made as follows: the results from the experts' information can be used to draw conclusions about the existing or available knowledge base which may or may not represent the true state of nature. In other words, the inferences that can be made are not necessarily relevant to truth. Also the inferences that can be made are not statistically based inferences. Example 18.1 illustrates this distinction.

Example 18.1: ***Expert Judgment Inference Versus Statistical Inference***

Statistical inference allows the analyst to draw conclusions from a statistically valid, or representative sample taken from a population which has a true value of some quantity of interest (parameter). The inference is made accompanied by a probability statement

specifying the chance that the conclusion is incorrect. For example, the parameter of interest, P, is the probability of an earthquake scaled at 8.0 or greater at a particular location. There is a true value for this probability in nature based on the entire history and future of the earth at this location. Ideally, the analyst draws a random sample of times, n, throughout the planet's history and future. He counts up the number of instances that such an earthquake occurred, x, and estimates the true probability of such a quake by x/n. This estimate would be representative of the true value from a statistically valid sample.

Such an example is ridiculous in reality. The information base required is not available. Some information about earthquakes and potential conditions for one are available to the expert. His knowledge and expertise are all that is available for estimating the probability at this location. Suppose that existing historical data is that no evidence or record of an earthquake exists. That does not mean that one never happened. The expert is still the primary source of all the information that is available. The expert is carefully interviewed using the techniques in this book. The information is recorded. His final estimate, p, is given under sets of conditions that are plausible based upon what is known.

The expert's estimate, p, is not the same as the mythical x/n. Also, p cannot be interpreted as representing the true value of P, whereas x/n can be interpreted for P.

What, then, is p? The value p is representative of the only existing information about P. It may be the best information that will ever be available in the history of man. It is subject to bias, to change, and to misinterpretation. It is not a statistical estimate of the parameter P. It may be considered a single datum from a sample of possible estimates. Analysts realize that statistical inferences are virtually impossible from a sample size of 1. However, p is relevant, useful information.

How is p useful if statistical inference is not possible? It is representative of the state of knowledge. It is information that was not previously known. It can be interpreted in conjunction with any caveats from conditions and assumptions that the expert made.

■

This limited ability to infer is bothersome to many analysts who are accustomed to drawing statistically based inferences about a population (the truth) from a statistical sample (the data). This limited ability to infer is also what leads many analysts and many experts to believe that expert information is not valid data and cannot be used. One of the misconceptions listed in chapter 2 deals with the issue of how to interpret expert judgment as valid data. In that section, the foundation for the entire book was laid with the claim that information from experts (data) was like any other data in that it must be carefully gathered, analyzed, and interpreted. Example 18.2 illustrates how an expert judgment application is identical to an experiment regarding the treatment of the data.

Example 18.2: ***Expert Judgment Data Versus Experimental Data***

A chemist is asked to determine the composition of a chemical mixture. He measures the easily identifiable elements and compounds first to determine composition. To do this, he uses his calibration standards, instruments, and test procedures. He then tackles the more difficult remaining items. He eliminates some and suspects some as being present based on his own experience of how elements and compounds could mix together. He completes the analysis and presents the results.

Most scientists, analysts, and laymen would consider the chemist's results as good experimental data. They would even make inferences (statistical ones) based upon the results. Is this a good idea? What are the problems with the above experimental data that affect the inference process? In answering these questions, it turns out that this experimental data is no better than any expert information. Here's why:

First, the chemist used his judgment and expertise to make decisions about which tests he would use. He went so far as to decide, based on his own experience, which items were likely to be there or not. These are the conditions and assumptions that the chemist made which are parallel to conditions and assumptions any expert makes in solving a problem.

Second, the instruments used have measurement problems. Calibration, standards used, and operator errors are commonplace sources of bias and lack of precision. The data recorded from these are not the ideal samples required to make statistically based inference. These biases and lack of representativeness of the data are parallel to the same things discussed in this book in expert data elicitation (see particularly *Pitfalls* in chapter 2).

Why then is the chemist's experimental data more acceptable than data from an expert's answer to a question? From this experimental description, the inferences possible for the chemist's data are no better than those for expert data. From a statistical perspective, neither data can be used for statistically based inference. However, both can be used to reflect the best state of knowledge available.

■

The above illustrations indicate a philosophy about data interpretation and inference. However, such a philosophy is necessary for expert judgment applications. It provides a logical and defensible (if required) structure for the need and use of expert information.

Improving the Inference Process

It is highly desirable to do everything possible to accurately elicit and analyze the information from the experts to get the best existing knowledge. Because inference stems from the interpretation of the results, careful interpretation becomes important. Care and improvements can be made from the design aspects of the elicitation and from the analyses done.

Design-Based Improvements

Proper experimental design is always the key to obtaining the most and best information for making inferences. This statement is true in experimental and expert judgment applications. The ways of accomplishing design-based improvements are given in the book and will be referred to by chapter numbers for this discussion. Basically, proper design includes coordinating elicitation with analysis methods by (1) structuring the questions and response modes, (2) monitoring granularity, (3) recording all information from the experts (conditions), and (4) performing quantification.

Synergism of elicitation and analysis

The interrelationship between parts II and III of this book may not appear obvious to the reader. However, the elicitation (part II) is designed with a focus on the experts, and the analysis (part III) is designed with a focus on the elicitation.

There are many different ways of analyzing expert answer data (usually quantitative in structure). The methods in part III are chosen specifically to match with the elicitation methods in part II. The choices made by the authors are mainly based on personal experiences while working with both parts, the elicitation and the analysis.

Too often these are not connected (if at all) until the elicitation is completed. The process of inference then becomes pure guesswork. The analyst is unfamiliar with the ways that the data were gathered or with the forms in which it was gathered. Also, the analyst probably has his own favorite methods that will more likely than not require assumptions or model formations that are not appropriate to the data or to the way it was gathered. This situation results in poor-to-bad inferences. This situation also contributes to the bad reputation of expert judgment data.

If the analyst is not the data gatherer, he should at least be involved with the person gathering the data at all stages of the project, from designing the questions (chapters 4 and 5), to selecting the elicitation components and tailoring the elicitation (chapters 6 and 7), to pilot testing and final elicitation (chapters 9 and 10). Decisions made by the data gatherer concerning the methods and components of the elicitation can provide important information on the problem to the analyst. Also, the analyst needs to know what form the gathered information will be in for the analysis. The analyst can help by avoiding during the elicitation some of the problems involving granularity, quantification, and conditionality.

Granularity

In the analysis part of expert judgment applications, monitoring the level of detail reflected in the information at each analysis step has been emphasized. Granularity can change within or between the analysis steps without notice. Granularity can also change in the elicitation. The best defense against changing granularity is proper recording of the elicitation to monitor such changes.

Inferences are made at the granularity that is the most general for all the steps. Therefore, it may not be possible to improve inferences made by monitoring granularity; however, inferences may be erroneous if proper attention is denied.

In the data-gathering and data-base-formation steps, it is likely that the data base (chapters 12 and 13) contains variables (information) with different granularities. The qualitative information can be thought of as more general for analysis purposes than the quantitative. The quantitative variables can be categories or ranks; categories being more general than ranks and both being more general than variables measured on a continuous numerical scale. In the model formation step (chapter 15), variables of differing granularities tend to be modeled together. Example 18.3 illustrates how the granularity should be monitored and what inferences can be made as a result of different granularities.

Example 18.3: ***Inference and Granularity***

In an expert judgment application, six experts were asked to provide estimates of how well experimental results matched results from a simulation code. The data base included variables on problem-solving features. Three variables indicated assumptions used, two indicated cues used, and one related to a definition. The data are as follows:

Expert	A_1	A_2	C_1	C_2	D_1	*ANSWER*
1	1	13	2	N	B	0.25
2	1	20	2	Y	A	0.10
3	2	33	1	Y	C	1.00
4	3	27	2	Y	A	0.50
5	3	25	1	N	B	0.45
6	2	30	2	Y	B	0.95

The variables A_1 and D_1 are category variables where the numbers 1, 2, and 3 and the letters A, B, and C refer to qualitative descriptions. Variable C_1 is a rank variable where the values 1 and 2 mean that 2 is twice as important as 1. Variable C_2 is also categorical where Y is *yes* and N is *no* indicating whether or not a certain cue was used. Variable A_2 is a numerical variable indicating the values assumed for a parameter important to the problem. The answers are on a continuous linear scale describing degree of agreement from 0.0 to 1.0 in value.

Using regression analysis (GLM), the single best predictor variable for *ANSWER* is A_2. If these 6 experts were a representative sample of all experts, then statistical inference about A_2 would be possible. However, no such claim can be made, and only the more limited inference is possible: A_2 may be an important variable for determining answers for these 6 experts. It also turns out that the variables D_1, C_1 and C_2 are significant variables in this regression. Their degree of importance as predictor variables is according to the order listed, D_1 being the most.

This set of 4 predictors from a regression analysis represents a mixture of granularities. With such a mixture and the lack of a statistically valid sample, only a limited inference is possible. In this case, the information gained is merely that these 4 variables are possible conditions for determining the answers. They are not to be used as predictors for other answers as would be the inference made with significant variables from a regression analysis. As possible conditional variables, they can be considered for use in aggregation; they can be used to answer the important question: Are the experts all solving the same problem?

■

Example 18.3, illustrated granularity problems with modeling. The aggregation step (chapter 16) has similar problems. However, aggregation involves combining final answers or answer distributions by some weighting scheme, and usually these answers are already at the same granularity. The same granularity might not hold if different aggregation methods were used on the same data set. Therefore, granularity must still be monitored.

In characterizing uncertainties (chapter 17), granularity can become a severe problem because the uncertainties may represent one level and the answers represent another. Also, the uncertainties from different experts may represent different granularities as illustrated in example 18.4.

Example 18.4: ***Granularity and Uncertainty Characterization***

Suppose that the six experts in example 18.3 provided the answers given above and provided uncertainty estimates as follows:

Expert	Answer	Uncertainties	Expert's Definition
1	0.25	(0.10, 0.50)	Range
2	0.10	(0.00, 0.20)	5th and 95th percentiles
3	1.00	(0.90, 1.00)	Range
4	0.50	(0.25, 0.75)	5th and 95th percentiles
5	0.45	(0.00, 1.00)	Absolute min. and max.
6	0.95	(0.90, 1.00)	Range

These uncertainty estimates do not have matching definitions (or matching granularities). The percentile definitions from experts 2 and 4 will not represent a 90% coverage interval. The definitions might represent more of a 30 to 40% coverage interval. The ranges from experts 1, 3, and 6 have no obvious interpretation. The values from expert 5 appear to be useless. However, if expert 5 claims that his range really represents his assessment of the true uncertainty, then those values are valid uncertainty characterizations even though they cover the entire possible range of the answers.

A common, but more general, definition for the uncertainty values is needed for all experts. One such solution would be to assume uniform distributions for each expert using the ranges provided as the upper and lower limits for the uniform distributions. Monte Carlo sampling from these uniform distributions provides a useful uncertainty analysis method in this case. However, the results of such an analysis would reinforce what is already evident from the ranges of values provided by the experts. The uncertainty is large enough to cover the entire possible range of values for the answer. In such a case, the granularity problem is overshadowed by the large uncertainty.

■

Modeling, aggregation, and uncertainty analyses are important for making inferences in ways other than those regarding granularity. These are discussed in another section below.

Quantification

Granularity is actually a part of quantification as it is broadly defined in this book. Quantification is defined as transformations from one type of data to another rather than as the traditional definition of transforming qualitative information to numbers. Qualitative information is considered more general than quantitative; ranks (or integers) more general

than (continuous) numbers; categories more general than ranks; descriptions more general than categories. The granularity/quantification connection affects the inference process.

In cases where data is sparse or missing for a set of variables, categories, ranks, or numerical values are summed or collapsed to form new variables or new categories for a variable. Summing or collapsing variables also changes granularity from specific to more general. The inferences must be made in terms of the new, more general variables and not in terms of the original information. An example of changing granularity is given in example 18.5 using one of the quantification methods, the Saaty method, which finds numerical weights from qualitative information.

Example 18.5: ***Granularity and Quantification Using Saaty's Method***

An expert is asked to compare and evaluate the likelihoods of seven different events. By using Saaty's pairwise comparison method, the relative evaluations (a general granularity) using the Saaty scale follow:

1 versus 2------- 0.50
1 versus 3------- 0.33
1 versus 4------- 1.00
1 versus 5------- 2.00
1 versus 6------- 3.00
1 versus 7------- 2.70

2 versus 3------- 0.50
2 versus 4------- 3.00
2 versus 5------- 4.00
2 versus 6------- 6.00
2 versus 7------- 5.00

3 versus 4--------2.00
3 versus 5--------3.00
3 versus 6--------3.00
3 versus 7--------3.00

4 versus 5--------2.00
4 versus 6--------3.00
4 versus 7--------2.00

5 versus 6--------2.00
5 versus 7--------1.00

6 versus 7--------0.50

The resulting relative weights are

$$(0.13, 0.27, 0.28, 0.13, 0.07, 0.05, 0.07) \quad .$$

Even though the weights are numerical, the interpretation must be made on a relative basis. For example, events 5 and 7 are *not* half as likely as events 1 and 4, nor are they one-fourth as likely as events 2 and 3. The only interpretation is that possible for the 7 events. The events judged most likely are 2 and 3, and the events judged least likely are 5, 6, and 7.

■

Conditionality

The concept of conditionality refers to the fact that the answers given by the experts are conditioned on many aspects of the problem, the elicitation, and the experts themselves. In chapters 12 and 13, analysis techniques are discussed that emphasize how to search for these conditions. Here their importance is in interpreting the answers.

The major focus for investigating conditions is to guarantee that the experts are all solving the same problem and that the problem being solved is indeed the one being asked. If conditions relating to problem-solving features such as definitions, assumptions, and heuristics used by the experts are found to have an effect on the answers given, then these features must be examined to determine if they somehow change the problem.

It is also recommended that some background information on the experts be gathered as possible sources for influential conditions. It has long been speculated that experts' backgrounds are a source of correlation among experts (Baecher 1979); however, recent studies (Booker and Meyer 1988a , Meyer and Booker 1987b) have indicated that evidence for influential background features is lacking.

Conditions can be found in the design features of the elicitation and in the environment. Questions must be carefully formulated (chapter 5) and the elicitation carefully implemented (chapter 8, 9, and 10) to minimize biases and other conditions that can affect the answers given by the experts. Other, uncontrollable conditions, such as the expert's mood, the room being uncomfortable, a disturbing recent event, should be noted and documented as possible important conditions.

Recording and monitoring conditions is essential in order to determine if they are important. It is also vital that the analyst and data gatherer not be responsible for inducing any additional conditions by the way they analyze and elicit the data. This warning includes the inference process. The analyst can impose his own views to the extent that he interprets the data in the way that he desires. Conditions that might not really be important can be used as excuses to disregard data that do not fit preconceived ideas. Even conditions that might be important can be falsely used for this purpose.

There are other ways in which inferences can be erroneously made by placing too much importance on conditions. While it is important not to ignore conditions, it is also important not to use them as excuses, cover-ups, or justifications. Conditions found to be significant or important may only be masks or indirect effects for some other effect that cannot be monitored. Conditions found to be important must also be relevant to the problem. Only then are interpretations made with conditions given as caveats or qualifiers.

The basic philosophy regarding possible effects from conditions is threefold. First, proper elicitation and analysis is designed to reduce any effects induced by the data gatherer and analyst. Second, the data gatherer and analyst should control conditions that are controllable, and they should record information on *any* conditions that are observable. Third, significant and relevant conditions are stated as part of the conclusions and as reflecting the inferences made.

Analysis-Based Improvements

The methods chosen in the analysis section reflect a conservative, cautious interpretation of the results. This caution is borrowed from the reactor design and probabilistic risk analysis communities. The use of redundancy and cross validation plays a major role in the analysis and design. In expert judgment analyses, this philosophy means that more than one technique is used to determine the results. The consistency of results is therefore validated. A simple way to cross validate is to use simulation techniques. Simulation has the added advantage of not relying on assumed distributions or on methods based on those assumptions.

Cross validation and redundancy

There are multiple ways of analyzing a data set. Many of the methods presented in part III of this book are similar and produce similar types of results that can be used together to validate each other or to provide redundancy of analyses.

A regression (GLM) analysis provides information on how condition variables affect the answers. Correlation also determines variable relationships. If two variables are significantly correlated, they will usually be significant terms in a regression model.

The various multivariate techniques are also interrelated. If a conditional (ancillary) variable is a good discriminator, then it will also be an effective predictor in a regression model. If cluster analysis reveals definite clusterings, it is possible to find a discriminating variable that is responsible for the clusterings. Factor analysis reveals, as can cluster analysis and correlation analysis, how the different variables are related to each other.

Therefore, the analyst can run different techniques and determine if the results are consistent. If the results are consistent, that lends strength to the conclusions. If the results are inconsistent, then there is either trouble with using the techniques (e.g., violation of the required assumptions) or trouble with the interpretations (e.g., a variable is not really important in determining the answers). Example 18.6 illustrates the use of redundant techniques.

Example 18.6: ***Inference Using Redundant Techniques***

Ten experts are asked to answer five questions, A_1 through A_5. For each question, a general problem-solving variable, P_1 through P_5, was found by summing up all the problem-solving features for that question. Each problem-solving variable was found to be significant in the regression analyses done for the answers. The individual problem-solving features and all other ancillary information (variables) were not found to be significant in the regressions.

Variables A_1, A_3 and A_4 were trimodal in structure. The three modes formed three clusters. Variables P_1, P_3, and P_4 were found to be significant discriminating variables for their respective answer variables, and they successfully predicted the three clusters for each answer variable.

Variables A_2 and A_5 had no identifiable structure. A cluster analysis of all the variables for questions 2 and 5 indicated that A_2 and P_2 formed a cluster, and A_5 and P_5 formed a cluster.

Variable A_1 was unimodal in structure. A correlation coefficient of A_1 and P_1 was found to be a significant value of 0.87.

Therefore, for each A_i variable, additional evidence was found to support the claim that the P_i variable might be important for determining A_i.

Having supportive evidence for the P_i variables as conditions for the answers, it is now necessary to specify this in the statement of conclusions. The answers given by the experts are conditioned on their general problem-solving processes (not specific problem-solving features). For aggregation, it would be desirable to examine the values of the P_i variables for weight determinations. For inference, it would also be desirable to use the P_i variables to determine whether or not all the experts were solving the same (the given) problem.

Suppose the additional evidence was not so clear. If the A_2, P_2 and A_5, P_5 variables did not cluster, then the conditionality would not necessarily apply for these questions. Other techniques would have to be tried. If none of the other techniques provided evidence of variable relationships between the A variables and the P variables, then the conditionality would be suspect. These potential conditional variables would have little or no importance for aggregation and interpretive purposes.

■

Because the inference process itself is weak in expert judgment applications, there must be strong supportive evidence for conditionality. Redundancy or cross validation helps to determine the degree of strength.

Simulation

Cross validation and redundancy checks can be provided by using simulation. There are at least three good reasons for using simulation to improve the inference capability in expert judgment applications: (1) sample sizes are typically small (less than 5 or 10 experts); (2) distributional forms for the answers do not follow convenient forms (such as the normal) and are usually multimodal or distribution mixtures; (3) estimations for variances, percentiles, or central measures (such as the median) are desired for inference but are difficult to obtain without specified distributional forms. At the very least, simulation allows the analyst the freedom to explore and check things in an empirical (data-based) manner.

It is difficult to make statistical inferences with small sample sizes. Most statistical techniques rely on asymptotic or theoretic results that require sample sizes of 30 or more. Designing for samples of 30 or more would be totally impractical in expert judgment applications. For one reason, the elicitation would be too time consuming and expensive. For another reason, there may not be 30 experts in existence.

It is equally difficult to make statistical inferences without specified distributional forms. Many statistical techniques require distributional forms such as normality. This is true of many of the multivariate analysis techniques such as discriminant analysis. Other techniques such as regression have less strict distributional requirements, requiring only that the residual or model errors be normal rather than the data itself.

It has been emphasized that all estimators (such as the mean or median) should be accompanied by variance estimates or interval (e.g., percentile values) estimates. Providing such estimates is part of the inference process and part of establishing the uncertainty. A single-valued estimator does not provide any information about the variability or uncertainty surrounding it. A single-valued estimator implies a precision in the results that is not present. A variance or interval estimator conveys the appropriate state of uncertainty and variability. It is difficult to estimate variances of estimators such as the median without distributional forms. In some cases it is difficult even with distributional forms because the formulas are not tractable.

Simulation can provide solutions to the difficulties from small sample sizes and required assumptions. It allows the analyst to make the most of a small sample size using such techniques as the bootstrap. Reliance upon asymptotic or theoretic results is not necessary, and simulation provides the way of obtaining estimates for variances and other quantities without theory, distributional assumptions, or difficult calculations.

Inferences with Modeling, Aggregation and Uncertainties

The relationship between the inference process and the results obtained from modeling (chapter 15), aggregation (chapter 16), and uncertainty (chapter 17) was introduced in the discussion of granularity above. These three steps in part III relate to the inference process in other ways.

The models in chapter 15 are used to identify relationships between the conditional (ancillary) variables and the answer variables. The major category of these models are the GLMs, which are the backbone of statistical modeling techniques. Other conditional models were also suggested that did not have a strong statistical basis. The purpose of both types of models is *not* to identify functional relationships among the variables to be used for prediction purposes; it is not even to specify model functions (equations); but the purpose is only to provide clues about variable relationships that could be verified by using other procedures (or other procedure's results could be verified using these models). The reason for such a limited purpose is that model assumptions may be lacking.

The purpose of aggregation is to formulate the final results as a combined estimate with a variance or as a distribution with central values (mean or median) and dispersion measures (variance or 5th and 95th percentiles). These final estimates or distributions are interpreted as the cumulative knowledge for the parameter of interest (the answer to the question). This interpretation is part of the inference process. These results do not have a statistical interpretation that relates the estimates to the true value. The dispersion estimates do have a valid interpretation relating to the uncertainty in the state of existing knowledge, and a conclusion can be made using the uncertainty characterization. However, the ideal goal of aggregation to form an estimate or distribution that accurately reflects the truth is not realizable. One could argue that this unrealistic goal negates the reason for aggregation; however, aggregation does provide a convenient summary or combination of all the available information.

As mentioned in the simulation section above, uncertainties are an integral part of the inference process. In representing or characterizing uncertainties, their existence is acknowledged as well as estimated. Drawing conclusions without accounting for uncertainties makes the information appear more precise than is true (bad inference). As emphasized in chapter 17, uncertainties are an important part of every experiment or application and cannot be ignored, especially when it is time to interpret the results.

Final Comments

With only weak inferences possible, a natural question becomes: Why take such care in gathering and analyzing the data? The answer to this was stated in Part I of the book: Expert judgment data is like any other data. It must be carefully gathered, analyzed, and interpreted. Careful interpretation, in this case, unfortunately translates to limited inferences. Trying to do otherwise violates the true content of the information gathered. A cliche is applicable here: *You can't squeeze blood out of a turnip*. This means that one cannot get better information than that which exists.

Other cliches and phrases are applicable to the philosophy of this book and to the inference process:

1. ***Rome wasn't built in a day.*** This means that more research is needed and that this book represents only a start in the research and efforts needed to resolve the many problems in expert judgment elicitation and analysis. Better inference will be possible with better techniques and understanding of the expert information.
2. ***Nothing good comes easily.*** This means that all the steps and suggestions offered here may seem tedious and unnecessary, but to obtain good quality data takes time and effort. Even limited inferences can only be made if a good job is done.
3. ***Take it with a grain of salt.*** This means that the results are accompanied by a list of caveats and conditions, and interpretation must include these.
4. ***Keep it simple.*** The methods presented are designed to be feasible and usable by data gatherers and analysts. Many other techniques exist, and some of them are referenced. No evidence exists that the more complex ones omitted are better than the simple ones offered. In fact, many of the more complex methods do not perform as well as the simpler ones do.

In conclusion, the inference process may be disappointing in that the results and conclusions available do not extend to the truth as is done in statistical inference. However, information has been gained that was previously unknown, and that is the sole reason for eliciting and analyzing expert judgment.

Appendices

Appendix A--SAATY

```
      program saaty
c
c  Uses Saaty's own FORTRAN subroutine for calculating the weights
c  for a single level (matrix) (Saaty 1982).  For more than 1 level
c  weights can be calculated using this code for each matrix and then
c  combined either by hand calculation or by modifying this code.
c
c  The input to this code can be done on the terminal or
c  by contructing input file SAATY.IN which has:
c    line 1: nfctr = number of experts (factors), maximum =10.
c    line 2: contains (nfactr-1) pairwise comparisons in f5.2 format,
c            comparing item 1 with items 2-nfct using scales **
c    line 3: contains (nfctr-2) pariwise comparisons in f5.2 format,
c            comparing item 2 with items 3-nfactr.
c    etc.
c    line (nfctr-1): contains 1 pairwise comparison in f5.2 format,
c         comparing the last two items [item nfctr to item (nfctr-1)].
c
c   **Values for pairwise comparisons can be taken from:
c   1)the Saaty integer scales: values = 2-9
c     where the first of the pair is better or more likely
c     than the second;
c     value = 1 where the pair is identical;
c     values = 1/2-1/9 where the first of the pair is worse or
c     less likely than the second.
c   2)the pairwise comparisons can be qualitative, using
c     a triplet of choices (better, same, worse) rather than the
c     numerical scale, the values for this triplet are:
c     (2.72, 1.00, 0.37), respectively.
c
c
      dimension w(60)
      open (unit=11,file='saaty.in',status='old')
      open (unit=22,file='saaty.out',status='new')
      write(*,'("To use input file SAATY.IN, type 1 else type 0: ",\)')
      read(*,*) infile
      if (infile.eq.1) then
        read(11,*) nfctr
        go to 61
      endif

      write (*,'("Enter the number of factors (up to 10): ",\)')
      read  (*,*) nfctr

   61 continue
      call mtxin(w,nfctr)
c      write (*,52) nfctr,(w(i),i=1,nfctr)
   52 format('Normalized weights for the',i3,' factors:',/,10f10.6)
```

```
      stop
      end
c
c
c
      subroutine mtxin(w,nfctr)
      dimension c(60,60),w(60),cw(60),w2(60),rct(10)
      data rct/.0,.0,.58,.90,1.12,1.24,1.32,1.41,1.45,1.49/
   16 continue
      if (infile.eq.0) write (*,9)
      nfctr1=nfctr-1
      do 10 i=1,nfctr1
    2 continue
      if(infile.eq.0)write (*,3) i
      i1=i+1
      if(infile.eq.0)read (*,4) (c(i,j),j=i1,nfctr)
      if(infile.eq.1)read(11,4) (c(i,j),j=i1,nfctr)
      if(infile.eq.0)write (*,5) i, (c(i,j),j=i1,nfctr)
      if(infile.eq.1)write (22,5) i, (c(i,j),j=i1,nfctr)
      if(infile.eq.0) write (*,1)
    4 format(10f5.2)
    5 format('Row',i3,' is:',10f5.2)
    1 format(' If not correct type 9, if correct hit return.')
    3 format(/,'Enter row',i3,' (use f5.2 format): ')
    9 format(/,'The upper triangular part of the matrix:')
   18 format(f1.0)
      if(infile.eq.1) yn=4
      if(infile.eq.0)read (*,18) yn
      if(yn.gt. 6) go to 2
      do 10 j=i1,nfctr
      if(c(i,j).lt.0.) go to 6
      if(c(i,j).ge.0.) go to 8
    6 c(i,j)=-(1.0/c(i,j))
    8 c(j,i)=1.0/c(i,j)
   10 continue

      do 14 i=1,nfctr
   14 c(i,i)=2.
      ts=0.
      do 24 i=1,nfctr
      s=0.
      do 22 j=1,nfctr
   22 s=s+c(i,j)
      w2(i)=s
   24 ts=ts+s
      do 26 i=1,nfctr
   26 w2(i)=w2(i)/ts
      k=0
   27 ts=0.
      k=k+1
      do 30 i=1,nfctr
      s=0.
      do 28 j=1,nfctr
   28 s=s+c(i,j)*w2(j)
      w(i)=s
   30 ts=ts+s
      d=0.
      do 38 i=1,nfctr
```

```
      w(i)=w(i)/ts
   38 d=d+abs(w(i)-w2(i))
      if(k.gt.10000) go to 42
      if(d.lt.1.e-15) go to 42
      do 37 i=1,60
   37 w2(i)=w(i)
      go to 27
   42 continue
      if(infile.eq.0) write(*,47)
      if(infile.eq.1) write(22,47)
   47 format(/,'The final matrix is:')
      do 40 i=1,nfctr
      c(i,i)=1
      if(infile.eq.1) write (22,41) (c(i,j),j=1,nfctr)
   40 if(infile.eq.0) write (*,41) (c(i,j),j=1,nfctr)
   41 format(10f6.3)
      do 46 i=1,nfctr
      s=0.
      do 44 j=1,nfctr
   44 s=s+c(i,j)*w(j)
       cw(i)=s
   46 continue
      s=0.
      do 48 i=1,nfctr
   48 s=s+cw(i)/w(i)
      ymax=s/nfctr
      ci=0.
      cr=0.
      if(nfctr.le.1) go to 49
      ci=(ymax-nfctr)/(nfctr-1)
      if(nfctr.le.2) go to 49
      cr=ci/rct(nfctr)
   49 continue
      if(infile.eq.0) write (*,50) (w(i),i=1,nfctr)
      if(infile.eq.1) write (22,50) (w(i),i=1,nfctr)
   50 format(/,'Normalized weights=',(10f6.3))
      if(infile.eq.0)write (*,52) ymax,ci,cr
      if(infile.eq.1)write (22,52) ymax,ci,cr
   52 format(' principle eigenvalue, lmax =',f6.3,/,
     *'Consistency index (deviation of lmax from n) =',f6.3,
     * /,' consistency ratio (should be < .10) =',f6.3)
      if(infile.eq.0) write (*,54)
   54 format(/,'If you want to redo this matrix,',
     * 'type 9, else hit return:')
      if(infile.eq.0)read (*,18) yn
      if(yn.ge.5) go to 16
      return
      end
```

Appendix B--MCBETA

```
      program mcbeta
c
c   Monte Carlo uncertainty analysis code
c    for beta distributions.
c
c   Important!!!!!!!!!!!! user must change the function for
c    combining the set of experts, events, or sequence as:
c    FUNCTION EVAL(PR)
c    where pr is array of probabilities of primary events,
c    and the function is aggregation estimator for all the betas.
c
c   Input is supplied using file MCBETA.IN, as unit 11,
c     see format below.
c   Output is sent to two files:
c     MCSTATS.OUT which contains the Monte Carlo results
c     BET.OUT which contains the fitted beta distributions results.
c     MCSTATS.OUT is unit 22 and BET.OUT is unit 33
c
c   Betas are fit with two supplied estimates either as:
c    1) 2 percentile estimates and levels (e.g. 0.05 & 0.95)
c    2) 1 percentile estimate and level and a mean value.
c
      parameter (nbmax=500,klmax=100,ndmax=100,nhmax=10000,nhp1=nhmax+1)
      dimension ppbig(10000)
      common/primary/pr(nbmax),param(klmax,ndmax,2)
      common/options/nb,nruns
      common/utility/dumy(nhp1)
      save /utility/, /options/, /primary/
      open (unit=22,file='mcstats.out',status='new')
      call input
      nout=0
      do 10 n=1,nruns
 5    call pgen
      prob=eval(pr)
      if (prob.ge.0..and.prob.lt.1.01) go to 8
      nout=nout+1
      write (22,1222) n,prob
      go to 5
 8    continue
      ppbig(n) = prob
 10   continue
      call finish (ppbig,nruns)
      write(22,1000) nout
 1000 format(//' Simulation generated',i5,
     *     ' values not in (0,1)')
 1222 format(' n=',i6,' prob=',e16.8)
      close (unit=22)
      stop
      end


      subroutine input
c
c   Read parameters from file 'MCBETA.IN'.
```

```
c
c     line    description
c
c     1       nb = number of primary events (le nbmax)
c             nr = number of monte carlo runs, max = 10000
c     2       iibet = 2 for 2 percentiles, = 1 for 1 percentile & mean
c                     next nb lines free format
c             two percentiles for the beta case:
c             upper estimate, lower estimate,
c             upper level (.95), lower level (.05)
c             or
c             mean, percentile value, percentile level.
c
      parameter (nbmax=500,klmax=100,ndmax=100)
      common/primary/pr(nbmax),param(klmax,ndmax,2)
      common/options/nb,nruns
      save /options/,/primary/
      dimension t(2)
      open (unit=11,file='mcbeta.in',status='old')
      open (unit=33,file='bet.out',status='new')
      read (11,*) nb,nr
      read(11,*) iibet
      nruns=nr
      do 10 i=1,nb
c
c  Subroutines twoper & meanper find the beta parameters, x0 fn0
c
        if(iibet.eq.2) call twoper(x0,fn0)
        if(iibet.eq.1) call meanper(x0,fn0)
        t(1)=x0
        t(2)=fn0
      param(i,1,1)=t(1)
      param(i,1,2)=t(2)
 10   continue
      close (unit=11)
      close (unit=33)
      return
      end



      subroutine pgen
      parameter (nbmax=500,klmax=100,ndmax=100)
      common/primary/pr(nbmax),param(klmax,ndmax,2)
      common/options/nb,nruns
      common/utility/p(klmax,ndmax)
      save /primary/, /options/, /utility/
      do 100 n=1,nb
      prm1=param(n,1,1)
      prm2=param(n,1,2)-prm1
  82  if(prm1.gt.1.0) gx1=gt(prm1)
      if(prm1.eq.1.0) gx1=gs(1.)
      if(prm1.lt.1.0) call gl(prm1,gx1)
      if(prm2.gt.1.0) gx2=gt(prm2)
      if(prm2.eq.1.0) gx2=gs(1.)
      if(prm2.lt.1.0) call gl(prm2,gx2)
      p(n,1)=gx1/(gx1+gx2)
      if(p(n,1).gt.1.0) go to 82
```

```
 100  continue
      do 200 i=1,nb
  200 pr(i)=p(i,1)
      return
      end
c
c ***************************************************************************
c
c   FUNCTION EVAL: must be changed for each new problem
c
c ***************************************************************************
c
      function eval(x)
      dimension y(150),x(500)
      ymedian = (x(1) + x(2) +x(3))/3.0
      eval =  ymedian
      return
      end
c
c
c
      function gs(alp)
c
c  Routine for generating gamma variates with
c   shape parameter less than 1.  )
c   Ahrens, J. H. and Dieter, U. (1974)
c   "Computer Methods for Sampling from Gamma,
c    Beta, Poisson, and Binomial Distributions,"
c   Computing, vol 12, p.223-246.
c
      data ex/2.718281828459045/
      bet=1.0
 1    u1=rnd(0.)
      b=(ex+alp)/ex
      p=b*u1
      if (p.gt.1.) go to 3
 2    x=exp(alog(p)/alp)
      u2=rnd(0.)
      if (u2.gt.exp(-x)) go to 1
      go to 10
 3    x=-alog((b-p)/alp)
      u3=rnd(0.)
      if (alog(u3).gt.(alp-1.)*alog(x)) go to 1
 10   gs=x*bet
      return
      end
c
c
c
      subroutine twoper(zpar,znpar)
c
c  This program computes the parameters of a beta distribution
c   given two percentiles of the distribution
c
c  The following values are to be read in from MCBETA.IN:
c     r2=upper percentile estimate
c     r3=lower percentile estimate
c     y=percentile level of r2 (e.g. .95)
```

```
c     w=percentile level of r3 (e.g. .05)
c
c  The followign values are calculated as the beta parameters:
c    zpar=the value of x0 in single precision
c    znpar=the value of n0 in single precision
c
      implicit real*8(a-h,o-y)
      common/per2sub/n,int,ifin
      save /per2sub/
      read(11,*)r2,r3,y,w
      a1 = .0001
      b1 = 1000.
      n = 0
      int=0
      ifin = 0
   30 call perint(an,bn,n)
      call betasub(an,bn,a1,b1,r2,r3,y,w,ppar,xnpar)
      if(ifin.ge.1)go to 20
      if(int.gt.2) go to 40
      if(int.ge.1) go to 30
   40 write(*,41)
   41 format(' parameter value is greater than 100000.' )
   20 continue
      zpar=ppar
      znpar=xnpar
      return
      end
c
c
c
      subroutine betasub(an,bn,a1,b1,r2,r3,y,w,ppar,xnpar)
      implicit real*8(a-h,o-y)
      common/per2sub/n,int,ifin
      save /per2sub/
c
c  Prints to output file for beta information, BET.OUT
c
      write(33,31)
   31 format(' RESULTS FOR THIS BETA DISTRIBUTION:',/,/)
      write(33,33) r2,y,r3,w
   33 format(' upper percentile= ',f12.6,' level=',f12.6,/,
     *' lower percentile=', f12.6,' level=',f12.6)
      x0=betpar(a1,b1,r2,an,y)
      ax= dbti(r3,an,x0)
      x1 = betpar(a1,b1,r2,bn,y)
      bx = dbti(r3,bn,x1)
   15 cn = (an + bn)/2.0
      x2 = betpar(a1,b1,r2,cn,y)
      cx = dbti(r3,cn,x2)
      if(cx-w) 50,60,60
   50 bn = cn
      go to 70
   60 an = cn
   70 dn = bn - an
      if(dn .lt. 10e-10) go to 80
      go to 15
   80 ppar = (an + bn)/2.
      qpar = x2
```

```
      prob = dbti(r2,ppar,qpar)
      diff = abs(prob-y) - .0001
      if(diff) 30,20,20
   30 prob = dbti(r3,ppar,qpar)
      diff = abs(prob-w) - .0001
      if(diff)35,20,20
   20 int = int+1
      go to 25
   35 xnpar = ppar + qpar
      bmean = ppar/xnpar
      ifin = ifin + 1
   91 format(' ',a4,/,a80)
      write(33,90) ppar,qpar,xnpar
   90 format(5x,'p = x0 = ',f12.6,5x,'q =  ',f12.6,5x,'  n0 = ',f12.6/)
      write(33,98) bmean
   98 format(2x,'the mean of the prior is  ',f12.8)
      var=(bmean*(1-bmean))/(xnpar+1.)
      std=var**.5
      write(33,85)var,std
   85 format(2x,'variance=',e14.6,'  std. dev.=',e14.6/)
      prob=dbti(r2,ppar,qpar)
      write(33,95) prob
   95 format(2x,'the upper percentile probability is  ',f16.13)
      prob = dbti(r3,ppar,qpar)
      write(33,96) prob
   96 format(2x,'the lower percentile probability is  ',f16.13,//)
   25 return
      end
c
c
c
      function betpar(a1,b1,r2,an,y)
      implicit real*8(a-h,o-y)
      a = a1
      b = b1
   10 c = (a + b)/2.0
      x = dbti(r2,an,c)
      if (x-y) 20,30,30
   30 b = c
      go to 40
   20 a = c
   40 d = b-a
      if (d.lt.10e-10) go to 50
      go to 10
   50 betpar = (a + b)/2.0
      return
      end
c
c
c
      subroutine perint(an,bn,n)
      n = n+1
      go to(10,20,30) n
   10 an=.0001
      bn=1000.
      go to 40
   20 an=1000.
      bn=10000.
```

```
      go to 40
   30 an=10000.
      bn=100000.
   40 return
      end
c
c
c

      subroutine meanper(z0,zn0)
c
c  This program computes the parameters of a beta
c   distribution given the mean and one percentile
c   of the distribution
c
c  The following values are read in from MCBETA.IN:
c       rmean = mean
c       perc = percentile
c       prob = percentile level.
c
c  The following values are calculated:
c       z0 = the value of x0 in single precision
c       zn0= the value of n0 in single precision
c
      implicit real*8(a-h,o-y)
      common/meansub/rmean,perc,prob,n,int
      save /meansub/
      read(11,*) rmean, perc, prob
      n = 0
      int = 0
   30 call parint(a,b,n)
      if(rmean.lt.perc)go to 10
      go to 20
   10 r1 = rmean
      r2 = perc
      y = prob
      go to 38
   20 r1 = 1. - rmean
      r2 = 1. - perc
      y = 1. - prob
   38 call betsub2(r1,r2,a,b,y,x0,fn0)
      if(int.eq.1) go to 35
      if(n.le.3)go to 30
      if(n.gt.3) go to 40
      go to 35
   40 write(*,42)
   42 format(' parameter value is greater than 100,000.' )
   35 continue
      z0=x0
      zn0=fn0
      return
      end
c
c
c
      subroutine betsub2(r1,r2,a,b,y,x0,fn0)
      implicit real*8(a-h,o-y)
      common/meansub/rmean,perc,prob,n,int
```

```
      save /meansub/
c
c  Prints beta information to BET.OUT
c
   92 format(' ',a4,/,a80)
      write(33,15) rmean,perc,prob
   15 format(2x,' the given mean is ',f8.5,//
     * '    the given percentile is  ',f8.5,//
     * '    the given percentile level  ',f8.5/)
      fn0=btpar(r1,r2,a,b,y)
      x0 = fn0*r1
      if(rmean.lt.perc)go to 30
      go to 40
   30 p = x0
      q = fn0 - x0
      go to 50
   40 q = x0
      p = fn0 - x0
      r2 = perc
      x0 = p
   50 prb = dbti(r2,p,q)
      diff = dabs(prb-prob) - .00001
      if(diff.ge.0.) go to 85
      int = int + 1
      write(33,60) fn0, x0
   60 format(2x,' n0 =  ',f12.6,'  x0 =  ', f12.6/)
      write(33,20) r2,prb
   20 format(/2x,' the probability at',f12.7,' is ',f19.17/)
      pm = dbti(rmean,p,q)
      write(33,70) rmean, pm
   70 format(2x,' the probability at the mean ',f7.5,' is',f12.8/)
      pmean = p/(p+q)
      var = (pmean*(1.-pmean))/(p+q+1)
      std = var**.5
      write(33,80) pmean
   80 format(2x,' the mean of the beta prior is x0/n0 = ',f8.5)
      write(33,86) var,std
   86 format(2x,' variance=',e14.6,'  std. dev.=',e14.6)
   85 return
      end
c
c
c
      function btpar(r1,r2,a,b,y)
      implicit real*8(a-h,o-y)
      common/meansub/rmean,perc,prob,n,int
      save /meansub/
   10 c = (a + b)/2.0
      x=c*r1
      q=c-x
      p=dbti(r2,x,q)
      if(p-y)20,30,30
   30 b=c
      go to 40
   20 a=c
   40 d=b-a
      if (d .lt. 1.0e-10)go to 50
      go to 10
```

```
   50 btpar=(a+b)/2.0
      return
      end
c
c
c
      subroutine parint(a,b,n)
      implicit real*8(a-h,o-y)
      n = n+ 1
      go to(10,20,30) n
   10 a = .0001
      b = 1000.
      go to 40
   20 a = 1000.
      b = 10000.
      go to 40
   30 a = 10000.
      b = 100000.
   40 return
      end
c
c
c
      subroutine gl(alp,x)
c
c  Finds gamma values for alpha less than 1.0
c
    1 u1=rnd(0.)
      b=(2.718281828+alp)/2.718281828
      p=b*u1
      if(p.gt.1.0) go to 3
    2 x=exp(alog(p)/alp)
      u2=rnd(0.)
      if(u2.gt.exp(-x)) go to 1
      return
    3 x=-alog((b-p)/alp)
      u3=rnd(0.)
      if(alog(u3).gt.(alp-1.)*alog(x)) go to 1
      return
      end
      function gt (alp)
c
c  Cheng, R. C. H. and Feast, G. M. (1979)
c   "Some Simple Gamma Variate Generators,"
c   Applied Statistics, vol 28, p. 290-295.
c   for alpha .gt. 0.5
c
      data aset/-1./
      if (aset .eq. alp) go to 1
      aset = alp
      a = alp - 0.5
      b = alp / a
      c = 2.0 / a
      d = c + 2.0
      s = sqrt(alp)
      h1 = (0.865 + 0.064/alp) /s
      h2 = (0.4343 - 0.105/s) / s
  1   u1 = rnd(0.)
```

```
      u = rnd(0.)
      if(u1.le.0. .or. u1.ge.1.)go to 1
      if(u.le.0.  .or.  u.ge.1.)go to 1
      u2 = u1 + h1*u - h2
      if (u2 .le. 0.) go to 1
      if (u2 .ge. 1) go to 1
  2   w = b * (u1/u2) * (u1/u2)
      if ( (c*u2-d+w+1./w) .le. 0.) go to 4
  3   if ( (c*alog(u2)-alog(w)+w-1.) .ge. 0.) go to 1
  4   gt = a * w
      return
      end
c
c
c
      function rnd(idum)
c
c  Generates random uniform numbers, use 0=idum
c
      save p,q,m,j,nn

      integer p,q,m(98)
      logical inuse
      data j/40/,nn/2147483647/,p/98/,q/27/
      data (m(k),k=1,50)/
     +1387256442, 539505633,   7126687,2115653676, 480642437,
     +1403109719, 898019591,1609472695, 742049136, 964528840,
     +1774590149, 531014893,1478060509, 224730595,1413365137,
     +1415397063, 370513614,1981855272,1672294721,1559669404,
     +1992066581, 440083042,1552169384, 949029171,1848294689,
     +1014369863,1226252978, 199445637, 552539314, 101995811,
     +1795618857,1468200845, 403608434, 466262418,1783034892,
     +2125486341,1437171068, 839437811, 685760609, 311739045,
     +1876584692, 223544964, 667792106,1829604735, 887026472,
     + 688815796,1153871680,1135467106,1975710098,1393037901/
      data (m(k),k=51,98)/
     + 330755675, 804762632, 393596594,1695657725,  50479950,
     +1039358666,1885424316, 400881551, 142829986, 187416368,
     + 821029919,1292641081, 415120294,1104581275,1258423968,
     + 304285054, 400491932,2014625087,1619263031, 750624285,
     +1996732699,  97476312,1250544934,2145510054,1510875684,
     + 262891578, 616032534,1316668730,1500747974,2138561534,
     + 809719156,1605036043, 510086967, 317411066,  54278455,
     +2052774305, 439191668,1881943474,1397167115,2046084812,
     + 644321591, 328615697,1004646018,1110120728,2007784487,
     + 992677826,1756605308, 796797739/
      if(idum)200,100,300
  100 j=j+1
      if(j.gt.p)j=1
      k=j+q
      if(k.gt.p)k=k-p
      m(j)=m(k).xor.m(j)
      rnd=float(m(j))/nn
      return
  200 iunit=100
  201 iunit=iunit-1
      inquire(unit=iunit,opened=inuse)
      if(inuse)goto 201
```

```
      open(unit=iunit,file='rnd.str',status='unknown')
      read(iunit,*)m,j
      close(unit=iunit)
      return
  300 iunit=100
  301 iunit=iunit-1
      inquire(unit=iunit,opened=inuse)
      if(inuse)goto 301
      open(unit=iunit,file='rnd.str',status='unknown')
      write(iunit,*)m,j
      close(unit=iunit)
      return
      end
c
c
c
      double precision function dbti (x,a,b)
c
c  Incomplete beta function value for x with a, b parameters
c   calls betac and gamln  -- double precision function
c
      real*8 betacf,gamln,a,b,x,bt,one,zero,two
      data one,zero,two/1.d0,0.d0,2.d0/
      if(x.lt.zero.or.x.gt.one) print *,'bad argument x in betai'
      if(x.lt.zero) x=zero
      if(x.gt.one) x=one
      if(x.eq.zero.or.x.eq.one) then
        bt=zero
      else
        bt=exp(gamln(a+b)-gamln(a)-gamln(b)
     *       +a*dlog(x)+b*dlog(one-x))
      endif
      if(x.lt.(a+one)/(a+b+two))then
       dbti=bt*betacf(a,b,x)/a
       return
      else
       dbti=one-bt*betacf(b,a,one-x)/b
       return
      endif
      end
c
c
c
      double precision function betacf(a,b,x)
      parameter(itmax=100,eps=3.d-7)
      implicit real*8 (a-h,o-z)
      data fone/1.d0/
      am=fone
      bm=fone
      az=fone
      qab=a+b
      qap=a+fone
      qam=a-fone
      bz=fone-qab*x/qap
      do 11 m=1,itmax
       em=m
       tem=em+em
       d=em*(b-m)*x/((qam+tem)*(a+tem))
```

```
      ap=az+d*am
      bp=bz+d*bm
      d=-(a+em)*(qab+em)*x/((a+tem)*(qap+tem))
      app=ap+d*az
      bpp=bp+d*bz
      aold=az
      am=ap/bpp
      bm=bp/bpp
      az=app/bpp
      bz=fone
      if(abs(az-aold).lt.eps*abs(az)) go to 1
   11 continue
      pause 'a or b too big, or itmax too small; Hit CR'
    1 betacf=az
      return
      end
c
c
c
      double precision function gamln(xx)
      save cof,stp,half,fone,fpf,x,tmp,ser
       real*8 cof(6),stp,half,fone,fpf,x,tmp,ser
      data cof,stp/76.18009173d0,-86.50532033d0,24.01409822d0,
     *    -1.231739516d0,.120858003d-2,-.536382d-5,2.50662827465d0/
      data half,fone,fpf/0.5d0,1.0d0,5.5d0/
      x=xx-fone
      tmp=x+fpf
      tmp=(x+half)*log(tmp)-tmp
      ser=fone
      do 11 j=1,6
       x=x+fone
       ser=ser+cof(j)/x
   11 continue
      gamln=tmp+log(stp*ser)
      return
      end
c
c
c
      subroutine sort (ra,n)
c
c  Sorts an array RA of length N into ascending numerial order
c   using the Heapsort algorithm.  N is input;  RA is replaced
c   by its sorted rearrangement.
c
      dimension ra(n)
      l = n/2 + 1
      ir = n
   10 continue
      if (l .gt. 1) then
         l = l - 1
       rra = ra(l)
       else
       rra = ra(ir)
       ra(ir) = ra(l)
       ir = ir - 1
       if (ir .eq. 1) then
          ra(l) = rra
```

```
            return
            endif
         endif
        i = l

        j = l + l
   20 continue
        if (j .le. ir) then
           if (j .lt. ir) then
            if (ra(j) .lt. ra(j+1)) j=j+1
            endif
         if (rra .lt. ra(j)) then
            ra(i) = ra(j)
            i = j
            j = j + j
            else
            j = ir + 1
            endif
         go to 20
         endif
        ra(i) = rra
        go to 10
        end
c
c sorts and outputs Monte Carlo results
c

        subroutine finish (pval,nn)
        dimension pval(10000),per(13)
        fnn=nn
        i01=fnn*.010001
        i05=fnn*.050001
        i10=fnn*.100001
        i20=fnn*.200001
        i30=fnn*.300001
        i40=fnn*.400001
        i50=fnn*.500001
        i60=fnn*.600001
        i70=fnn*.700001
        i80=fnn*.800001
        i90=fnn*.900001
        i95=fnn*.950001
        i99=fnn*.990001
        ss=0.0
        sum=0.0
        do 412 i=1,nn
        sum=sum+pval(i)
        ss=ss+pval(i)**2
  412 continue
        avg=sum/fnn
        var=(ss-sum**2/fnn)/(fnn-1.)
        stdev=sqrt(var)
        call sort (pval,nn)
        fmin=pval(1)
        fmax=pval(nn)
        per(1)=pval(i01)
        per(2)=pval(i05)
        per(3)=pval(i10)
```

```
      per(4)=pval(i20)
      per(5)=pval(i30)
      per(6)=pval(i40)
      per(7)=pval(i50)
      per(8)=pval(i60)
      per(9)=pval(i70)
      per(10)=pval(i80)
      per(11)=pval(i90)
      per(12)=pval(i95)
      per(13)=pval(i99)
      i51=i50+1
      fmedian=(pval(i50)+pval(i51))/2.
      write(22,41) nn
   41 format(/,'Monte Carlo results for ',i6,' samples:',/)
      write(22,42) fmin,fmax
   42 format(' minimum value = ',e12.6,/,' maximum value = ',e12.6)
      write(22,43) avg,var,stdev
   43 format(' mean = ',e12.6,/,' variance = ',e12.6,/,
     *' standard deviation = ',e12.6)
      write(22,44) fmedian,(per(k), k=1,13)
   44 format(' median = ',e12.6,/,/,'  1st percentile = ',e12.6,
     */,'  5th percentile = ',e12.6,/,' 10th percentile = ',e12.6,
     */,' 20th percentile = ',e12.6,/,' 30th percentile = ',e12.6,
     */,' 40th percentile = ',e12.6,/,' 50th percentile = ',e12.6,
     */,' 60th percentile = ',e12.6,/,' 70th percentile = ',e12.6,
     */,' 80th percentile = ',e12.6,/,' 90th percentile = ',e12.6,
     */,' 95th percentile = ',e12.6,/,' 99th percentile = ',e12.6)
      return
      end
```

Appendix C--EMPIRICAL

```
      program empirical
c
c
c forms empirical distribution functions for a given set
c of percentiles for multiple experts.
c
c uses simulation to combine weighted aggregations of these
c distributions according to a specified weighting function
c
c empirical cumulative distribution functions (for each expert)
c are sampled in the simulation using lines connecting the
c individual points of the distribution.  The more percentiles
c provided by the experts, the less influence this linear
c approximation has on the results.  A step function version
c of this code is available from the authors.
c
c inputs can be made directly from the terminal or through a
c file called emp.in (on unit 52).
c outputs are on a file called emp.out (on unit 59).
c
c emp.in file has the following lines and formats:
c  line 1 idim = no. of experts (distributions) - free format
c  line 2 iper = no. of percentiles for each distribution - free format
c     for each expert & DM do lines 3,4,5,6
c  line 3 pe array = estimates of the iper percentiles - free format
c  line 4 pl array = levels (e.g. 0.95) for the percentiles - free
      format
c  line 5 pmin = minimum value for the estimates - free format
c  line 6 pmax = maximum value for the estimates - free format
c     last 3 lines are:
c  line 7 nn = number of simulations (e.g. 1000) - free format
c  line 8 ifun = 1 for equal weights, = 2 for unequal - free format
c  line 9 wt array = weights for experts & DM - free format
c
c
      dimension pe(20,20), pl(20,20), fx(21), fy(21)
      dimension val(20), fxa(20,21), fya(20,21),plx(20,21)
      dimension pmin(20),pmax(20),pval(10000),per(13),wt(20)
      open (unit=59,file='emp.out',status='new')
      open (unit=52,file='emp.in',status='old')
      write(*,'("To use input file EMP.IN, type 1; else type 0: ",\)')
      read(*,*) ifile
      if(ifile.eq.1) go to 50

      write(*,'("Enter the sum of the experts and DM: ",\)')
      read(*,*) idim
      write( *,'("Enter the # of percentiles for an expert/DM: ",\)')
      read(*,*) iper
      do 10 i=1,idim
      write (*,11) iper, i
   11 format ('Enter the ',i3,' percentile estimates for expert ',i3)
      read(*,*) (pe(i,j), j=1,iper)
```

```
      write (*,12) iper, i
   12 format (' Enter the ',i3,' percentile levels ',
     *'(e.g. .95 for 95th percentile) for expert ',i3)
      read(*,*) (pl(i,j), j=1,iper)
      write(*,'("Enter the absolute minimum value possible: ",\)')
      read(*,*) pmin(i)
      write(*,'("Enter the absolute maximum value possible: ",\)')
      read(*,*) pmax(i)
   10 continue
      write( *,'("Enter the number of samples for the simulation ",\)')
      read(*,*) nn
      write(*,48)
   48 format(/,'Specify the aggregation function to be used:')
      write(*,'("Enter 1 for equal weights, else enter 2:  ",\)')
      read(*,*)ifun
      if(ifun.eq.2) then
      write(*,'("Enter the weights for experts & DM, including 0s: ")')
      do 8 i=1,idim
      write(*,7) i
    7 format('  weight for person ',i2,' = ',\)
      read(*,*) wt(i)
    8 continue
      endif
      sumw=0.0
      do 6 i=1,idim
    6 sumw=sumw+wt(i)
      epsilon=0.0001
      if(abs(sumw-1.0).gt.epsilon) then
      write(*,'("The weights will be normalized to 1.0")')
      do 5 i=1,idim
    5 wt(i)=wt(i)/sumw
      endif
      write(*,'("Output is on file EMP.OUT")')
      go to 40

   50 continue
      read(52,*) idim
      read(52,*) iper
      do 19 i=1,idim
      read(52,*) (pe(i,j), j=1,iper)
      read(52,*) (pl(i,j), j=1,iper)
      read(52,*) pmin(i)
      read(52,*) pmax(i)
   19 continue
      read(52,*) nn
      read(52,*) ifun
      read(52,*) (wt(i), i=1,idim)
      sumw=0.0
      epsilon = 0.0001
      do 4 i=1,idim
    4 sumw=sumw+wt(i)
      if(abs(sum-1.0).gt.epsilon) then
      write(59,'("The weights are normalized to 1.0")')
      do 3 i=1,idim
    3 wt(i)=wt(i)/sumw
      endif
   40 continue
      write(59,13) idim,iper,nn
```

```
   13 format ('Number of experts=',i5,/,'Number of percentiles=',i5,
     */,'Number of samples=',i5)
      if(ifun.eq.1) write(59,49)
   49 format('The aggregation function uses equal weights.')
      if(ifun.eq.2) then
      write(59,47)
   47  format('The aggregation function has weights as:')
      write(59,*) (wt(i), i=1,idim)
      endif
      do 14 i=1,idim
      write(59,16) i
   16 format(/,'Estimates for expert ',i3,':')
      write(59,15) pmin(i),pmax(i),(pl(i,j),j=1,iper)
   15 format(' min & max=',2f6.3,/,' levels ',5f7.4,/,'        ',6f7.4)
      write(59,9)(pe(i,j),j=1,iper)
   9  format(' estimates',5f7.4,/,'           ',6f7.4)
   14 continue
c
c
      do 20 i=1,idim
      call distmake (i,pe,pl,iper,pmin(i),pmax(i),fx,fy)
      iper1=iper+1
      do 30 j=1,iper1
      fxa(i,j)=fx(j)
      fya(i,j)=fy(j)
   30 continue
   20 continue
c
c  forms the nn samples of the product distribution
c
      call monte (fxa,fya,idim,iper,nn,pval,plx,pl,ifun,wt,pmax,pmin)
c
c  calculates stats for pval
c
      call calc(nn,pval,avg,fmedian,var,stdev,fmin,fmax,per)
c
c  print results
c
      write(59,41) nn
   41 format(/,'Monte Carlo results for ',i5,' samples:')
      write(59,42) fmin,fmax
   42 format(' minimum value = ',e12.6,/,' maximum value = ',e12.6)
      write(59,43) avg,var,stdev
   43 format(' mean = ',e12.6,/,' variance = ',e12.6,/,
     *' standard deviation = ',e12.6)
      write(59,44) fmedian,(per(k), k=1,13)
   44 format(' median = ',e12.6,/,'  1st percentile = ',e12.6,
     */,'  5th percentile = ',e12.6,/,' 10th percentile = ',e12.6,
     */,' 20th percentile = ',e12.6,/,' 30th percentile = ',e12.6,
     */,' 40th percentile = ',e12.6,/,' 50th percentile = ',e12.6,
     */,' 60th percentile = ',e12.6,/,' 70th percentile = ',e12.6,
     */,' 80th percentile = ',e12.6,/,' 90th percentile = ',e12.6,
     */,' 95th percentile = ',e12.6,/,' 99th percentile = ',e12.6)
      close (unit=59)
      close (unit=52)
      end
c
c  Forms empirical distribution function for each expert & DM: fx and
```

```
c  fy.
c
      subroutine distmake (ie,pe,pl,n,pmn,pmx,fx,fy)
      dimension pe(20,20), pl(20,20), fx(21), fy(21)
      ibin=n+1
      fy(1) = pl(ie,1)/(pe(ie,1)-pmn)
      fy(ibin) =(1.0-pl(ie,n))/(pmx-pe(ie,n))
      do 100 i=2,n
      j=i-1
      fy(i)=(pl(ie,i)-pl(ie,j))/(pe(ie,i)-pe(ie,j))
  100 continue
      do 110 i=1,n
C      fx(i)=pl(ie,i)
      fx(i)=pe(ie,i)
  110 continue
      fx(ibin)=1.
      write(59,113) ie
  113 format(/,'Empirical distribution; fx,fy for person:',i4)
      do 114 j=1,ibin
  114 write(59,*) fx(j),fy(j)
      return
      end
c
c  Performs the monte carlo simulation of the supplied functions
c
      subroutine monte (fxa,fya,idim,iper,nn,pval,plx,pl,ifun,wt,
     *pmax,pmin)
      dimension fxa(20,21),fya(20,21),val(20),plx(20,21),wt(20)
      dimension pval(10000),pl(20,20),pmax(20),pmin(20)
      n=iper+1
      do 223 i=1,idim
      do 222 j=1,iper
      plx(i,j)=pl(i,j)
  222 continue
      plx(i,n)=1.0
  253 format(' plx=',6e12.6)
  223 continue
      do 200 isamp=1,nn
      do 201 i=1,idim
      n=iper+1
      idum = 1
      val(i)=0.0
      if (isamp.eq.1.and.i.eq.1) idum = -iabs(487320587)
      t=ran3(idum)
      if(t.eq.0.0) val(i)=0.0
      if(t.le.plx(i,1).and.t.gt.0.0) then
        b1 = plx(i,1)/(fxa(i,1)-pmin(i))
        b0 = plx(i,1)-b1*fxa(i,1)
        val(i)=(t-b0)/b1
      endif
      if(t.gt.plx(i,iper)) then
        b1 = (1.0-plx(i,iper))/(pmax(i)-fxa(i,iper))
        b0 = 1.0-b1*pmax(i)
        val(i) = (t-b0)/b1
      endif
      do 202 k=2,iper
      km1=k-1
      if(t.le.plx(i,k).and.t.gt.plx(i,km1)) then
```

```
        b1 = (plx(i,k)-plx(i,km1))/(fxa(i,k)-fxa(i,km1))
        b0 = plx(i,k)-b1*fxa(i,k)
        val(i) = (t-b0)/b1
      endif
  202 continue
  213 format(' isamp',i4,' expert',i2,' rand#',e12.6,' value',e12.6)
  201 continue
      dim=idim
      if(ifun.eq.1) go to 275
      if(ifun.eq.2) go to 265
  265 continue
      sumy = 0.0
      do 266 i=1,idim
      sumy = sumy + val(i)*wt(i)
  266 continue
      pval(isamp)=sumy
      go to 270
  275 sumy=0.0
      do 276 i=1,idim
      fi=idim
      sumy=sumy+val(i)/fi
  276 continue
      pval(isamp)=sumy
      go to 270
  270 continue
  214 format('isamp',i4,' pval',e12.6)
  200 continue
      return
      end

      function ran3 (idum)
c
c         Returns a uniform random deviate between 0.0 and 1.0.
c         Set IDUM to any negative value to initialize or
c         reinitialize the sequence.
c
      parameter (mbig=1000000000,mseed=161803398,mz=0,fac=1./mbig)
c
c         According to Knuth, any large MBIG, and any smaller (but still
c         large) MSEED can be substituted for the above values.
c
      save     inext,     inextp,      ma
      dimension      ma(55)
      data           iff /0/
      if (idum.lt.0 .or. iff.eq.0) then
        iff = 1
        mj = mseed - iabs(idum)
        mj = mod(mj,mbig)
        ma(55) = mj
        mk = 1
        do 11 i = 1, 54
          ii = mod(21*i,55)
          ma(ii) = mk
          mk = mj - mk
          if (mk .lt. mz) mk = mk + mbig
          mj = ma(ii)
   11   continue
        do 13 k=1, 4
```

```
          do 12 i=1, 55
           ma(i) = ma(i) - ma(1+mod(i+30,55))
           if (ma(i) .lt. mz) ma(i) = ma(i) + mbig
   12     continue
   13     continue
        inext = 0
        inextp = 31
        idum = 1
      endif
      inext = inext + 1
      if (inext .eq. 56) inext = 1
      inextp = inextp + 1
      if (inextp .eq. 56) inextp = 1
      mj = ma(inext) - ma(inextp)
      if (mj .lt. mz) mj = mj + mbig
      ma(inext) = mj
      ran3 = mj*fac
      return
      end
c
c
c
      subroutine calc(nn,pval,avg,fmedian,var,stdev,fmin,fmax,per)
      dimension pval(10000),per(13)
      fnn=nn
      i01=fnn*.010001
      i05=fnn*.050001
      i10=fnn*.100001
      i20=fnn*.200001
      i30=fnn*.300001
      i40=fnn*.400001
      i50=fnn*.500001
      i60=fnn*.600001
      i70=fnn*.700001
      i80=fnn*.800001
      i90=fnn*.900001
      i95=fnn*.950001
      i99=fnn*.990001
  401 format(' 1,5,10,50,90,95,99',7i4)
      scale=0.0
      ss=0.0
      sum=0.0
      do 400 i=1,nn
      scale=scale+pval(i)
  400 continue
      scale=1.0
      do 411 i=1,nn
      pval(i)=pval(i)/scale
  411 continue
  422 format('scale',e12.6)
      do 412 i=1,nn
      sum=sum+pval(i)
      ss=ss+pval(i)**2
  433 format ('pval =' ,e12.6)
  412 continue
      avg=sum/fnn
      var=(ss-sum**2/fnn)/(fnn-1.)
      stdev=sqrt(var)
```

```
      call sort (nn,pval)
      fmin=pval(1)
      fmax=pval(nn)
      per(1)=pval(i01)
      per(2)=pval(i05)
      per(3)=pval(i10)
      per(4)=pval(i20)
      per(5)=pval(i30)
      per(6)=pval(i40)
      per(7)=pval(i50)
      per(8)=pval(i60)
      per(9)=pval(i70)
      per(10)=pval(i80)
      per(11)=pval(i90)
      per(12)=pval(i95)
      per(13)=pval(i99)
      i51=i50+1
      fmedian=(pval(i50)+pval(i51))/2.
      return
      end


      subroutine sort (n,ra)
c
c         Sorts an array RA of length N into ascending numerial order
c         using the Heapsort algorithm.  N is input;  RA is replaced
c         by its sorted rearrangement.
c
      dimension ra(n)
      l = n/2 + 1
      ir = n
   10 continue
      if (l .gt. 1) then
         l = l - 1
         rra = ra(l)
         else
         rra = ra(ir)
         ra(ir) = ra(1)
         ir = ir - 1
         if (ir .eq. 1) then
          ra(1) = rra
          return
          endif
      endif
      i = l
      j = l + l
   20 continue
      if (j .le. ir) then
         if (j .lt. ir) then
          if (ra(j) .lt. ra(j+1)) j=j+1
         endif
       if (rra .lt. ra(j)) then
          ra(i) = ra(j)
          i = j
          j = j + j
          else
          j = ir + 1
          endif
```

```
      go to 20
      endif
     ra(i) = rra
     go to 10
     end
```

Appendix D--BOOT

```
      program boot
c
c  Constructs bootstrap samples from the original sample of size n.
c
c  Each sample is randomly formed on its own.
c
c  Input is done on the terminal or file BOOT.IN
c  BOOT.IN has the following lines in free format:
c     line 1 iabb = 0 if all simulated values are printed, =1 if not
c     line 2 irange = 1 if the sample of n experts each supplies two
c                     ranges values (3 values per person), = 0 if not **
c     line 3 n = number of values supplied = number of persons or
c                                            3 times that
c     line 4 m = number of simulations
c     line 5 x array = the n values supplied
c                      (a space between each value)
c
c  Output is sent to file BOOT.OUT
c
c  ** if each expert provides a best estimate and an upper & lower range
c     value, then irange = 1 and there are 3 times as many values as
c     experts.  for this case, the sample size used in the program
c     is changed to n/3 = number of experts, not the number of values.
c
c
      character title*75
      dimension x(65),xsamp(65),xbig(20000),xnew(65)
      dimension xmed(1000)
      open (unit=11,file='boot.in',status='old')
      open (unit=22,file='boot.out',status='new')
      write(*,'("to use input file boot.in, type 1 else type 0: ",\)')
      read(*,*) infile
      if(infile.eq.1) go to 61

      write(*,31)
   31 format('to print all simulated values, type 0; else type 1: ',\)
      read(*,*) iabb
      write(*,32)
   32 format('if range values are included, type 1; else type 0: ',\)
      read(*,*) irange
      if(irange.eq.1) write(22,33)
      if(irange.eq.0) write(22,34)
   33 format(' irange=1, ranges with sample values are assumed')
   34 format(' no range values included')
      write(*,'("enter the sample size: ",\)')
      read(*,*) n
      write(*,'("enter the number of simulations (e.g.1000): ",\)')
      read(*,*) m
      fn=n
      write(*,39)
   39 format('enter the sample values with a space between each:')
      read(*,*) (x(i),i=1,n)
      go to 65
```

```
   61 continue
      read(11,*) iabb
      read(11,*) irange
      if(irange.eq.1) write(22,63)
      if(irange.eq.0) write(22,64)
   63 format(' irange=1, ranges with sample values are assumed')
   64 format(' no range values included')
      read(11,*) n
      read(11,*) m
      fn=n
      read(11,*) (x(i),i=1,n)

   65 continue
      write(22,66) m
   66 format(' the number of simulations is ',i5)
      write(22,37) n, (x(i),i=1,n)
   37 format(' the ',i2,' sample values are: ',/,10e9.2)
c
c  Calculate simple statistics for the sample only
c
      slog=0.0
      sum=0.0
      ss=0.0
      do 51 i=1,n
      sum=sum+x(i)
      if(x(i).eq.0.0) go to 51
      slog=slog+alog10(x(i))
      ss=ss+x(i)**2
   51 continue
      xbar=sum/fn
      xgbar=10.0**(slog/fn)
      xvar=(ss-sum**2/fn)/(fn-1.)
      xstd=sqrt(xvar)
      do 52 i=1,n
      xnew(i)=x(i)
   52 continue
      call sort(xnew,n)
      k=n/2
      k1=k+1
      kind=mod(n,2)
      if(kind.eq.0) xtild=(xnew(k)+xnew(k1))/2.0
      if(kind.eq.1) xtild=xnew(k1)
      write(22,53)xbar,xtild,xgbar,xvar,xstd,xnew(1),xnew(n)
   53 format(/,'simple statistics for the original sample:',/,
     *' sample mean = ',e9.2,/,' sample median = ',e9.2,/,
     *' sample geometric mean = ',e9.2,/,' sample variance = ',
     *e9.2,/,' sample standard deviation = ',e9.2,/,
     *' sample minimum = ',e9.2,/,' sample maximum = ',e9.2,/)
c
c   Begin bootstrap sampling of x
c   All bootstrap samples are randomly formed and stored in xbig
c
      do 50 i=1,m
      do 50 j=1,n
      k=(i-1)*n+j
      idum=1
      if(i.eq.1.and.j.eq.1) idum = -iabs(5739784770)
   12 t1=rand(idum)
```

```
      temp=t1*fn
      do 13 kk=1,n
      kk1=kk-1
      if(temp.eq.0.0) xbig(k)=x(1)
      if(temp.le.kk.and.temp.gt.kk1) xbig(k)=x(kk)
      if(temp.gt.n) go to 12
   13 continue
   50 continue
      if(irange.eq.1) n=fn/3.0
c
c  Loop for the 3 estimators: median, mean, geomean
c
      iwrite=0
      do 100 ii=1,3
      do 40 i=1,m
      do 41 j=1,n
      k=(i-1)*n+j
      xsamp(j)=xbig(k)
   41 continue
      call sort(xsamp,n)
c
c  Median calculations
c
      if (ii.eq.1) then
        k=n/2
        k1=k+1
        kind=mod(n,2)
        if(kind.eq.0) xmed(i)=(xsamp(k)+xsamp(k1))/2.0
        if(kind.eq.1) xmed(i)=xsamp(k1)
      endif
c
c  Mean calculations
c
      if (ii.eq.2)then
        xmean=0.0
        do 101 ij=1,n
  101   xmean=xmean+xsamp(ij)/float(n)
        xmed(i)=xmean
      endif
c
c  Geometric mean calculations
c
      if (ii.eq.3)then
        xgeo=0.0
        do 102 ij=1,n
        if(xsamp(ij).eq.0.0) go to 102
        xgeo=xgeo+alog10(xsamp(ij))
  102   continue
        if(xgeo.eq.0.0) then
         iwrite=iwrite+1
         xmed(i)=0.0
         go to 107
        endif
        xmed(i)=10.0**(xgeo/float(n))
  107   continue
      endif
   40 continue
      if(iwrite.gt.0.0) write(22,104)iwrite
```

```
  104 format(/,/,'warning: ',i4,' samples had no geometric mean,',
     *' results may be biased low')
c
c  End sampling, sort resulting estimators
c
      call sort(xmed,m)
c
c  Calculate percentiles, estimators, variances
c   Print results
c
      if(ii.eq.1) write(22,27)
      if(ii.eq.2) write(22,28)
      if(ii.eq.3) write(22,29)
   27 format(' ',//,'medians')
   28 format(' ',//,'means')
   29 format(' ',//,'geomeans')
      if(iabb.ne.1) write(22,43)
   43 format('the sample estimator values from the simulation')
      if(iabb.ne.1) write(22,21) (xmed(i),i=1,m)
   21 format(' ', 6f10.6)
      smed=0.0
      ssmed=0.0
      slmed=0.0
      do 45 i=1,m
      smed=smed+xmed(i)
      ssmed=ssmed+xmed(i)**2
      if(xmed(i).eq.0.0) go to 45
      slmed=slmed+alog10(xmed(i))
   45 continue
      amed=smed/float(m)
      vmed=(ssmed-(smed**2)/float(m))/float(m-1)
      sdmed=sqrt(vmed)
      gmed=10.0**(slmed/float(m))
      delta=float(m)/100.0
      m99=99.0*delta
      m95=95.0*delta
      m90=90.0*delta
      m50=50.0*delta
      m20=20.0*delta
      m30=30.0*delta
      m40=40.0*delta
      m60=60.0*delta
      m70=70.0*delta
      m80=80.0*delta
      m1=1.0*delta
      m5=5.0*delta
      m10=10.0*delta
      write(22,23) xmed(1),xmed(m)
   23 format(' minimum value = ',f10.6,/,' maximum value = ',f10.6)
      write(22,26) xmed(m1),xmed(m5),xmed(m10),xmed(m20),xmed(m30),
     *xmed(m40),xmed(m50),xmed(m60),xmed(m70),xmed(m80),xmed(m90),
     *xmed(m95),xmed(m99)
   26 format(' percentiles: ',/,11x,'  1:',f10.6,/,11x,'  5:',f10.6,/,
     * 11x,' 10:',f10.6,/,11x,' 20:',f10.6,/,
     * 11x,' 30:',f10.6,/,11x,' 40:',f10.6,/,
     * 11x,' 50:',f10.6,/,11x,' 60:',f10.6,/,
     * 11x,' 70:',f10.6,/,11x,' 80:',f10.6,/,
     * 11x,' 90:',f10.6,/,11x,' 95:',f10.6,/, 11x,' 99:',f10.6)
```

```
      write(22,22) amed,vmed,sdmed,gmed
   22 format(' mean = ',f10.6,/,' variance',f10.6,/,
     *' standard deviation = ',f10.6,/,
     *' geometric mean = ',f10.6)
  100 continue
      stop
      end

      subroutine sort (ra,n)
c
c   Sorts an array ra of length n into ascending numerial order
c   using the heapsort algorithm.  n is input;  ra is replaced
c   by its sorted rearrangement.
c
      dimension ra(n)
      l = n/2 + 1
      ir = n
   10 continue
      if (l .gt. 1) then
         l = l - 1
       rra = ra(l)
       else
       rra = ra(ir)
       ra(ir) = ra(1)
       ir = ir - 1
       if (ir .eq. 1) then
          ra(l) = rra
          return
       endif
      endif
      i = l
      j = l + l
   20 continue
      if (j .le. ir) then
         if (j .lt. ir) then
          if (ra(j) .lt. ra(j+1)) j=j+1
         endif
         if (rra .lt. ra(j)) then
          ra(i) = ra(j)
          i = j
          j = j + j
         else
          j = ir + 1
         endif
      go to 20
      endif
      ra(i) = rra
      go to 10
      end


      function rand (idum)
c
c   Returns a uniform random deviate between 0.0 and 1.0.
c   set idum to any negative value to initialize or
c   reinitialize the sequence.
c
      parameter (mbig=1000000000,mseed=161803398,mz=0,fac=1./mbig)
```

```
c
c   According to Knuth, any large mbig, and any smaller (but still
c   large) mseed can be substituted for the above values.
c
      save     inext,     inextp,      ma
      dimension      ma(55)
      data           iff /0/
      if (idum.lt.0 .or. iff.eq.0) then
         iff = 1
        mj = mseed - iabs(idum)
        mj = mod(mj,mbig)
        ma(55) = mj
        mk = 1
        do 11 i = 1, 54
          ii = mod(21*i,55)
          ma(ii) = mk
          mk = mj - mk
          if (mk .lt. mz) mk = mk + mbig
          mj = ma(ii)
   11     continue
         do 13 k=1, 4
           do 12 i=1, 55
              ma(i) = ma(i) - ma(1+mod(i+30,55))
              if (ma(i) .lt. mz) ma(i) = ma(i) + mbig
   12      continue
   13    continue
         inext = 0
         inextp = 31
         idum = 1
      endif
      inext = inext + 1
      if (inext .eq. 56) inext = 1
      inextp = inextp + 1
      if (inextp .eq. 56) inextp = 1
      mj = ma(inext) - ma(inextp)
      if (mj .lt. mz) mj = mj + mbig
      ma(inext) = mj
      rand = mj*fac
      return
      end
```

Glossary of Expert Judgment Terms

ADVISORY EXPERT: The in-house employee or consultant who is considered expert in the subject matter and who assists the project personnel in developing the questions that will be later asked of the external experts.

AGGREGATION: See Behavioral or Mathematical Aggregation.

ALPHA LEVEL: See Significance Level.

ANCHORING BIAS: The indvidual's failure to adjust sufficiently from his first impression in solving a problem. Sometimes this bias is explained in terms of Bayes Theorem as the failure to adjust a judgment in light of new information as much as it would be adjusted in terms of Bayes mathematical formula.

ANCHORING AND ADJUSTMENT HEURISTIC: This effect occurs when an individual reaches a final answer by starting from an initial value and adjusting from it. The initial value can be supplied with the question, or it can be reached by the expert through his impressions or computations. Usually, the use of this heuristic skews the answer toward the initial value.

ANCILLARY DATA OR INFORMATION: Any information or data gathered as part of the elicitation that is not the expert's answer. For example, information on the expert's background and problem-solving processes (expert data) is ancillary information. This information has the potential for being related to the the answers. Thus ancillary information can form conditional variables.

ANALYSIS OF VARIANCE: A statistical technique for testing the equivalence or lack of equivalence of mean values from several different groups or classes of data. The test is done by comparing variation between the groups to the variation within the groups. The within-group variation represents the background-, error-, or noise-level variation. Groups are determined prior to the study so that the test can be made from a minimal number of measurements. Such efficient planning is called experimental design.

ANALYST: Member of the project personnel who analyzes the expert data and judgment. The analyst may have other roles, such as interviewer.

ANSWER-ONLY DOCUMENTATION: A written record of only the experts' answers.

ANSWER PLUS PROBLEM-SOLVING DOCUMENTATION: A written record of the experts' answers and how they arrived at these answers.

ANSWERS OR EXPERT ANSWERS: The expert's final response to a technical question. This term includes responses given in quantitative (estimates) or qualitative form (solution).

AVAILABILITY BIAS: Differing ease with which events can be retrieved from long-term memory. Data involving catastrophic, familiar, concrete, or

recent events may be easier to recall. Availability bias affects people's ability to accurately estimate frequencies and to recall other aspects of the event.

BACKGROUND INFORMATION: Information that the expert needs to interpret the question or problem. Background information includes the sequence of events leading up to the event in question, pictorial representations of the question (e.g., flow charts), and decompositions of the question.

BAYESIAN METHODS/APPROACH: A technique for combining information of various types or from various sources. The combining calculations are based on Bayes Theorem, which defines the probability distribution function of the data as conditioned on its parameters. These parameters are also assumed to have a probability distribution called a prior distribution. Combining the data and prior distributions produces a posterior distribution. The expected valued of this posterior distribution is the desired final estimate. The philosophy of this method is to use and combine all available information to form the final estimate rather than to rely only on the data from a single study or experiment.

BEHAVIORAL AGGREGATION: A means of obtaining one answer from multiple experts through the use of behavioral techniques that encourage consensus. For example, group-think bias can be fostered to create pressures toward unanimity.

BIAS: Bias can be defined as occurring when (1) expressions of the expert's thinking do not match his actual thinking at the time of the elicitation, and (2) the expert's estimates do not follow normative statistical or logical rules. An example of the first would be if the expert judged a particular event to be extremely rare but had to select from response options that did not extend as far as his judgment. An example of the second would be if the expert claimed that A was better than B in some respect, B better than C, and C better than A. Sources of bias can be a person's needs (*motivational bias*) or thought processes (*cognitive bias*).

BOOTSTRAP: A data-based simulation technique useful for finding estimates and distributions of estimates when statistical distribution theory is not applicable. This technique is based on forming multiple (1000) random samples with replacement from the original sample data, calculating the estimator of interest in each sample, and forming the distribution of that estimator from these calculated 1000 values.

CLIENT: The person who has requested the gathering of expert judgment. That is, the client is the one whose needs the project will serve. Often the client is the person funding the study. Whether the client is the funder or not, the client can usually say what the purpose and goals of the project are , what information is needed from the experts, and what resources will be available for the project.

CLUSTER ANALYSIS: Techniques for identifying how values or a variable is grouped numerically or how variables themselves are grouped according to shared information. For example, the data set of values (0.2, 0.25, 0.30,

0.77, 0.80, 0.95) would cluster into two groups. Various grouping methods are used to determine group size and membership. Most are based on some measure of distance between values and groups of values.

COGNITION: The mental activity, the processing of information, that humans do when they solve problems.

COGNITIVE BIAS: Biases whose source is the limitations of the human mind. Anchoring bias is one example.

COGNITIVE DISSONANCE: Cognitive dissonance occurs when an individual finds a discrepancy either between his beliefs or between his beliefs and his actions (Festinger 1957). For example, an individual may find he holds an opinion that conflicts with that of the other group members, and if he has a high opinion of the intelligence of the group, he may resolve the discrepancy by unconsciously changing his judgment to be in agreement with that of the group.

CONDITIONAL VARIABLE: An ancillary quantity (variable) found to be influential or important for determining the answers given by the experts. This determination is made using more than one analysis technique.

CONDITIONALITY: The description of the phenomenon where one variable has an influence on or a significant relationship to another variable. Conditionality also includes the case where several variables are influential over several other variables.

CORRELATION: See Dependence.

CUMULATIVE DISTRIBUTION FUNCTION (CDF): The function resulting from the accumulation of area under the probability distribution function (pdf), in other words, the integral of the pdf. The function value, $F(x)$, is the probability that the random variable takes on values less than or equal to the value at x.

DATA GATHERER: See Interviewer or Knowledge Engineer.

DECISION ANALYSIS: Structured means for conceptualizing and resolving complex problems. This structure involves breaking the problem into parts to make it more tractable. Often the decision structure includes the initial options or acts, the possible consequences of these values, and uncertainty measures. A variety of structural forms are used, ranging from event trees to hierarchies of factors. Decision analysis is applied mainly to business problems such as plant siting, procurement, and portfolio analysis. It has also been used in crisis management, international negotiations, intelligence analyses, and labor-management negotiations (Peaslee 1981).

DECOMPOSITION: The breaking of a problem or question into its component parts to make it easier to solve. This technique has been shown to increase accuracy.

DELPHI METHOD: An elicitation method developed by the Rand Corporation to limit the biasing effects of interaction. In a true Delphi, the experts do not interact with one another and only interact with the moderator in a limited way. The experts, in isolation from one another, give their judgments and, perhaps, some of their reasons for making these judgments. The moderator collects these judgments, makes the judgments anonymous, distributes

these judgments to the individual experts, and allows each of them to revise their previous judgments. This process can be repeated for as many times as desired, such as until consensus is achieved.

DEPENDENCE: Expert's estimates are conditioned on some factor and affected by this conditioning. In this handbook, *dependence* is used interchangeably with *correlation.* Dependence can refer to the data from different experts (between-expert dependence) or it can refer to dependencies of estimates given by the same expert (within-expert dependence).

DEPENDENT VARIABLE: The quantity of interest that contains the expert answers. These answers are usually conditioned and therefore dependent on other ancillary variables or information.

DESIGN OF ELICITATION METHOD: Planning of the method in terms of (1) the project's constraints--e.g., time, budget and personnel; (2) goals--e.g., for obtaining particular data; and (3) additional considerations--e.g., the logistics and cost of meeting together versus separately, the structuring of the elicitations, the treatment of bias, the presentation of the problems, and the documenting of the elicitation). The reason for designing the elicitation is to create the optimal combination of techniques for a particular situation. Different techniques possess differing advantages and abilities to control for particular factors, such as those which would introduce bias. One of the main techniques used in designing the data gathering is *structuring*, or placing controls on the elicitation (see Structuring). An example of structuring applied to interactive groups is to have the natural leader present his or her views last, so as to prevent the follow-the-leader effect (see Group Think Bias).

DETAILED VERBATIM DOCUMENTATION: See Verbatim Documentation.

DETAILED STRUCTURED DOCUMENTATION: See Structured Documentation.

DISAGGREGATION: See Decomposition.

DISCRIMINANT ANALYSIS: A multivariate statistical technique for determining how well a chosen variable discriminates or classifies each datum from a data set into specified groups.

DISPERSION MEASURE: An estimate of how much spread or dispersion is in the sample data. Dispersion measures may be in the form of percentiles, variances, ranges, or error bars.

DISTRIBUTION: See Probability Distribution Function.

ELICITATION: Process of gathering expert judgment in a specially designed manner. (See Delphi, Individual Interview, and Interactive Group.)

ESTIMATE: The expert's answer encoded in the response mode. Estimates specifically refers to answers given in numerical form, such as probabilities or ratings.

ESTIMATOR: The formula or procedure for calculating or determining the value of a property of a population such as the median or the parameter of a distribution such as the mean.

ETHNOGRAPHIC TECHNIQUE: An interviewing technique from cultural anthropology that involves restating the subject's words into questions. This method avoids the danger of having the interviewer bias the subject's account by using the subject's own words. This technique is used to pursue in greater depth information that the subject has mentioned.

EXPERT: A person who has background in the subject matter at the desired level of detail and who is recognized by his peers or those conducting the study as being qualified to answer questions. The expert is sometimes referred to as the *external expert.*

EXPERT DATA: See Problem-Solving Data.

EXPERT JUDGMENT: Judgments of those with expertise or knowledge in the area. Expert judgment is usually elicited when experimental data is sparse or lacking. In this book, expert judgment refers to a combination of the expert's answer, his data on how this answer was reached (e.g., definitions, assumptions, and algorithms), and ancillary information on the expert himself (e.g., educational background and work experience).

FACTOR ANALYSIS: A multivariate statistical technique for determining how a set of variables share common information. The original variables are transformed to a set of new variables called factors. Some factors, called common factors, are formed from shared information of more than one of the original variables. Other factors, called unique factors, are formed from information from only one variable.

GENERAL LINEAR MODEL, GLM: See Regression.

GRANULARITY: Level of generality used in gathering, examining, or analyzing data. For example, data on expert's problem solving could be viewed at a coarse granularity, such as the of type of heuristic used, or at a fine granularity, such as the actual calculations performed as a part of each heuristic.

GROUP THINK BIAS: The tendency to modify a judgment so that it is in agreement with that of the group or of the group leader. Generally, the individual is unaware that he has modified his judgment to be in agreement. This bias is classified as a motivational bias because it stems from the human need to be accepted and respected by others. Individuals are more prone to group think if they have a strong desire to remain a member, if they are satisfied with the group, if the group is cohesive, and if they are not a natural leader in the group.

HEURISTIC: A short cut used to reduce the mental effort of solving a complex problem. A common heuristic is that of anchoring and adjusting. Instead of doing many detailed calculations, the individual adjusts in small increments from his initial impression of the answer.

HUMAN RELIABILITY ANALYSIS: This analysis "models events that are primarily due to human actions or inactions, often called errors, analyzes their effects, and quantifies their impact" (Doughterty et al., 1986: 3-2). Human reliability analysis is often a component of risk analysis because human events can contribute to the initiation, prevention, or mitigation of damage states.

IMPRESSION MANAGEMENT BIAS: A type of social pressure bias that occurs when the subject is responding to the reactions of those not physically present. For example, on a survey the subject tries to answer in such a way as to bring the most approbation (e.g., from society in the abstract or from the question writer in particular).

INCONSISTENCY BIAS: Inability to maintain the same problem-solving heuristic, definitions, or assumptions through time because of the limited information-processing capacity of the human mind.

INDEPENDENT VARIABLE: A quantity that is fixed, measured, recorded, or determined before or during the study. This quantity is possibly related to the answers given. See also Ancillary Information and Conditional Variables.

INDIVIDUAL INTERVIEW: One of the three basic methods of elicitation. One individual is interviewed at a time, usually in a face-to-face situation. The interviewer can structure the elicitation to any degree. An unstructured interview resembles a conversation; a structured one an interview driven by prepared questions. Often the separate responses are mathematically combined in some way, hence its other names *staticized* or *nominal group.*

INFERENCE (GENERAL): The process of drawing conclusions from information for interpretation on a general or universal scale.

INFERENCE (STATISTICAL): The process of drawing conclusions about the population of interest or study from the results of statistically valid sampling and analysis.

INTENSIVE PILOT TEST: A type of pilot testing that combines structured interviews and observations. The intensive pilot test provides two kinds of feedback: (1) how the expert progresses through the elicitation, his general impressions, when and why he decides to respond to particular questions; and (2) how the expert specifically interprets each direction, statement of the question, or response mode option.

INTERACTIVE GROUP METHOD: One of the three basic methods of elicitation. In the interactive group method, the participants are in a face-to-face situation with one another and a session moderator. The participants' interactions with one another can be structured to any degree. A totally unstructured group resembles a typical meeting; a highly structured group is carefully choreographed as to when the participants present their views and when there is open discussion.

INTERVIEWER: The person who elicits the expert judgment. (See also Knowledge Engineer.)

KNOWLEDGE ACQUISITION: Part of artificial intelligence connected with "extracting, structuring, and organizing knowledge from some source, usually human experts, so it can be used in a program." (Waterman 1986:392).

KNOWLEDGE-BASED COGNITION OR BEHAVIOR: The level of thinking that most of expert judgment involves. It is interpretive, analytical, high-level, conscious activity caused by thinking about rare or uncertain phenomena.

KNOWLEDGE ENGINEER: The knowleedge engineer is similar to the interviewer in that both elicit information from the expert. However, the term *knowledge engineer* refers to someone who, in addition to interviewing, represents and enters the expert knowledge into a computer system with the goal of creating a knowledge-based system.

LIMITED INTERACTIVE GROUP: See Delphi Method.

LIMITED PILOT TEST: A type of pilot testing done with a very small sample of experts. The limited pilot test is done after the intensive pilot test to provide a time estimate of the duration of each part of the elicitation.

LINEAR SCALE: A continuous line of numbers on a scale that is linear in structure; that is, each number on the line is separated from its neighbor by a value equal to the difference between it and its neighbor. For example, 4.0 would be twice as far from 0.0 as 2.0.

LONG-TERM MEMORY (LTM): Memory of large capacity and relatively permanent duration. A portion of what is processed in the individual's short-term memory is stored in this type of memory.

MATHEMATICAL AGGREGATION: The use of mathematical means to combine multiple expert's answers into one answer, usually when a single estimate is needed. Multiple expert's answers or distributions can also be combined into a single distribution. Some mathematical methods weight the expert's answers equally, such as the mean; others use more complex weighting schemes.

MEAN: The numeric average of a set of values (sample) calculated by summing the values and dividing by the value of how many there are in the set. The mean of a population is the theoretically derived expected value.

MEDIAN: The middle value of a sample or a distribution. The median is the 50th percentile of a distribution. It is the value such that half of the sample or the distribution is larger, and half of the sample or distribution is smaller. The median is a measurement of the center of the sample or distribution.

MILLER'S NUMBER: The number of things that the average person can mentally juggle--7 plus or minus 2.

MISINTERPRETATION: The altering of the expert's thoughts as a result of the methods of elicitation and documentation. See Training Bias.

MISREPRESENTATION: The altering of the expert's thoughts and answers as a result of modeling or analyzing them. See Tool Bias.

MODE: The most frequent value in a data set (sample) or distribution. The mode is the hump of the distribution. A distribution or data set with more than one hump is bimodal (for 2), trimodal (for 3), or multimodal (more than 2).

MODES OF COMMUNICATION: Communicating with the expert in person, by mail (electronic or postal), or by telephone. Each mode has its advantages and disadvantages. For example, the mail mode is the least expensive but the most time consuming).

MONITORING ELICITATION: Real-time observations of the elicitation process, usually for detecting bias.

MONTE CARLO SIMULATION: A computational technique for investigating properties and behavior of a variable by repeated random sampling from a known or assumed distribution (e.g., normal) representing the variable.

MOTIVATIONAL BIAS: Biases that have as their source, the emotional needs and wishes of the subject. Group think bias is one example.

MULTIVARIATE ANALYSIS: A statistical analysis technique that allows two or more variables of interest to be considered simultaneously.

NOMINAL GROUP METHOD: See Individual Interview.

NONPARAMETRIC TECHNIQUE: Statistical analysis techniques that does not require assuming that the sample or population follow a particular distribution, such as the normal distribution. These are sometimes referred to as distribution-free techniques.

NORMAL DISTRIBUTION: A particular probability distribution function that is symmetrically shaped about the mean, has identical values for mean, median, and mode, and has a bell-like shape. It is also known as the Gaussian distribution or a bell-shaped curve.

NORMATIVE EXPERTISE: Expertise in the statistical or mathematical principles of the response mode.

OUTLIERS: Extreme-valued observations in a sample or data set. It is unlikely (not probable) that these observations belong to the same distribution as the rest of the sample.

PAIRWISE COMPARISONS: Establishing the relative ratings of a set of objects, events, or criteria by comparing them two at time. If there is a set of n things, then in order to make all possible pairwise comparisons, $n(n-1)/2$ comparisons are required. Comparing object A to object B is the reciprocal of comparing object B to object A.

PERCENTILE: The value from a distribution of a random variable that divides the area under the distribution curve into the specified percentages. For example, the 5th percentile is the value of the distribution such that 5% of the distribution is smaller and 95% of the distribution is larger.

PILOT TESTING: A type of practice involving taking a sample of the larger expert population, presenting these experts with an aspect of the elicitation, obtaining their feedback, and revising the elicitation accordingly. (See also Intensive and Limited Pilot Testing.)

POPULATION: The entire existing or theoretical set of items under study. Examples are (1) all the people on the earth--past, present and future; (2) all earthquakes on the earth; (3) all earthquakes in a particular location on the earth; and (4) failures and successes of all components of a certain type.

POSTERIOR DISTRIBUTION/DENSITY FUNCTION: The resulting distribution or density function from a Bayesian analysis. This distribution is the combination of the prior information and the data.

PRACTICING THE ELICITATION: Rehearsing the elicitation to detect any problems in its design before its use. One type of practice is pilot testing. (See Pilot Testing.)

PRIOR DISTRIBUTION/DENSITY FUNCTION: The distribution or density for the parameters of interest. This represents the information known prior

to the gathering of the data. It is combined with the data to form the posterior.

PROBABILITY: Refers to the chance of something occurring. One important property of probabilities is that they are values from 0.0 to 1.0. A probability equal to 0.0 means that the event never happens. A probability equal to 1.0 means that the event always happens. A probability equal to 0.5 means that the event happens half of the time. Another important property is that the probabilities of all exhaustive (all events in a set), mutually exclusive events (nonoverlapping events) must sum to the value 1.0.

PROBABILITY DISTRIBUTION FUNCTION (PDF) OR PROBABILITY DISTRIBUTION: The mapping of a random variable onto a functional representation of probabilities for the possible values of that random variable. The nonnegative function, f, is mathematically represented by an equation in terms of the values of the random variable, x. The area under the entire curve resulting from this equation is 1.0. Sectional areas under the curve correspond to the probabilities of the corresponding x values. The values of the equation, f, are probabilities per unit interval, dx.

PROBLEM-SOLVING DATA: information relating to the expert's solution of the problem such as his definitions, assumptions, or algorithms.

PROBLEM-SOLVING ELICITATION: Techniques used to elicit how the subject solved the problem. These elicitation techniques can be used to deliver data of differing levels of detail. For example, in a risk analysis study, a few sentences on the expert's reasoning might be all that is needed. For an artificial intelligence project, more detailed information might be needed to model each step of the expert's thinking.

PROBLEM-SOLVING METHOD OR PROCESS: The means by which the expert solves the problem. These means could include the expert's interpretation of the problem, assumptions, definitions, and algorithms.

PUTATIVE INTERVAL: An interval estimate calculated from a simulated or computed distribution, usually from a bootstrap or Monte Carlo simulation. For example, the 5th and 95th percentiles of the simulated distribution of the median for a sample would form the central 90% putative interval for that distribution.

QUALITATIVE DATA OR INFORMATION: Any nonnumeric data such as verbal descriptions, classifications, categories, or preferences.

QUANTIFICATION: The process of transforming qualitative information into quantitative or numerical forms. In addition, quantification is used here to refer to the process of transforming raw data, in any original form, to a desired numerical form. This transformation can change the granularity of the data from coarse to fine.

QUANTITATIVE DATA OR INFORMATION: Any numerical data such as integers, ranks, or values on the real number line.

QUESTION: The concrete, detailed points within the question areas to which the expert is asked to respond. Questions are also referred to as *problems*. An example of a question or problem is "what is the leak rate in gallons per

minute as a function of time to seal failure due to loss of cooling to the pump shaft?" The test for whether a query qualifies as a question is whether the expert finds it sufficiently specific to be answered.

QUESTION AREA: The specific issue for investigation or the general area in which the experts will be questioned. Question areas are developed by considering such information as the goal of the project and the client's directives. For example, for an application whose goal was the provision of likelihoods and consequences of severe accidents in light-water reactors, separate question areas exist for front- and back-end phenomena, the different plant manufacturers, and so on.

QUESTION PHRASING: The wording of the question and the response mode done to maximize the chances that the expert understands it and is not unduly influenced by the wording. Payne (1951) has shown that different word choices and orderings can change the answer reached by 4 to 15%.

RANDOM VARIABLE: The quantity of interest that can take on any of a set of possible values or outcomes of an experiment or observation. The symbol for a random variable is X; the symbol for a generic value of a random variable is x.

RANKS: A set of numeric or descriptive values assigned to an original set of values or descriptions. Ranks are usually cardinal (i.e., composed of integer values, in ascending or descending order, and equally spaced 1, 2, 3, etc.). Ranks can also be ordinal, or descriptive in nature (worst, better, best). Ratings are usually assigned numbers from a chosen scale (e.g. from 1 to 10). The numbers are assigned by the user or the analyst according to some criteria.

RATINGS: See Ranks.

REGRESSION: The analysis that finds the best fit line for a dependent variable, y, in terms of the independent variables, x_i. The form of the model is $y = b_0 + b_1x_1 + b_2x_2 + ... + \varepsilon$ where the b_0 is the intercept term, the other bs represent slopes, and ε is the residual or remaining error not accounted for in the model. The regression line is fit such that the squared distances between the data points and the line are minimized. Regression is a subset of the analyses known as general linear models (GLMs) where linear relationships are determined using various techniques.

RELIABILITY ANALYSES: Studies of process or equipment failure or operability. An example of a reliability study would be an analysis of "how frequently a chemical reactor might overheat due to malfunctioning pumps, heat exchangers, human operators, control systems, and other plant equipment..." (Henley and Kumamoto, 1981: 8).

RESPONSE MODE: The form in which the subject is asked to give his judgment. Some numeric response modes that are commonly used are probabilities, odds, intervals, ratings, logs, and pairwise comparisons. Nonnumeric, qualitative, response modes include verbal or written descriptions, classifications, categories, or preferences.

RISK ANALYSIS: Risk analysis includes the data and techniques used to quantify risk. These analyses usually include a model of events leading to risks and their consequences. A risk analysis of a nuclear plant could describe major factors relating to accident events; frequency and uncertainty ranges of accident events; major factors and accident phenomena leading to these damage events; and consequences and risks of these to the public.

SAMPLE: A subset of a population of items that is chosen for examination. A statistically valid sample is chosen using sampling technique designed for representativeness of the population and for random selection of the items.

SHORT-TERM MEMORY: A memory of limited capacity and intermediate duration. Ericsson and Simon (1980) depict short-term memory as being where information is processed in problem solving.

SIGNIFICANCE OR SIGNIFICANCE LEVEL: The result of a statistical test or technique is said to be significant if the conclusions indicate a difference between the data and the assumed normal state of the world. For example, the test indicates two experts are correlated where it is assumed that experts are not correlated. Because nothing is 100% certain, there is a chance that the conclusion drawn from a test is incorrect. The probability of a significant result being incorrect is the significance level. This level is chosen by the analyst prior to the test and indicates the chance that he is willing to take that the conclusion is incorrect. Usually this level is 5% or less, but the choice is always indicated in the statement of the conclusion. For example, "the positive correlation between expert 1 and 2 is significant at the 5% level," means that experts 1 and 2 are positively correlated, and there is a 5% chance that they are not. This 5% is sometimes called the alpha level or the type I error.

SIMULATION: See specific simulation techniques: Bootstrap and Monte Carlo.

SOCIAL PRESSURE: An effect that induces individuals to slant their responses or to silently acquiesce to the views that they believe will be acceptable to the interviewer, their group, supervisors, organization, or society in general. This altering of an individual's thoughts can take place consciously or unconsciously. The social pressure can come from those physically present or from the subject's internal evaluation of how others would interpret their responses. People's need to be loved, respected, and recognized induces them to behave in a manner that will bring affirmation.

SOLUTION: Expert judgments that are given as descriptive text or diagrams, as opposed to numerical estimates.

STANDARD DEVIATION: the square root of the variance.

STATICIZED GROUP: See Individual Interview.

STRUCTURED DOCUMENTATION: A detailed type of record of the expert's answers and problem-solving processes. The person tasked with providing the documentation is usually provided with a format of what should be recorded. The format lists those aspects deemed to be the most important (e.g., answers and uncertainty levels, assumptions, and algorithms) and the level of detail at which the information is desired.

STRUCTURING ELICITATION: The amount of controls placed on the elicitation process. The interaction between the experts is one aspect of an elicitation method that is typically structured. Varying degrees of structure can be imposed, ranging from none to a high degree of structuring. No structuring would allow spontaneous interaction between the experts; a high degree would produce carefully choreographed communications. (See *Designing on Paper--Planning the Elicitation* for a description of the larger process of which structuring is a part.)

SUBSTANTIVE EXPERTISE: Expertise stemming from the expert's experience in the field in question, such as in the rupture rate of Westinghouse pipes.

SUMMARY DOCUMENTATION: A type of record of the expert's answers and problem-solving processes. Typically, it provides a few sentences or paragraphs on the experts' thinking, such as the sources of information that they used, their major assumptions, and their reasons for giving particular answers.

THINK ALOUD METHOD: See Verbal Protocol.

TOOL BIAS: The misrepresentation of the expert's data as a result of forcing these to fit the tools selected for analysis. The analyst, and people in general, tend to use those models or methods with which they are most comfortable. Then, they are often unable to objectively judge whether they have used the tool appropriately (e.g., the model required that the data have a normal distribution, and the data may not have).

TRAINING BIAS: The tendency of the data gatherer to introduce bias into the expert's data by misinterpreting it. It is an unconscious human tendency to interpret incoming information in terms of what is already believed, such as what has been learned through professional training. For example, it is common for a data gatherer to define a term using those definitions that he or she has learned rather than to elicit the expert's definitions.

TYPE I ERROR: See Significance Level.

UNDERESTIMATION OF UNCERTAINTY BIAS: The tendency to underestimate the true amount of uncertainty in giving an answer. For example, when people are asked to put a range around their answer such that they are 90% sure that the range encompasses the correct answer, their ranges only cover 30-60% of the total.

VARIABLE: See Random Variable.

VARIANCE: A measure of dispersion based on the squared differences between individual values and their mean or expected value. The standard deviation is the square root of the variance.

VERBAL PROBE: A method from educational psychology used to elicit information on the subject's problem solving. There are different types of verbal probes that vary in when and how they are asked. This book uses verbal probe is used to refer to a question which has a nonleading wording that is asked while the expert is still attending to the subject of the question.

VERBAL PROTOCOL METHOD: A method from educational psychology involving having the subject *think aloud* as he works through the problem. The verbal protocol is used in face-to-face interviews.

VERBATIM DOCUMENTATION: A record of the expert's answers and problem-solving processes. Obtaining a verbatim account is usually done by mechanically recording the expert's elicitation sessions and then transcribing them. This type of documentation is more frequent in artificial intelligence than in traditional expert judgment applications.

VOLUNTEERED DISPERSION MEASURE: A type of dispersion measure that the experts volunteer without being asked. This dispersion marks a spread of values around the expert's best estimate.

WISHFUL THINKING BIAS: A tendency that occurs when an individual's hopes influence his judgment. For example, people typically overestimate what they can produce in a given amount of time. In general, the greater the subject's involvement and the more he stands to gain from the answer, the greater this bias. Also called *conflict of interest* bias.

References

Amos, C. N., Benjamin, A. S. Boyd, G. J., Kunsman, D. M., Lewis, S. R., Smith, L. N., and Williams, D. C. (1987), "Evaluation of Severe Accident Risks and the Potential for Risk Reduction: Grand Gulf, Unit 1," NUREG/CR-4551, Vol. 4, SAND86-1309, Sandia National Laboratories, Albuquerque, NM.

Armstrong, J. S. (1981), Long-Range Forecasting: From Crystal Ball to Computer, Wiley-Interscience, New York, NY.

Armstrong, J. S., Denniston, W. B., Jr., and Gordon, M. M. (1975), "Use of the Decomposition Principle in Making Judgments," Organizational Behavior and Human Performance, **14**, pp. 257-263.

Ascher, W. (1978), Forecasting: An Appraisal for Policymakers and Planners, John Hopkins University Press, Baltimore, MD.

Baecher, G. B. (1979), "Correlations Among Experts' Opinions," Department of Civil Engineering, Massachusetts Institute of Technology, Boston, MA.

Barclay, S., Brown, R. V., Kelley, C. W., III, Peterson, C. R., Phillips, L. D., and Selvidge, J. (1977), "Handbook for Decision Analysis," Decisions and Designs, Inc., McLean, VA.

Baron, R. A. and Bryne, D. (1981), Social Psychology: Understanding Human Interaction, Allyn and Bacon Inc., Boston, MA.

Benjamin, A. S., Kunsman, D. M., Williams, D. C., Boyd, G. J., and Murfin, W. B. (1987), "Evaluation of Severe Accident Risks and the Potential for Risk Reduction: Surry Power Station, Unit 1," NUREG/CR-4551, Vol. 1, SAND86-1309, Sandia National Laboratories, Albuquerque, NM.

Bernreuter, D. C., Savy, J. B., Mensing, R. W., Chen, J. C., and Davis, B. C. (1985), "Seismic Hazard Characterization of the Eastern United States," UCID-20421, Lawrence Livermore National Laboratory, Livermore, CA.

Booker, J. M. (1978), "Mean Wheat Yield Prediction When Predictor Variables Are Subject to Error," Ph.D. dissertation, Texas A&M University, College Station, TX.

Booker, J. M. and Bryson, M. C. (1985), "Decision Analysis in Project Management: An Overview," IEEE Transactions in Engineering Management, EM-32, pp. 3-9.

Booker, J. M., Bryson, M. C., and McWilliams, T. P. (1984), "Decision Analysis Package for R&D Project Selection and Evaluation," LA-UR-84-41, Los Alamos National Laboratory, Los Alamos, NM.

Booker, J. M. and Meyer, M. A. (1985), "Sources and Effects of Correlation of Expert Opinion," LA-UR-85-1879, Los Alamos National Laboratory, Los Alamos, NM.

_____ (1988a), "Sources and Effects of Interexpert Correlation: An Empirical Study," (LA-UR-87-2998), IEEE Transactions on Systems, Man, and Cybernetics, **18**, (1), pp. 135-142.

_____ (1988b), "Determining the Independence of Experts Using Simulation," LA-UR-88-301, Los Alamos National Laboratory, Los Alamos, NM.

Boose, J. H. and Gaines, B. R. (1988), "Knowledge Acquisition for Knowledge-Based Systems," AAAI-88 Tutorial Notes, Seventh National Conference on Artificial Intelligence, St. Paul, MN, August 22.

Boose, J. H. and Shaw, M. (1989), "Knowledge Acquisition for Knowledge-Based Systems," AAAI-89 Tutorial Notes, Eleventh International Conference on Artificial Intelligence, Detroit, MI, August 20.

Brachman, R. J. and Levesque, H. J. (1985), Readings in Knowledge Representation, Morgan Kaufman Publishers, Los Altos, CA.

Bruckner, L. A. and Martz, H. F. (1987), "Beta Product," Statistics Group draft document, available from the Statistics Group, MS F600, Los Alamos National Laboratory, Los Alamos, NM.

Capen, E. C. (1975) "The Difficulty of Assessing Uncertainty," Society of Petroleum Engineers and the American Institute of Mining, Metallurgy, and Petroleum Engineers 50th Annual Fall Conference and Exhibit, Dallas, TX, September 28-October 1.

Cattell, R. B. (1963), "The Personality and Motivation of the Researcher From Measurements of Contemporaries and From Biographies," in Scientific Creativity: Its Recognition and Development, C. W. Taylor and F. Barron, eds., John Wiley, New York, NY, pp. 119-138.

Clancy, W. J. (1989), "Viewing Knowledge Bases as Qualitative Models," IEEE Expert, **4**, pp. 9-24.

Cleaves, D. A. (1986), "Cognitive Biases and Corrective Techniques: Proposals for Improving Elicitation Procedures for Knowledge-Based Systems," Proceedings from the AAAI sponsored 2nd Annual Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada, November 1986.

Clemen, R. T. (1986), "Calibration and the Aggregation of Probabilities," Management Science, **32** (3), pp. 312-314.

Clemen, R. T. and Winkler, R. L. (1985), "Limits for the Precision and Value of Information from Dependent Sources," Operations Research, **33**, pp. 427-442.

Club of Rome (1974), Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind, Universe Books, New York, NY.

Cochran, W. (1963), Sampling Techniques, John Wiley & Sons, New York, NY.

Comer, M. K., Seaver, D. A., Stillwell, W. G., and Gaddy, C. D. (1984), "Generating Human Reliability Estimates Using Expert Judgments," NUREG/CR-3688, Vol. 1-2, SAND84-7115, Sandia National Laboratories, Albuquerque, NM.

Conover, W. J. (1971), Practical Nonparametric Statistics, John Wiley & Sons, New York, NY.

Cook, I. and Unwin, S. D. (1986), "Controlling Principles for Prior Probability Assignments in Nuclear Risk Assessment," NUREG/CR-4514, SAND85-1323, Sandia National Laboratories, Albuquerque, NM.

Dalkey, N. C (1969), "An Experimental Study of Group Opinion: The Delphi Method," Futures, **1**, pp. 403-406.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989), "Clinical Versus Actuarial Judgment," Science, **243**, pp. 1668-1673.

Denning, P. J. (1986), "The Science of Computing--Will Machines Ever Think?" American Scientist, **74**, pp. 344-46.

_____ (1988), "The Science of Computing--Blindness in Designing Intelligent Systems," American Scientist, **76**, pp. 118-120.

Dhaliwal, J. S. and Benbasat, I. (1989), "A Framework for the Comparative Evaluation of Knowledge Acquisition Tools and Techniques," Proceedings of the AAAI-sponsored 4th Annual Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada, pp. 9-1 to 9-24.

Dougherty, E. M., Jr., Fragola, J. R., and Collins, E. P. (1986), "Human Reliability Analysis," SAIC/NY-86-1-OR, Science Applications International Corporation, Oak Ridge, TN 37831, April.

Dreyfus, H. L. and S. E. (1986), Mind Over Machine, Free Press, NY.

Duran, B. S. and Odell, P. L. (1974), Cluster Analysis: A Survey, Springer-Verlag, New York.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," Annals of Statistics, **7**, pp. 1-26.

Efron, B. and Gong, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation," The American Statistician, **37**, pp. 36-48.

Egan, D. E. and Schwartz, B. J. (1979), "Chunking in Recall of Symbolic Drawings," Memory and Cognition, **7**, pp. 149-158.

Elston, A., Emerson, J. D., Meyer, M. A., Osborn, M. M., and Stoddard, M. L. (1986), "Report of the Fort Knox AOBC Prototype I Data Collection," S-6:86-U-429, Military Systems Group, Los Alamos National Laboratory, Los Alamos, NM.

Ericsson, K. A. and Simon, H. A. (1980), "Verbal Reports as Data," Psychological Review, **87**, (3), pp. 215-250.

_____ (1984), Protocol Analysis: Verbal Reports as Data, MIT Press, Cambridge, MA.

Faust, D. and Ziskin, J. (1988), "The Expert Witness in Psychology and Psychiatry," Science, **241**, pp. 31-35.

Feller, W. (1957), An Introduction to Probability Theory and Its Applications, Vol. 1, John Wiley & Sons, New York, NY.

Festinger, L. (1957), A Theory of Cognitive Dissonance, Stanford University Press, Palo Alto, CA.

Fogel, L. J. (1967), Human Information Processing, Prentice-Hall, Englewood Cliffs, NJ.

French, S. (1986), "Calibration and the Expert Problem," Management Science, **32**, (3), pp. 315-321.

Gaines, B. R. and Boose, J., eds. (1988), Knowledge Acquisition for Knowledge-Based Systems, Academic Press, San Diego, CA.

Gaines, B. R. and Shaw, M. L. (1989), "Comparing the Conceptual Systems of Experts," Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, August, Morgan Kaufmann Publishers, Detroit, MI, pp. 633-638.

Genest, C. and Zidek, J. V. (1986), "Combining Probability Distributions: A Critique and an Annotated Bilbilography," Statistical Science, **1**, (1), pp. 114, 148.

Goffman, E. (1959), The Presentation of Self in Everyday Life, Doubleday Anchor Books, Garden City, NY.

Gold, P. E. (1987), "Sweet Memories," American Scientist, March-April, pp. 151-155.

Gorden, R. L. (1980), Interviewing: Strategy, Techniques, and Tactics, Dorsey Press, Homewood, IL.

Gough, R. (1975), "The Effect of Group Format on Aggregate Subjective Probability Distributions," in Utility, Probability, and Human Decision Making, D. Wendt and C. Viek, eds., Dordrecht-Holland, Reidel.

Gustafson, D., Shukla, R., Delbecq, A., and Walster, G. (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**, pp. 280-291.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969), "Analysis of Categorical Data by Linear Models," Biometics, **25**, pp. 489-504.

Harrison, J. M. (1977), "Independence and Calibration in Decision Analysis," Management Science, **24**, (3), pp. 320-328.

Hayes-Roth, B. (1980), "Estimation of Time Requirements During Planning: Interactions Between Motivation and Cognition," N-1581-ONR, Rand Corporation, Santa Monica, CA.

Henley, E. J. and Kumamoto, H. (1981), Reliability Engineering and Risk Assessment, Prentice-Hall, Englewood Cliffs, NJ.

Henrion, M. and Cooley, D. R. (1987), "An Experimental Comparison of Knowledge Engineering for Expert Systems and for Decision Analysis," Proceedings of the 6th National American Association for Artificial Intelligence held in Seattle, July, Morgan Kaufmann Publishers, Los Altos, CA, pp. 471-476.

Hirely, W. (1989), "Survey Probes Tensions Between Science and Democracy," The Science Observer, American Scientist, **77**, pp. 24-27.

Hoffman, R. R. (1987), "The Problem of Extracting the Knowledge of Experts from the Perspective of Experimental Psychology," Artifical Intelligence Magazine, Summer, pp. 53-67.

Hogarth, R. (1975), "Cognitive Processes and the Assessment of Subjective Probability Distributions," Journal of the American Statistical Association, **70**, (350), pp. 271-291.

_____ (1980), Judgment and Choice: The Psychology of Decisions, Wiley-Interscience, Chicago, IL.

Janis, I. C. (1972), Victims of Group Think: A Psychological Study of Foreign Policy Decisions and Fiascos, Houghton Mifflin, Boston, MA.

Johnson, M. E. (1987), Multivariate Statistical Simulation, John Wiley & Sons, New York, NY.

Kahneman, D. and Tversky, A. (1982), "Subjective Probability: A Judgment of Representativeness," Judgment Under Uncertainty: Heuristics and Biases,

Kahneman, D., Slovic, P., and Tversky, A., eds., Cambridge University Press, Cambridge, MA, pp. 32-47.

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

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

Knapp, R. H. (1963), "Demographic, Cultural, and Personality Attributes of Scientists," Scientific Creativity: Its Recognition and Development. C. W., Taylor and F. Barron, eds., John Wiley, New York, NY, pp. 205-216.

Kouts, H., Cornell, A., Farmer, R., Hanauer, S. and Rasmussen, N. (1987), "Methodology for Uncertainty Estimation in NUREG-1150 (Draft): Conclusions of a Review Panel," a letter sent to the U.S. Nuclear Regulatory Commission, Division of Reactor Accident Analysis, Office of Nuclear Regulatory Research and received August 21, 1987. Also published as NUREG/CR-5000, BNL-NUREG-52119 in December 1987, Department of Nuclear Energy, Brookhaven National Laboratory, Upton, NY.

Krupka, M. C., Peaslee, A. T., Jr., and Laquer, H. L. (1983), "Gaseous Fuel Safety Assessment for Light-Duty Automotive Vehicles," LA-9829-MS, Los Alamos National Laboratory, Los Alamos, NM.

Kshirsagar, A. M. (1972), Multivariate Analysis, Marcel Dekker, New York, NY.

LaFrance, M. (1988), "The Knowledge Acquisition Grid: A Method for Training Knowledge Engineers," Knowledge Acquisition for Knowledge-Based Systems, Vol. 1, Gaines, B. R. and Boose, J. H., eds., Academic Press, London, U.K.

Lichtenstein, S. and Fischhoff, B. (1977), "Do Those Who Know More Also Know More about How Much They Know?" Organizational Behavior and Human Performance **20**, pp. 159-183.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982), "Calibration of Probabilities: The State of the Art to 1980, Judgment Under Uncertainty: Heuristics and Biases, Kahneman, D., Slovic, P., and Tversky, A., eds., Judgment Under Uncertainty: Heuristics and Biases, Cambridge University Press, Cambridge, MA, pp. 306-334.

Lindley, D. V. and Singpurwalla, N. D. (1984), "Reliability and Fault Tree Analysis Using Expert Opinions," GWU/IRRA/TR-84/10, George Washington University, Washington, DC.

Mahoney, Michael (1976), The Scientist as Subject: the Psychological Imperative, Ballinger Publishing Co., MA.

Martz, H. F., Beckman, R. J., Campbell, K., Whiteman, D. E., and Booker, J. M. (1983), "A Comparison of Methods for Uncertainty Analysis of Nuclear Power Plant Safety System Fault Tree Models," LA-9729-MS, Los Alamos National Laboratory, Los Alamos, NM, April.

Martz, H. F., Bryson, M. C., and Waller, R. A. (1985), "Eliciting and Aggregating Subjective Judgments--Some Experimental Results," Proceedings

of the Tenth Annual SAS Users Group International Conference, SAS Institute Inc., Cary, NC.

Martz, H. F. and Waller, R. A. (1982), Bayesian Reliability Analysis, John Wiley & Sons, New York, NY.

Mathiowetz, N. A. (1987), "Response Error: Correlation Between Estimation and Episodic Recall Tasks," in the Survey Research Methods Proceedings of the Joint Statistical Meetings, San Francisco, CA, pp. 430-435.

McGraw, K. L. and Harbison-Briggs, K. (1989), Knowledge Acquisition: Principles and Guidelines, Prentice Hall, Englewood Cliffs, NJ.

Meyer, M. A. (1986), "Human Factors Affecting Subjective Judgments," in Proceedings of the Thirty-First Conference on the Design of Experiments in Army Research and Development, Madison, WI. LA-UR-84-3176, Los Alamos National Laboratory, Los Alamos, NM.

_____ (1987), "Eliciting Data on Subject's Problem Solving of Computerized Exercises: A Cognitive Interview," A-1:87-U-601, Statistics Group internal report, Los Alamos National Laboratory, Los Alamos, NM.

Meyer, M. A. and Booker, J. M. (1987a), "Problems with Expert Data in Studies of Inter-Expert Correlation," in the Social Statistics Proceedings of the Joint Statistical Meetings, San Francisco, CA, pp. 88-97.

_____ (1987b), "Sources of Correlation Between Experts: Empirical Results from Two Extremes," NUREG/CR-4814, LA-10918-MS, Los Alamos National Laboratory, Los Alamos, NM.

_____ (1989), "A Practical Program for Handling Bias in Knowledge Acquisition," Proceedings of the AAAI-sponsored 4th Annual Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada, pp. 23-1 to 23-20.

Meyer, M. A., Booker, J. M., Cullingford, H. S., and Peaslee, A. T., Jr. (1981), "A Data-Gathering Method for Use in Modeling Energy Research, Development and Demonstration Programs," S-4/81-11, Technology Assessment Group, Los Alamos National Laboratory, Los Alamos, NM.

_____ (1982), "A Data-Gathering Method for Use in Modeling Energy Research, Development and Demonstration Programs," Energy Programs, Policy, and Economics: Alternative Energy Sources IV, Vesiroglu, T. N., ed., Butterworth Publishers, FL., pp. 421-430.

Meyer, M. A. and Johnson, E. R. (1985), "Proceedings from the Technology Control Panel Workshop," LA-UR-85-4078, Los Alamos National Laboratory, Los Alamos, NM.

Meyer, M. A., Mniszewski, S. M., and Peaslee, A. T., Jr. (1989), "Using Three Minimally Biasing Elicitation Techniques for Knowledge Acquisition," Knowledge Acqusition, **1**, pp. 59-72.

Meyer, M. A., Peaslee, A. T., Jr., and Booker, J. M., (1982), "Group Consensus Methods and Results," LA-9584-MS, Los Alamos National Laboratory, Los Alamos, NM.

Miller, G. A. (1956), "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," Psychological Review, **63**, pp. 81-97.

Morris, P. A. (1977), "Combining Expert Judgments: A Bayesian Approach," Management Science, **23**, (3), pp. 679-693.

_____ (1986), "Observations on Expert Aggregation," Management Science, **32** (3), pp. 321-328.

Mumpower, J. L., Phillips, C. D., Renn, O., and Uppuluri, V. R. R. (1987), Expert Judgment and Expert Systems, NATO ASI series, Springer-Verlag, NY.

Ortiz, N. R., Wheeler, T. A., Meyer, M. A., and Keeney, R. L. (1988), "The Use of Expert Judgment in NUREG-1150," to appear in the Proceedings of the 16th Water Reactor Safety Information Meeting, Gaithersburg, ML, October. Sand88-2253C, Sandia National Laboratories, Albuquerque, NM.

Payne, S. (1951), "The Art of Asking Questions," Princeton University Press, Princeton, NJ.

PERT Coordinating Group (1963), "PERT Guide for Management Use," Department of Defense, Washington, D.C.

Raiffa, H. (1970), Decision Analysis: Introductory Lectures on Choices Under Uncertainty, Addison-Wesley, Reading, MA.

Ripley, B. A. (1987), Stochastic Simulation, John Wiley & Sons, New York, NY.

Roe, A. (1952), The Making of a Scientist, Greenwood Press, Westport, CT.

_____ (1963), "Personal Problems and Science," in Scientific Creativity: Its Recognition and Development, C. W. Taylor and F. Barron, eds., John Wiley, New York, NY, pp. 132-138.

Rosenfield, I. (1988), The Invention of Memory: A New View of the Brain, Basic Books, NY.

Saaty, T. L. (1980), The Analytic Hierarchy Process: Planning, Priority Setting, and Resource Allocation, McGraw-Hill, NY.

_____ (1982), Decision Making for Leaders: Lifetime Learning Publications, Belmont, CA.

Schervish, M. J. (1986), "Comments on Some Axioms for Combining Expert Judgements," Management Science, **32**, pp. 306-312.

Seaver, D. A. (1976), "Assessments of Group Preferences and Group Uncertainty for Decision Making," Social Science Research Institute, University of Southern California, Los Angeles, CA.

_____ (1978), "Assessing Probability with Multiple Individuals: Group Interaction Versus Mathematical Aggregation," SSRI Research Report 78-3, Social Science Research Institute, University of Southern California, Los Angeles, CA.

Seaver, P. A. and Stillwell, W. G. (1983), "Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations," NUREG/CR-2743, SAND82-7054, Sandia National Laboratories, Albuquerque, NM.

Shaw, M. L. G. and Woodward, J. B. (1989), "Mental Modes in the Knowledge Acquisition Process," Proceedings of the AAAI-sponsored 4th Annual Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada, pp. 29-1 to 29-24.

Skuse, D. and Sowa, J. F. (1988), "Knowledge Representation: Design Issues," AAAI Tutorial notes, Seventh National Conference on Artificial Intelligence, St. Paul, MN, August 21.

Snedecor, G. W. and Cochran, W. G. (1978), Statistical Methods, sixth edition, Iowa State University Press, Ames, IO.

Sowa, J. F. (1984), Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA.

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

Spradley, J. P. (1979), The Ethnographic Interview, Holt, Rhinehart, and Winston, NY.

Stael von Holstein, C. A. (1971), "Two Techniques for Assessment of Subjective Probability Distributions: An Experimental Study," Acta Psychologia, **35**, pp. 378-394.

Stillwell, W. G., Seaver, D. A., and Schwartz, J. P. (1982), "Expert Estimation of Human Error Probabilities in Nuclear Power Plant Operations: A Review of Probability Assessment and Scaling," NUREG/CR-2255, SAND81-7140, Sandia National Laboratories, Albuquerque, NM.

Stoto, M. A. (1988), "Dealing with Uncertainty: Statistics for an Aging Population," American Statistician, **42**, pp. 103-110.

Stroud, M. (1980), "Professional Writing: Strategies to Improve the Transfer of Knowledge," Communications Strategies of Advanced Technologies, Albuquerque, NM.

Tietjen, G. L. (1986), A Topical Dictionary of Statistics, Chapman and Hall, New York, NY.

Tversky, A. and Kahneman, D. (1974), "Judgments Under Uncertainty: Heuristics and Biases," Science, **185**, pp. 1124-1131.

_____ (1981), "Framing of Decisions and the Psychology of Choice," Science, **211**, pp. 453-458.

U.S. Nuclear Regulatory Commission (NRC), Office of Nuclear Regulatory Research (1983), "PRA Procedures Guide: A Guide to the Performance of Probabilistic Risk Assessments for Nuclear Power Plants," NUREG/CR-2300, Vol. 1-2, prepared under the auspices of the American Nuclear Society and the Institute of Electrical and Electronic Engineers under a grant from the Nuclear Regulatory Commission, Washington, D. C.

U.S. Nuclear Regulatory Commission (NRC), Office of Nuclear Regulatory Research (1989), "Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants," (formerly entitled "Reactor Risk Reference Document") NUREG-1150, Vol. 1-2, second draft for peer review, Washington, DC.

Van de Ven, A. and Delberq, A. (1974), "The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes," Academic Managment Journal, **17**, pp. 605-621.

Vedder, R. G. and Mason, R. O. (1987), "An Expert System Application for Decision Support in Law Enforcement," Decision Science , **18**, pp. 400-414.

Waern, Y. (1987), "Mental Models in Learning Computerized Tasks," in Psychological Issues of Human-Computer Interactions in the Work Place, M. Frese, E. Ulich, and W. Dzida, eds., North-Holland Press, Amsterdam. pp. 275-94.

Waterman, D. A. (1986), A Guide to Expert Systems, Addison-Wesley Publishing, Reading, MA.

Weissenberg, P. (1971), Introduction to Organizational Behavior: A Behavioral Science Approach to Understanding Organizations, Intext Educational Publishers, Scranton, OH.

Welbank, M. (1983), A Review of Knowledge Acquisition Techniques for Expert Systems, Martlesham Consultancy Services, British Telecom Research Laboratories, Martlesham Heath, Ipswich, England, IP5 7RE.

Wheeler, T. A., Hora, S. C., Cramond, W. R., and Unwin, S. D. (1989), "Analysis of Core Damage Frequency from Internal Events: Expert Judgment Elicitation," NUREG/CR-4550, Vol. 2, SAND86-2084, Sandia National Laboratories, Albuquerque, NM.

Winkler, R. L. (1968), "The Consensus of Subjective Probability Distributions," Management Science, **15**, (2), pp. B-61 through B-75.

_____ (1981), "Combining Probability Distributions from Dependent Information Sources," Management Science, **27**, pp. 987-997.

Winkler, R. L. (1986), "Expert Resolution," Management Science, **32**, (3), pp. 298-303.

Zimbardo, P. G. (1983), "To Control a Mind," Stanford Magazine, Winter, pp. 59-64.

NRC FORM 335 (8-87) NRCM 1102, 3201, 3202

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

SEE INSTRUCTIONS ON THE REVERSE

1. REPORT NUMBER (Assigned by PPMB: DPS, add Vol. No., if any)

NUREG/CR-5424
LA-11667-MS

2. TITLE AND SUBTITLE

Eliciting and Analyzing Expert Judgement:
A Practical Guide

3. LEAVE BLANK

4. DATE REPORT COMPLETED

MONTH	YEAR
December	1989

5. AUTHOR(S)

M. A. Meyer, J. M. Booker

6. DATE REPORT ISSUED

MONTH	YEAR
January	1990

7. PERFORMING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)

Los Alamos National Laboratory
Los Alamos, NM 87545

8. PROJECT/TASK/WORK UNIT NUMBER

9. FIN OR GRANT NUMBER

A7225

10. SPONSORING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)

Division of Systems Research
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, DC 20555

11a. TYPE OF REPORT

Technical

b. PERIOD COVERED (Inclusive dates)

12 SUPPLEMENTARY NOTES

13. ABSTRACT (200 words or less)

In this book we describe how to elicit and analyze expert judgment. Expert judgment is defined here to include both the experts' answers to technical questions and their mental processes in reaching an answer. It refers specifically to data that are obtained in a deliberate, structured manner that makes use of the body of research on human cognition and communication. Our aim is to provide a guide for lay persons in expert judgment. These persons may be from physical and engineering sciences, mathematics and statistics, business, or the military. We provide background on the uses of expert judgment and on the processes by which humans solve problems, including those that lead to bias. Detailed guidance is offered on how to elicit expert judgment ranging from selecting the questions to be posed of the experts to selecting and motivating the experts to setting up for and conducting the elicitation. Analysis procedures are introduced and guidance is given on how to understand the data base structure, detect bias and correlation, form models, and aggregate the expert judgments.

14 DOCUMENT ANALYSIS - a KEYWORDS/DESCRIPTORS

expert judgment, expert opinion, elicitation, subjective judgment, knowledge acquisition

b. IDENTIFIERS/OPEN-ENDED TERMS

15. AVAILABILITY STATEMENT

Unlimited

16. SECURITY CLASSIFICATION

(This page) Unclassified

(This report) Unclassified

17. NUMBER OF PAGES

18. PRICE