LEGIBILITY NOTICE

A major purpose of the Technical Information Center is to provide the broadest dissemination possible of information contained in DOE's Research and Development Reports to business, industry, the academic community, and federal, state and local governments.

Although a small portion of this report is not reproducible, it is being made available to expedite the availability of information on the research discussed herein.

LA-UR -90-2604

CONF-9009226--1

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

LA-UR--90-2604

DE90 014975

TITLE COMMON PROBLEMS IN THE ELICITATION AND ANALYSIS OF EXPERT OPINION AFFECTING PROBABILISTIC SAFETY ASSESSMENTS

AUTHOR(S) Mary A. Meyer, A-1
Jane M. Booker, A-1

SUBMITTED TO To be presented at the CSNI Conference in Santa Fe, September 4-6, 1990

DISCLAIMER

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

1/25/90

COMMON PROBLEMS IN THE ELICITATION AND ANALYSIS OF EXPERT OPINION AFFECTING PROBABILISTIC SAFETY ASSESSMENTS

Mary A. Meyer and Jane M. Booker
Los Alamos National Laboratory, MS F600, Los Alamos, NM, 87544, USA

ABSTRACT

Expert opinion is frequently used in probabilistic safety assessment (PSA), particularly in estimating low probability events. In this paper, we discuss some of the common problems encountered in eliciting and analyzing expert opinion data and offer solutions or recommendations. The problems are:
(1) that experts are not naturally Bayesian. People fail to update their existing information to account for new information as it becomes available, as would be predicted by the Bayesian philosophy.
(2) that experts cannot be fully calibrated. To calibrate experts, the feedback from the known quantities must be immediate, frequent, and specific to the task.
(3) that experts are limited in the number of things that they can mentally juggle at a time to 7 ± 2.
(4) that data gatherers and analysts can introduce bias by unintentionally causing an altering of the expert's thinking or answers.
(5) that the level of detail in the data, or granularity, can affect the analyses.
(6) that the conditioning effect poses difficulties in gathering and analyzing of the expert data. The data that the expert gives can be conditioned on a variety of factors that can affect the analysis and the interpretation of the results.

INTRODUCTION

What Is Expert Opinion?

Expert opinion is information given by an expert in response to a technical problem. This information is often the only available or supporting data for rare or never observed events. For this reason, we will refer to the information provided by experts as *data*.

An expert is a person who has background in the subject area and is recognized by his or her peers or those conducting the study as qualified to answer questions.

Expert opinion can vary in form from being an answer (e.g., an estimate of the probability of an occurrence of a nuclear reactor accident of a particular type) to a description of the expert's thought processes in arriving at an answer.

Expert opinion has also been called expert judgment, subjective judgment, expert forecast, best estimate, educated guess, and most recently, expert knowledge. Whatever it is called, expert opinion is more than a guess. It is an informed judgment based on the expert's training and experience, and it is useful data for analysis and interpretation purposes.

When Expert Opinion is Used

Expert opinion data has been widely used, especially in technical fields. This type of data provides information when other data sources, such as measurements, experimentation, observations, 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 opinion was used where

"experimental or observational data or validated computer models were not available or not widely agreed upon" (Ortiz, Wheeler, Meyer & Keeney, 1988, p. 4).

Expert opinion has been specifically gathered 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 estimating the likelihood of rare reactor accidents. Another would be estimating the safety of designs for the new production reactor.
- To forecast future events. In general, when applicable data are unavailable, predicting future events or actions requires use of expert opinion. To make these predictions, the experts must adjust, sometimes radically, from the status quo or past patterns. For instance, the demand for various utilities in the United States, such as electricity, may come from experts' projections (Ascher, 1978).
- To integrate or interpret existing data. Expert opinion 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 the expert's reasons for giving an answer (e.g. his assumptions, definitions, and problem-solving processes), the expert's answer encoded in descriptive categories or preference scales (e.g., poor, moderate, and good), or additional information regarding existing data (e.g., theory T suggests that system A is more reliable than system B).

  Quantitative data are numeric data such as estimates of probabilities, physical phenomenon (e.g., temperature), simple ranks or ratings (e.g., 1-5), and error bounds on any such estimates of probability, physical phenomenon, or ranks or ratings (e.g., 0.75 ± 0.25).
- To learn an expert's problem-solving process or a group's decision-making processes. The experts often do not know how they solve a problem or reach a decision because their thoughts have become so automatic that the process is difficult to recall. However, their problem-solving techniques are needed to support their answers, to improve current practices, to train others, or to create systems that provide expert advice.
- 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 two benefits: it can serve as a complement to the current state of knowledge or as motivation for further study.

Expert opinion is often used to address more than one of the above-mentioned needs. Such was the case in the reactor risk project, NUREG-1150 (Wheeler, Hora, Cramond & Unwin, 1989), where the expert opinion met all of the above-mentioned purposes.

In addition, gathering expert opinion often provides side benefits: one of the most common benefits is the facilitation of communication. The experts readily see how their opinions 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 opinion can be viewed as a representation, a snapshot, of the expert's knowledge at the time of response to the technical question (Keeney & von Winterfeldt, 1989). As Ascher (1978, p. 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 opinion legitimately can and should change as the expert receives new information. In addition, because the opinion reflects the expert's knowledge and learning, the experts can validly differ in their opinions.

An expert frequently gives answers 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 opinion 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 processes 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 & Collins, 1986, p. 4-2) Thus, knowledge-based cognition is often invoked by the situations for which expert opinion is sought.

The quality of expert opinion varies according to how the data are gathered; the data can be obtained in a variety of ways, ranging from the unconscious to the deliberate. Expert opinion is often gathered unconsciously such as in technical projects. Analysts typically make decisions in defining problems, establishing boundary conditions, and screening data without being aware that they have used their own expert opinion. For example, analysts or experts select particular reactor environments or operating conditions, thereby eliminating conditions judged exotic or unimportant.

Expert opinion 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 opinions *off the top of their heads*. The informal means of gathering expert opinion has been a source of current controversies involving the accuracy of expert opinion.

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 opinion practices and cognitive limitations common to human beings" (Faust & Ziskin, 1988, p. 33). In other words, experts are subject to the same cognitive and motivational biases of other humans if certain preventative measures are not used in the elicitation.

By *bias* is meant a skewing of the expert's opinion from some reference point. With *cognitive biases*, the expert opinion is considered skewed from the standpoint of mathematical or logical rules, usually because of the ways in which the expert mentally processes information. With *motivational biases*, the expert opinion is skewed from the standpoint of the expert's thoughts. The expressions of the expert's thinking do not match

either the expert's thoughts or answers at the time of the elicitation because the expert has altered the reports of his or her thoughts or because the interviewer has misunderstood the expert's reports.

Formal means of gathering expert opinion 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 opinions. Because the quality of expert opinion is often evaluated in terms of the methods used to gather the opinions, the greater time and effort associated with the formal approach is an advantage.

Second, the formal approach lends itself more readily to documentation than those methods used unconsciously or less formally. That is, records are usually kept of the elicitation methods used and of how the experts arrived at their final opinions. Such a record allows the formal method and its results to be scrutinized and the results updated to reflect new information. Thus, the formal approach is more likely to advance through the review process (Ortiz et al., 1988).

COMMON PROBLEMS IN ELICITATION AND ANALYSIS

In this section, we address a few misconceptions concerning expert opinion and aspects of elicitation and analysis that have typically caused problems. We hope that this information will alert readers and prevent their falling into the difficulties described.

Experts Are Not Naturally Bayesian

Many analysts and theorists in the decision, reliability and safety analysis community advocate the Bayesian analysis approach. In these communities, this approach has led to a philosophy regarding the evaluation of gathered data as they are conditioned on other events or circumstances (e.g., other variables). Given that the data are conditioned on other variables, the Bayesian philosophy implies that as these conditions change, the data change. In other words, data are updated with changing conditions. However, a major problem occurs when applying a Bayesian approach to expert opinion because experts are not naturally Bayesian (Kahneman & Tversky, 1982). Human cognitive processes do not seem to follow Bayesian philosophy.

Humans are not Bayesian for a variety of reasons demonstrated in both laboratory settings and actual applications. The studies of Kahneman and Tversky (1982) 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 $P(A|B)=P(A|C)$; that is, 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; that is, $P(A|C)=P(A)$ and $P(A|B)=P(A)$. However, in estimating probabilities it is unlikely that any event would be totally independent of all conditions.

Other characteristics of human cognition prevent experts from being Bayesian. Some of these characteristics include the inability of humans to grasp (1) the effects of sample size, (2) the frequencies of truly rare events, (3) the meaning of randomness, (4) and the effects of variability (Hogarth, 1975).

We offer examples of each of these characteristics below.

(1) Humans will estimate the same event failure frequency whether they observed one failure in 100 times or one failure in 1000 times.

This shows a failing to account for changing sample size.
(2) Humans will estimate the frequency of a rare event to be higher if they have personally experienced that event.
(3) Humans consider a set of values with *any* sort of pattern, repetition, or clustering to be nonrandom. This natural tendency results in truly random processes being considered nonrandom.
(4) Humans underestimate the variance of a distribution of values. If an expert is asked to estimate the variance or standard deviation of the temperatures necessary for a component to overheat, that expert will provide a variation that can be anywhere from one-half to one-fourth of the actual variability.

These same inabilities also contribute to human difficulties in estimating probabilities in general (Kahneman & Tversky, 1982). Human cognition does not follow the axioms (rules) of probability, such as all probabilities lie in the [0,1] interval and the sum of mutually exclusive and exhaustive event probabilities must be 1. The probabilities elicited from a human do not represent a true, mathematical, probability measure.

To counter these many human tendencies, careful attention is needed in selecting the appropriate method for gathering the expert's answers. In particular, the mode in which the expert is asked to respond should be one that the expert understands and one that avoids asking for quantities that humans do not accurately estimate (e.g., the variance). Frequently, experts will need to be trained in the use of the chosen response mode. For example, if the question asks the experts to estimate probabilities of an event, the expert must be trained in the meaning of the term *probability*.

Hogarth (1980, p. 149) offers 8 keys to aid the expert in the use of probability thinking and probability distributions.

(1) Think in terms of two different sources of variation around the mean value. The first source is a natural or background noise level variation; the second source is variation from a lack of reliability or from measurement errors.
(2) Remember that the variability in the raw data is smaller than the variance of the mean of that data. Variances of the mean are affected by both the number of observations in the data set and by the variation in the data set. The larger the number of observations, the smaller will be the variance of the mean.
(3) Ask "what is the base-rate?" (e.g., what is the standard of comparison).
(4) Ask "what is the validity of of the information source?" and "how does it related to the predictive target?"
(5) Question the reliability of the information source. Imperfect reliability implies less predictive ability. Avoid extreme predictions based on poor or extreme data sources.
(6) Distinguish between information sources that overlap (i.e., dependent) and sources than differ (i.e, independent).
(7) Ask to what extent your data could be explained by a random process.
(8) Ask if it is possible to test your predictions.

While these keys help in probability predicting, they rely on the experts' knowledge of the concepts of variation, sample size, Bayesian updating and randomness. As we have already noted, experts must be trained and monitored to counter biases resulting from misuse and misunderstanding of these concepts.

Experts Cannot Be Fully Calibrated

The concept of calibration is basic to the scientific method. We define *calibration* to mean the comparison of an unknown (instrument or process) with a defined standard or a correct

procedure to adjust the unknown until it matches the standard. Until recently, calibration applied to measuring devices or processes for which standards or known quantities were available. The concept of calibrating experts was attractive because people are known to drift from particular standards (e.g., to be prone to motivational biases, to fail to adequately update their estimates in light of new information, to fail to account for the effects of sample size and variability, and to fail to follow the axioms of probability theory).

However, 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 that hamper calibration. For feedback to be effective as a calibration tool, it must be immediate, frequent, and specific to the task (e.g., the data received by meteorologists in weather forecasting). Such feedback cannot often be given in the case of risk, reliability, and safety assessments because the outcomes lie in the future and are unknown.

While this situation of uncalibrated experts and unknown outcomes may seem problematic, many users of expert opinion (e.g., decision makers) do not worry about biases arising from their experts. Instead, they have faith in experts because they perceive them as being very knowledgeable (Morris, 1986). This faith in the expert's opinion is not to imply that researchers ignore the calibration problem. On the contrary, calibration issues have led many decision analysts to focus on problems that arise from expert-decision maker interactions. 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 or her own when reaching a final decision.

Decision makers also affect calibration through the evaluation of their own as well as the expert's calibration. For example, decision makers who see themselves as miscalibrated can induce additional biases by over-compensating for their perceived lack of calibration. Furthermore, they may not perceive independence when it actually exists (Harrison, 1977). Thus, awareness of miscalibration and overcompensation for it, just as ignorance of it, can acerbate calibration problems.

Mathematical model techniques such as those from Winkler (1968) combine the decision maker's information with that of his experts. Some of these models incorporate calibration terms to adjust for miscalibration. However, using these models requires estimating or determining the model terms (e.g., the miscalibration, the correlation among the experts and the decision maker). Therefore, the use of these models is limited.

An alternative solution to complex mathematical models is to handle miscalibration with the use of broad uncertainty measures and characterizations (Meyer & Booker, chapter 17, 1989). The uncertainty measures provide an envelope around the experts' estimates that attempts to capture the true estimate by accounting for the bias induced by the lack of calibration.

Another more basic approach is to use elicitation methods that will minimize the biases that contribute to miscalibration. For example, interviewers can select elicitation methods and phrasings of the questions that reduce the potential for motivational bias by avoiding *leading* the expert. In addition, the data gatherers can monitor the elicitation for signs of the occurrence of various biases, as described in Meyer, Booker, and Bradshaw (1990).

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. A classic paper by Miller (1956) identifies the number of things that people can accurately discriminate. In these studies, the subjects differentiated things on the basis of one attribute, such as the volume of a sound. For example, when subjects listened to a sound played at varying levels of loudness, they could accurately discern about 7 levels. Other experiments included differentiating the size of drawn squares, determining the saltiness of various solutions, and distinguishing between musical notes. From many such experiments, Miller determined a limit of 7 as our processing capacity because the number of errors increases greatly after that point.

The number 7 does not represent 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 opinions in a row.

Multidimensional data is the input that we simultaneously receive from our five senses, assess, and act on as functioning human beings. As in their daily opinions, the subjects also exceeded the limit of 7 in experiments using multidimensional attributes. For example, in an experiment that produced combinations of six acoustical variables, subjects readily discerned 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 opinions, "we can make relatively crude opinions of several things simultaneously" (Miller, 1956, p. 88).

We can also exceed the limit of 7 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 an individual consecutively makes several absolute opinions, the information is stored in short-term memory. Memory has its limits, such as in 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 capability is called *chunking*. For example, a person learning radiotelegraphic code first hears each dot and dash 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 & Schwartz, 1979).

The information mentioned in this section has several implications for expert opinion. At the very least, it suggests that the interviewer avoid creating an elicitation situation in which the experts have to mentally juggle more than 7 items at a time. This suggestion applies also to response modes. When rating scales are used, the interviewer may wish to limit their gradations to 7 or less because more gradations impair an expert's ability to make fine discriminations.

If the project demands that a high number of distinctions be made simultaneously, the data gatherers

should take into account that the experts will judge these distinctions more crudely than if they had considered them separately or in pairs. In such situations, the interviewer may consider having the experts compare two items at a time. There are techniques such as Saaty's analytical hierarchy process (Saaty, 1980) designed for making pairwise comparisons.

Data Gatherers and Analysts Can Introduce Bias

The data gatherers and analysts can unintentionally introduce bias into expert opinion. Bias in this case refers to motivational bias: an altering of the expert's responses because of the influence of the data gatherer (interviewer or knowledge engineer) or analyst. Specifically, the data gatherers and analysts can cause bias through *misinterpretation* or *misrepresentation* of the expert data.

Data gatherers can misinterpret the expert data when they listen and record an expert's thoughts; factors like personal knowledge, training, and experience influence their understanding of the expert's words. For example, when an engineer, economist, and decision analyst met initially with military experts on a manufacturing matter, they each interpreted the information in terms of their own training. The engineer perceived the problem as an engineering one, the economist a cost/benefit one, and the decision analyst a multiattribute decision-theory one. They each questioned the experts to obtain the additional information that they needed to apply their individual orientation in greater depth. They each believed that they had received information that confirmed the applicability of their own training to treating the matter. For this reason, we also refer to this source of bias as *training bias*.

Data misrepresentation occurs during the later stages of an expert opinion project; specifically, misrepresentation can occur when the data is represented, modeled, or analyzed. When performing analyses, analysts tend to force the data into models or methods with which they are most comfortable or familiar. For this reason, misrepresentation is also referred to as *tool bias*.

A parallel example would be to use a hammer not only to drive nails into wood but screws and bolts as well. In all cases, the tool would perform its function, in some instances better than others. If the hammer continued to perform its function through time, the user would probably never realize the tool's shortcomings nor would he or she seek an alternative. The same can be said for models. For example, analysts may wish to use a model that requires an independence assumption or assumes a particular distribution for the data (e.g., a normal distribution). To use the model, they will probably assume that the data meet these requirements (or hope that the modelling technique is robust to the violation of assumptions). Analysts may not question the validity of these assumptions because of some of the social and psychological mechanisms discussed below.

Notice that training and tool bias are related. The connection exists because each of us in our fields have inherent values that predispose us toward particular approaches and methods. For example, knowledge acquisition, a subfield of artificial intelligence (Henrion & 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 & 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. Knowledge acquisition and cultural anthropology favor methods 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 humans prone to such subtle but pervasive biases? Why do we selectively take in data that support what we already know and believe that it can be handled by the approaches, models, or methods we prefer? First, it should be noted that all 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 a Western scientific culture would take the (visible) 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. However, 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 *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 handle the problem leads us to use those tools that we know best and then to believe that they worked. A certain psychological mechanism prevents us from realizing when our beliefs and perceptions do not match, such as when the use of a favored method proves 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 & Byrne, 1981). This tendency inhibits people's ability to update old information in light of new (i.e., Bayesian updating).

Scientists are not immune to this tendency (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 best prevent our own tendency to introduce bias? First, we can strive to remain aware of this tendency. Second, we can select elicitation methods that minimize the role of, and hence the opportunity for interpretation of, the interviewer or knowledge engineer. These methods, used for obtaining data on the expert's answer or problem-solving processes, place the emphasis on the expert. By focusing on learning the expert's thoughts and words and using these to pursue questioning, the data gatherer is less likely to have his or her views intrude (Meyer, Mniszewski & Peaslee, 1989). In addition, the data gatherers can try to act like

a blank slate to avoid translating the expert's data into their own concepts.

Third, the analyst can select analysis methods (e.g., simulation, data-based methods, and non-parametric statistical methods) that require the making of minimal assumptions on the data and avoid fitting the data to restrictive models. We also advocate the use of multiple analysis techniques to cross-validate conclusions and results.

Finally, we suggest that the analyst exercise care in the type of inference (conclusions) that are drawn from this data. Because experts do not provide a random sample of estimates from an underlying population of estimates, statistical inference about that population is not possible. In other words, the experts' estimates cannot be used to make conclusions about the entire true population of values. The inference possible from the expert data is only a general inference concerning the state of knowledge existing at that time by these particular experts.

The Level of Detail in the Data (Granularity) Can Affect the Analyses

The term *granularity* has its origins in fields such as numerical analysis and artificial intelligence. In numerical analysis, granularity refers to the computational grid size used for defining the level at which the computations are made. In artificial intelligence, granularity is defined as "the level of detail in a chunk of information" (Waterman, 1986).

Granularity ranges from coarse (e.g., outlining the basic functions of a nuclear power plant) to fine (e.g., determining the functions of a particular nuclear power plant component). 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 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 testing of a new component, the goal of the problem defines the granularity at the component level. If the goal is to determine the component's performance in a system which is not critical to reactor operations, it would not be necessary to gather information on the component's behavior in multiple environments and conditions. However, if the component is in a critical system, this information and more might be required. The latter goal is at a more specific level, and questions necessary to obtaining the required information 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 also depends upon the complexity of the problem. On simpler questions, such as those whose answers can be verified (e.g., almanac questions), the subject tends to use more structured and detailed problem-solving techniques. Thus, the data from simpler questions are easier for the interviewer to record and for the analyst to model in full detail. On complex problems, the subject's problem-solving information tends to be more plentiful but less structured or clear. The subject and the interviewer may encounter the limitations of information processing mentioned in the previous section, *Experts are Limited in the Number of Things that They Can Mentally Juggle*. The subject resorts 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 subject'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.

Granularity greatly affects the data gathering and aggregation processes. In complex problems, there are many different variables to consider and different data sources to combine. For example, one component in a system is well tested and there is a large amount of information available on its performance. However, all the other components in the system are rare with little known about performance. Therefore, the entire system is a mixture of granularities. To be consistent in the data gathering for this system, the system granularity must be determined by the coarsest level of available information (i.e, the level of the rarest component).

The level of detail greatly affects the analyses, particularly the formation of models, their interpretation, and the drawing of conclusions. For example, different models can be formed from the gathered data, depending on the chosen level of granularity. Typically, the analyst constructs a model whose level of detail depends on the data content of the subject who has provided the least amount of or the most general information.

Granularity is also an issue in the interpretation of the data. Analysts *see* data from their own perspective, which is not necessarily the same perspective as that of the subject from whom 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 preconceived model or hypothesis (tool 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 are highly qualitative, with high uncertainties, as is often the case with expert opinion.

Two studies of interexpert correlation show how granularity affects conclusions (Booker & Meyer, 1988; Meyer & Booker, 1987). In the first study (Booker & Meyer, 1988), which dealt with the problem-solving techniques of statisticians, the experts were asked simply constructed questions. Because their data contained specific problem-solving features, the analyst was able to compare the statisticians using general linear models. The result from this comparison was that experts who used 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 & Booker, 1987), nuclear engineers were asked questions with a more complex structure. The specific heuristics and assumptions that the nuclear engineers used were so varied that the design matrix for use in general linear models was prohibitively sparse. Thus, the problem-solving models had to be constructed at a more general level by combining specific problem-solving features. When these more general models were constructed, which mirrored the ways that the experts processed the 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 detailed level of the first study, no evidence for any interexpert correlation would have been found. Therefore, even though both studies

concluded that experts' answers were correlated according to their problem-solving processes, the models for these problem-solving process had different granularities. The effect of granularity on correlation results ocurred 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 and currently at the University of Kansas, called this relationship to our attention.) In sum, conclusions can differ depending on the granularity of the models chosen.

Because granularity can change at all stages (from question design, to data gathering, to analysis, to interpretation), we recommend that it be carefully monitored throughout these stages. It may not be possible to choose a granularity before the study and keep that level throughout. The expert can change the level and so can the analyst. When granularities change, the most general level should dominate for the remaining stages. The results should always be stated in terms of the granularity used.

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 which include the wording of the problem, the elicitation setting and reference materials made available, 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, and the expert's skill at articulating his or her thoughts. We believe that expert data are highly conditioned on these other factors and this complicates the study of expert data.

The conditioning effect creates problems in both the elicitation and analysis of expert opinion. In the elicitation, the researcher may not have control over the intrusion of factors that influence expert opinion. 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. The conditioning effect also complicates the analysis; the factors often overlap and cannot be separated for analysis of their effects on the expert data (Meyer & Booker, 1987).

The conditioning effect relates to the problem of bias in expert data. Some conditioning effects could be labeled as sources of bias. That is, they lead to an altering of the expert's responses or to opinions that do not obey mathematical and logical 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. In addition, the expert's use of a shortcut in problem solving, such as using the present as a baseline from which to estimate future patterns, could bias his or her answer (Hogarth, 1980).

We recommend a two-step approach for handling the conditioning effect and its offshoot, bias: (1) control those factors that can be controlled, and (2) gather as much data as possible on those factors that cannot be controlled so that the expert data may be either analyzed later for their effect or to annotate the results according to the conditions. 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. In contrast, the

project personnel cannot control for other factors, such as the expert's internal state, personality attributes, and professional background. However, data can be gathered on these factors by asking a series of demographic questions before or after the expert solves the problem. In these two ways, the effects of conditioning can be examined, if not reduced.

In general, we have found structuring to be an effective means of controlling factors and gathering data on those which are not easily controlled. *Structuring* means imposing controls on the elicitation process. It can include presenting the expert with a clear and assimilable statement, 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. Structuring can be done to varying degrees to different aspects of the elicitation process. In general, 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.

A structured elicitation process can be more easily monitored for the intrusion of various factors. For example, a question is frequently structured by decomposing it into is component parts. Question decomposition eases the cognitive burden of solving complex problems and has been found to lead to more accurate answers (Armstrong, Denniston & Gordon, 1975; Hayes-Roth, 1980). When a question is decomposed, the expert provides estimates on each part and the data gatherer can control or record conditioning factors for each part. In general, we have found monitoring for bias easier when experts verbalize their thoughts and answers.

GENERAL RECOMMENDATIONS

In addition to the specific recommendations given in the above six common problems, we recommend a general approach to elicitation and analysis as a means of avoiding or reducing problems in expert opinion.

We advocate that the elicitation be designed to fit the experts and the way that humans think rather than force the experts to adapt to convenient or standard methods. We propose the research on human limitations and tendencies toward bias be taken into account in selecting the methods. For example, if the interviewer selecting the elicitation methods does not consider people's limitations in comparing more than 7 things at once, the resulting data will be less credible. If in the former case the expert estimates are being used to develop a model or decision process, there is the danger of *garbage in, garbage out*.

We also advocate the practice of eliciting as much of the information on the expert's problem-solving processes as possible. 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 & Meyer, 1988; Meyer & Booker, 1987). The expert's definitions and assumptions frequently explain how the expert arrived at one particular answer and not another. In addition, problem-solving data will prove valuable later if multiple expert's estimates are to be mathematically combined to form a single estimate. The expert data can also 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 problem-solving process allows the opinions to be more

easily updated as new information becomes available.

We also suggest controlling for the factors that can enter into the elicitation process and influence the expert's problem-solving process. For example, the phrasing of the problem, the interviewer's responses, and other participant's responses can affect the answer an expert reaches. For those influences that can not be easily controlled, such as the expert's tendency to anchor to his or her first impression, we recommend gathering as much data as possible to analyze their effects.

We recommend using the decomposition principle to obtain the best estimates from the expert, to minimize biases, and to help monitor granularity and conditions affecting the answers.

The approach to analysis that we recommend complements the elicitation philosophy mentioned above. Just as the elicitation approach allows the experts' capabilities to shape the data-gathering methods, the analysis philosophy should allow the data to dictate which analytic methods are appropriate. 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 analyst avoids (where possible) blindly assuming particular properties of expert opinion (e.g., that the expert opinion data is normally distributed, that the answers of multiple experts are independent, or that the experts are perfectly calibrated). Instead, we suggest that the analyst use 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 do not depend on an assumed distribution of the data.

We also recommend using a variety of methods to address the multivariate structure of expert opinion data. The data is multivariate because it includes answers to multiple problems, information on conditioning factors, information on the experts' problem-solving processes, and information about the experts' backgrounds. Therefore, the data is a mixture of qualitative and quantitative information. Multivariate analysis techniques allow the simultaneous consideration of two or more variables of interest. Many of these techniques accommodate the mixture of qualitative and quantitative data types. Thus, they can be used to investigate some of the more important properties of the data, such as the dependence of experts and identifying important conditioning factors. Caution is required in the use of standard statistical multivariate techniques because these have strict assumptions about the data. Again, simulation and some data-based techniques such as the bootstrap may be more appropriate for expert opinion data analysis.

REFERENCES

Armstrong, J. S. (1981). *Long-Range Forecasting: From Crystal Ball to Computer.* New York, New York: Wiley-Interscience.

Armstrong, J. S., Denniston, W. B., Jr. & Gordon, M. M. (1975). Use of the decomposition principle in making judgments. *Organizational Behavior and Human Performance,* 14, 257-263.

Ascher, W. (1978). *Forecasting: an Appraisal for Policymakers and Planners.* Baltimore, Maryland: John Hopkins University Press.

Baron, R. A. & Bryne, D. (1981). Understanding human interaction. *Social Psychology.* Boston, Massachusetts: Allyn and Bacon Inc.

Booker, J. M. & Meyer, M. A. (1988). Sources and effects of interexpert correlation: an empirical study.

IEEE Transactions on Systems, Man, and Cybernetics, **18**, 135-142.

Dougherty, E. M., Jr., Fragola, J. R. & Collins, E. P. (1986). Human reliability analysis. Science Applications International Corporation Report SAIC/NY-86-1-OR, Oak Ridge, Tennessee.

Egan, D. E. & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory and Cognition*, **7**, 149-158.

Faust, D. & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science*, **241**, 31-35.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Palo Alto, California: Stanford University Press.

French, S. (1986). Calibration and the expert problem. *Management Science*, **32**, 315-321.

Harrison, J. M. (1977). Independence and calibration in decision analysis. *Management Science*, **24**, 320-328.

Hayes-Roth, B. (1980). Estimation of time requirements during planning: interactions between motivation and cognition. Rand Corporation Report N-1581-ONR, Santa Monica, California.

Henrion, M. & 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*, July, Seattle, Washington, pp. 471-476.

Hogarth, R. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association*, **70**, 271-291.

_____ (1980). *Judgement and Choice: The Psychology of Decisions*. Chicago, Illinois: Wiley-Interscience.

Kahneman, D. & Tversky, A. (1982). Subjective probability: A judgment of representativeness. In Kahneman, D., Slovic, P. & Tversky, A. Eds., *Judgment Under Uncertainty:* Heuristics and Biases, pp. 32-47. Cambridge, Massachusetts: Cambridge University Press.

Keeney, R. L. & von Winterfeldt, D. (1989). On the uses of expert judgment on complex technical problems. *IEEE Transactions on Engineering Management*, **36**, 83-86.

Lichtenstein, S., Fischhoff, B. & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky Eds., *Judgment Under Uncertainty: Heuristics and Biases*, pp. 306-334. Cambridge, Massachusetts: Cambridge University Press.

Mahoney, M. (1976). *The Scientist as Subject: The Psychological Imperative*. Cambridge, Massachusetts: Ballinger Publishing Co.

Meyer, M. A. & Booker, J. M. (1987). Sources of correlation between experts: empirical results from two extremes. Los Alamos National Laboratory Report LA-10918-MS, Los Alamos, New Mexico and Nuclear Regulatory Commission Report NUREG/CR-4814, Washington, D.C.

Meyer, M. A. & Booker, J. M. (1989). Eliciting and analyzing expert judgment: A practical guide. Los Alamos National Laboratory Report LA-11667-MS, Los Alamos, New Mexico and Nuclear Regulatory Commission Report NUREG/CR-5424, Washington, D.C.

Meyer, M. A., Booker, J. M., & Bradshaw, J. M. (1990). A flexible six-step program for defining and handling bias in knowledge elicitation. In B. Wielinga et al. Eds., *Current Trends in Knowledge Acquisition*, pp. 237-256. Amsterdam, The Netherlands: IOS Press.

Meyer, M. A., Mniszewski, S. M. & Peaslee, A. T., Jr. (1989). Using three minimally biasing elicitation techniques for knowledge acquisition. *Knowledge Acquisition*, **1**, 59-72.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, **63**, 81-97.

Morris, P. A. (1986). Observations on expert aggregation. *Management Science*, **32**, 321-328.

Ortiz, N. R., Wheeler, T. A., Meyer, M. A. & Keeney, R. L. (1988). The use of expert judgment in NUREG-1150. *Proceedings of the 16th Water Reactor Safety Information Meeting.* Oct., Gaithersburg, Maryland. In Nuclear Regulatory Commission Report NUREG/CP-0097, Vol. 5, pp. 1-25.

Saaty, T. L. (1980). *The Analytic Hierarchy Process: Planning, Priority Setting, and Resource Allocation.* New York: McGraw-Hill.

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), Vol. 1-2, second draft for peer review, Nuclear Regulatory Commission Report NUREG-1150, Washington, D.C.

Waterman, D. A. (1986). *A Guide to Expert Systems.* Reading, Massachusetts: Addison-Wesley Publishing.

Wheeler, T. A., Hora, S. C., Cramond, W. R. & Unwin, S. D. (1989). Analysis of core damage frequency from internal events: expert judgment elicitation, Vol. 2. Sandia National Laboratories Report SAND86-2084, Albuquerque, New Mexico and Nuclear Regulatory Commission Report NUREG/CR-4550, Washington, D.C.

Winkler, R. L. (1986). Expert resolution. *Management Science,* 32, 298-303.