THE QUALITY OF EXPERTS' PROBABILITIES OBTAINED THROUGH FORMAL ELICITATION TECHNIQUES

Prepared for

Nuclear Regulatory Commission
Contract NRC-02-88-005

Prepared by

Robert L. Winkler
Stephen C. Hora
Robert G. Baca

Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas

September 1992

CONTENTS

CONTENTS (Cont'd)

FIGURES

TABLES

ACKNOWLEDGMENTS

This report was prepared to document work performed by the Center for Nuclear Waste Regulatory Analyses (CNWRA) for the U.S. Nulcear Regulatory Commission under Contract No. NRC-02-88-005. The activities reported here were performed on behalf of the NRC Office of Nuclear Material Safety and Safeguards, Division of High-Level Waste Management. The report is an independent product of the CNWRA and does not necessarily reflect the views or regulatory position of the NRC.

1 INTRODUCTION

1.1 OBJECTIVE

Expert judgment is expected to play an important role in several activities related to the characterization, design, and evaluation of the proposed high-level waste repository. For the purpose of postclosure performance assessments, formal techniques for eliciting expert judgments may prove to be very useful in developing: (i) subjective probabilities for hydrogeologic events, conditions, and processes; and (ii) probability curves for parameters representing site and waste form characteristics (e.g., hydraulic properties, sorption coefficients, and solubilities).

Obtaining probability information through elicitation techniques is motivated largely by one basic and compelling fact. That is, in the earth sciences, the time and space scales are so large that obtaining sufficient quantities of data to compute a probability estimate (i.e., relative frequency) or probability curve will not, in general, be feasible. In addition, the current scientific knowledge about coupled geologic, hydrologic, thermal and geochemical phenomena is very limited. The elicitation approach permits the formal incorporation in a performance assessment of the information that is available. This information is primarily in the form of expert judgments.

While formal elicitation techniques are well developed and have been extensively used in decision analysis and risk analysis applications (primarily in business, but also in the public sector), relatively few applications have been made in the earth sciences. One of the fundamental concerns about using these elicitation techniques is the question: "How reliable is expert judgment?" In the terminology of the expert judgment analyst, the question is one evaluating the "quality" of the expert's judgment. Other concerns about expert elicitation include such aspects as how the experts are selected, how they are trained, approaches for avoiding bias, and aggregation of expert judgments.

This report focuses on the issue of the quality of experts' probabilities as obtained through a formalized elicitation process. A broad review of the scientific literature was performed to survey studies where expert judgment was compared with actual data. The review included the fields of:

- Science and engineering,
- Risk analysis,
- Weather forecasting,
- Medicine,
- Business, and
- Psychology.

1.2 REPORT ORGANIZATION

This brief report is organized into five major sections. In Section 2, the general steps in the expert elicitation process are outlined. Various interpretations of probabilities are presented and explained in Section 3. In Section 4, the results of the literature review are presented. An appendix is also included which describes various procedures for evaluating experts' probabilities.

2 OBTAINING EXPERT JUDGMENTS AS PROBABILITY DISTRIBUTIONS

2.1 INTRODUCTION

Expert judgment can provide important information in decision analysis, risk analysis, and policy analysis to supplement other information that might be available. For example, important sources of information for the performance assessment of nuclear waste disposal systems include the results of experiments and observations, mathematical/computer models of physical and geochemical processes, and expert judgments. Although expert judgments may take many forms, such as recommendations and criticisms, scenarios, value judgments, model-building choices, and estimates, the discussion in this report is limited to the acquisition of judgments in the form of probability distributions.

The evaluation of risks for the purpose of policy and decision making has led to the development of formal methods for the collection of expert judgment through the elicitation of probabilities. Examples of the use of such methods and discussions of such methods include Morgan et al., (1984), Electric Power Research Institute (EPRI) (1986), Richmond (1987), Bonano et al., (1989), Hora and Iman (1989), Keeney and von Winterfeldt (1989, 1991), Merkhofer and Runchal (1989), Nuclear Regulatory Commission (NRC) (1989), Whitfield and Wallsten (1989), Morgan and Henrion (1990), Cooke (1991a), and Whitfield et al., (1991). Judgments given as probabilities are often called subjective probabilities or degrees of belief to distinguish them from relative frequencies and other interpretations of probability.

2.2 ACQUISITION AND USE OF EXPERT JUDGMENT

The acquisition and use of expert judgment encoded as probabilities has long been a vital part of decision analysis (e.g., Raiffa, 1968; Spetzler and Stael von Holstein, 1975). Many of the techniques and procedures developed and used in decision analysis are also used in risk assessment and performance assessment. However, special aspects of the latter situations necessitate different emphasis on some issues. The primary focus may not be on a particular decision, but on representing the state of information about something and characterizing certain risks of interest. Moreover, the public nature of most risk assessment problems dictates extra care to provide a justifiable and well-documented process. Otway and von Winterfeldt (1992) discuss the importance of making the use of expert judgments more formal, explicit, and documented.

Expert judgment is often used to obtain probability distributions for uncertain quantities and probabilities of potential events in order to assess risks. Expert judgment, however, should not be viewed as a substitute for experimentation, modeling, and observation. Instead, it complements these activities. For example, when alternative sources of information are available but conflict, the judgments of experts may be the preferred method for integrating the information into a single probability distribution. Experts may be able to assess the uncertainty inherent in the various sources of information and, additionally, may be able to make adjustments to account for biases in the data. Experts therefore may provide a "calibrating" and "integrating" mechanism to account for differences in applications, environments, and other factors.

Scaling up from an experiment to a projection of the behavior of the real process is another problem. The geochemistry of experiments conducted in a laboratory environment is less complicated

than the real systems. Estimating solubilities in a laboratory and then scaling up to a nuclear waste disposal system involves making a transition from a laboratory experiment with a relatively simple environment to a system that is spatially differentiated by many microenvironments. Expert judgment is one option to make this transition.

Expert judgments may change with time, of course. This is only to be expected. As new information becomes available, the state of knowledge is modified and opinions change. The role of the experts is to summarize the available information and to express both what is known and what is not known. Of course, probabilities are the natural medium for expressing such uncertainties. Moreover, expressing judgments in terms of probabilities allows these judgments to be combined with other sources of information, thus evoking the power of mathematical manipulation. Such manipulations are not possible when judgments are given qualitatively. A final point is that experts' probabilities can help to pinpoint where additional information (more data collection or experimentation, further model-building, consulting different experts) can be valuable.

2.3 THE SELECTION OF EXPERTS

Who is an expert? An expert may be someone who has special skills and training resulting in superior knowledge about a particular field and access to that knowledge (Bonano et al., 1989). The identification of experts is an important stage in the process of acquiring expert judgments. Quite aside from the danger of selecting an expert who is not "well qualified," performance assessment is in the public view. Therefore, because the stakes go beyond science, the criteria and process of selecting experts become critical.

Experts can be identified through literature searches, registries of professional organizations, consulting firms, research laboratories, government agencies, and universities. A formal nomination process is sometimes used, particularly when controversy is possible. The nomination process should be designed to preclude bias in selection. A first step is inviting stakeholders and interested parties to nominate experts. A second step is using an independent external selection panel to evaluate the nominees.

The criteria for selection should be specific and documented, including:

- Evidence of expertise, such as publications, research findings, degrees and certificates, positions held, awards, etc.;
- Reputation in the scientific community, including knowledge of the quality, importance, and relevancy of the nominee's work and the nominee's ability to provide the desired judgments;
- Availability and willingness to participate;
- Understanding of the general problem area;
- Impartiality, including the lack of an economic or personal stake in the potential findings; and
- Inclusion of a multiplicity of viewpoints.

Biases from economic or other stakes can affect an expert's judgment. Yet, excluding an expert because of potential bias may prevent relevant information from being discussed. A solution to this dilemma was used in NUREG-1150 (Wheeler et al., 1989; Hora and Iman, 1989), where potentially biased authorities were allowed and encouraged to give testimony. In one instance, an employee of a pump manufacturer testified on the performance of a pump under severe stress. The employee had access to relevant information that the members of the expert panel did not have. However, inclusion of the employee on the expert panel could have been perceived as a conflict of interest. Presentation of the employee's testimony allowed the expert panel to judge the value of the testimony.

2.4 PREPARATION FOR PROBABILITY ELICITATION

The elicitation process is much more than just asking experts to assess some probabilities. The experts must prepare and be prepared for the experience. Often, experts in a substantive field such as geology or health may not be experienced in expressing their beliefs in the form of probability distributions and may not understand why their probabilities are of interest and how they will be used. Training the experts, then, is a crucial step in the process.

It is important that analysts with expertise and experience in eliciting experts' probability judgments design and conduct the elicitation process (Keeney and von Winterfeldt, 1991). The analysts' expertise involves probability elicitation, in contrast to the experts' substantive knowledge regarding the events or variables of interest. Since the analysts will plan and direct the process, the selection of qualified analysts is just as important as the choice of substantive experts. As Keeney and von Winterfeldt point out, in complex technological problems it may also be helpful to have the analysts joined by individuals with general knowledge about the field of interest (e.g., geology or seismology) to help in structuring the problem and communicating with the experts.

Training the experts has multiple objectives. One is to motivate the experts and provide an overview of the process, including how the experts' judgments will be used. Another objective is to develop the experts' confidence in their ability to express their judgments as probabilities and in the entire process. Lack of understanding or confidence can undermine the entire effort. Yet another objective is to assure that the experts have access to relevant background information and evidence specific to the questions of interest (reports of pertinent studies, etc.) and sufficient opportunity to review this information and think about its implications for the elicitation.

Motivating the experts is an important aspect of the process. Experts may object to the formal elicitation of judgments as probabilities because they believe that "opinion" is being substituted for "objective" scientific research. However, the experts' role is not creating new knowledge, but synthesizing disparate and often conflicting interpretations of information to produce an integrated picture. Experts who appreciate their role from this perspective are likely to be cooperative and to find the entire process enjoyable and educational.

The fundamental objective of elicitation training, of course, is to help the experts learn to express their judgments in probabilistic form. Training introduces experts to the notion of probability and to methods for probability assessment. It should also instill awareness of potential cognitive and motivational biases and suggestions for how to counteract such biases. Finally, it provides practice at actual probability assessment. Evidence suggests that practice improves elicitation, as will be discussed in subsequent chapter.

The description of the issues should be accompanied by information to assist the expert. Included should be references and data sources pertinent to the issues. This approach was used in a study of the relation of $SO_4^=$ concentrations and mortality (Morgan et al., 1984), where two-page summaries of articles were prepared for the experts. In Environmental Protection Agency (EPA) studies of health risks associated with exposure to lead (Whitfield and Wallsten, 1989), and ozone (Whitfield et al., 1991), documents summarizing relevant scientific evidence and some useful calculations were provided to all of the experts. In the NUREG-1150 study (Hora and Iman, 1989), large numbers of reports and studies were provided to the experts.

An expert may have depth of knowledge in a related field, but perhaps limited knowledge about the specific issues of interest. In recent work on the solubilities of transuranic elements in brine (Trauth et al., 1991) done for the Waste Isolation Pilot Plant (WIPP) performance assessment, one of the experts was from the field of ocean chemistry. This expert had substantial knowledge of solubilities, but he did not have direct experience with the brines expected to fill the mined salt repository after closure nor was the expert initially knowledgeable about the relevant chemistry, including the pH and the ionic strength of the brine, in the repository. To participate effectively, this expert required substantial information about the chemistry expected within the repository.

Experts should not be asked questions without allowing time to study the issues and consider possible answers. A good procedure is to introduce the experts to the questions to be addressed, provide background briefings, and allow an interim study period to consider the issues. Such a period of study could range from several days to several months depending upon the number and complexity of the issues. During this time the experts might ask for additional information (e.g., data sets, articles, reports). At the end of the study period, the experts should be prepared to state and defend the rationales supporting their elicited judgments.

2.5 THE ELICITATION OF EXPERT JUDGMENT

An important step in an expert judgment process is identifying the issues to be addressed and formalizing the quantities to be elicited. While this may appear to be a simple and obvious step, experience has shown that developing questions for experts that are mutually understood by those asking the questions and those answering the questions can be a difficult task.

Achieving an accurate, logically complete, and understandable description of an issue to be addressed by experts is critically important. The description of the issue under assessment must be complete and without unstated assumptions; everything that can be disagreed upon must be made explicit. Spetzler and Stael von Holstein (1975) have suggested the test of "clairvoyance." If, after reading the description of the issue, a "clairvoyant" would be able to answer the question without asking for any additional information, the description is complete. Often, analysts will make contextual assumptions that are not obvious. Conversely, experts often make assumptions that were not intended by the analysts.

A principle espoused by Winker et al., (1978) is that of asking only questions about observable, or at least theoretically observable, quantities. To understand this assertion more fully, consider a situation where a two-dimensional (2D) model is used to predict the transport of fluids through the earth. Perhaps one parameter in this model is the spacing of fractures. In reality, fractures come in many sizes, lengths, and orientations. Spacings vary greatly between nearest neighbors and are not constant along the length of fractures for any pair of fractures. The parameter in the model is, then, a convenient

summarization that has no physical meaning and cannot be measured, even conceptually. In this situation, the experts should be asked questions about physically-measurable quantities, such as average aperture or fracture permeability. From these responses the uncertainties about the model parameter can be derived.

The questions for probability elicitation should be presented in such a way as to reduce tendencies toward bias on the part of the expert. Experts can be biased by presentations of issues that hint at or suggest a particular answer. Also, awareness of possible cognitive biases (Kahneman et al., 1982) should stimulate efforts by the analyst (and, to some degree, the expert as well) to structure the questions in an attempt to avoid such biases. For example, probing the extremes of an expert's probability distribution before asking questions about the middle part of the distribution helps to counter any tendency to anchor on central values.

2.6 INTERACTION AMONG EXPERTS

There are alternative approaches to organizing a group of experts. These approaches vary with the respect to the scope of the issues being addressed, the amount and type of interaction among the experts, the amount of redundancy, and the role of the experts in defining objectives.

The simplest organization is either one expert or several experts working in isolation from each other. When there are several experts addressing the same issues, there is some useful redundancy because multiple experts' alternative viewpoints increase the potential for representing the appropriate range of uncertainty. The difficulty with isolated experts is that information is not shared, thus reducing the opportunity for learning through the exchange of information.

When there are multiple experts addressing the same questions, panels may be organized in which experts work together sharing information and approaches to the issues. The study of nuclear reactor safety (Hora and Iman, 1989) allowed for two meetings of the experts, with an intervening study period of about eight weeks. Issues and background information were presented in the first meeting while the experts presented their analysis in the second meeting.

Merkhofer and Runchal (1989) and Whitfield et al., (1991) employed a combination of individual elicitation procedures and group interaction. After judgments were assessed independently in probabilistic form by each expert, the resulting probability distributions and rationales were shared among the experts, who were then allowed to revise their individual probabilities.

Panels may provide a "group assessment" or individual assessments; each has advantages. When experts work together and obtain a group assessment, there is no need to combine assessments. However, group interactions can sometimes be dysfunctional, with some individuals dominating the discussion. Also, opinions may vary so greatly that a consensus cannot be reached, in which case it may be necessary to combine distributions mechanically using a quantitative rule such as the arithmetic or geometric mean. In contrast, when experts provide individual assessments, the potential is better for capturing the full uncertainty about the issue, and provide a traceable and auditable basis for individual opinions (as opposed to opinions of an anonymous group). However, the combination of the disparate probabilities may be difficult in certain cases.

Isolated panels of experts are efficient with respect to the experts' time but require coordination. The views, assumptions, and findings of a panel may shape the issues addressed by another. When coordination between experts or panels of experts is necessary, the work may be accomplished sequentially. An example of such a situation is the coordination between a panel of experts judging how fast and how high pressure will rise in a reactor containment vessel, while a second panel considers the pressure at which the vessel will fail. Knowing the relevant range of pressure-rise parameters will help establish the assumptions for the second panel's assessments of containment failure probabilities.

Another strategy for the analysis of complex issues consists of using multi-disciplinary teams of experts. This approach is relevant when the issue to be addressed is difficult to decompose into a series of smaller independent or conditional issues. A difficulty with decomposition is that there are significant linkages or information requirements among the issues. Redundant teams were used in the EPRI (1986) study of seismicity in the Eastern United States. A main limitation of using several teams is that their organization is difficult. A second problem is bringing the team members together to perform their analyses. This is costly. In the seismicity study, the teams were formed within companies and institutions to facilitate communication. Each team was allowed some flexibility in determining how to decompose the problem into the individual experts' specializations. Coordination assured that the assessments made by the various teams would be compatible with the overall study objectives.

2.7 PROCESSING JUDGMENTS

The goal of processing judgment is to: (i) produce a usable product for the ensuing analysis; and (ii) to preserve intact the expert's judgments. Judgments often require some processing to put them in a usable form. Assessments obtained using indirect methods, for example, must be translated into probabilities or densities. Distributions for continuous quantities are most often assessed by obtaining several points on the distribution function and then fitting and/or interpolating to obtain the remainder of the distribution. Frequently a member of a certain family of distributions, such as the normal family, is used as an approximation to the experts' judgments.

Occasionally, it may be necessary to extrapolate beyond the given range of values. This can occur when the expert has provided values, such as the 0.01 and 0.99 quantiles, but it is necessary to obtain the endpoints of the distribution. These values are often the most important in the analysis because they lead to the most serious consequences.

Another type of processing is the aggregation of judgments from multiple experts (see Sections 5 and 6). Aggregation is often justified by one or more reasons (Bonano et al., 1989):

- An aggregated distribution provides a better appraisal of knowledge than the individual distributions (a sample mean is better than one observation);
- The aggregated distribution is sometimes thought of as representing some sort of consensus; and
- It is easier to use a single distribution in further analysis.

However, when judgments are combined, the individual judgments should be retained to show the range of opinion on an issue, as well as provide a traceable record of individual judgments.

2.8 DOCUMENTATION

Regardless of how well an expert judgment elicitation process is designed and implemented, adequate documentation is required. The entire expert elicitation process should include documentation of the procedures and criteria for selecting experts and issues, descriptions of the elicitation issues, copies of all supporting materials, and the results of the elicitation sessions. Most importantly, the detailed rationales for the assessments, the methods, and results of any post-elicitation processing of the judgments or re-elicitations by the experts should be provided. Moreover, as new evidence becomes available, understanding the rationale for probability distributions will allow the judgments to be reinterpreted instead of being discarded. For example, Sandia National Laboratories (SNL) (Camp et al., 1990) has undertaken the updating of some distributions obtained in the NUREG-1150 study. Without explicit rationales, updating these distributions would be difficult.

2.9 SUMMARY

Although expert judgment pervades all scientific inquiry, it is often disguised as implicit assumptions while the primary attention focuses on data, models, and analysis (all of which require expert judgment to plan, conduct, and interpret). In the WIPP study of waste isolation, for example, a great deal of time and effort has been spent on the geology and hydrology of the area. Extensive computer algorithms have been constructed and preliminary risk analyses made. Only recently, however, has the focus shifted to the effect of human intrusion, an area where the physical sciences have less to say. Yet human intrusion is apt to be the dominant contributor to risk. Those risks that are analyzed using expert judgment are often the most crucial risks, but the effort applied to obtaining good expert judgments is small when compared to efforts elsewhere.

The elicitation of expert judgments in the form of probabilities makes the role of expert judgment explicit and impossible to ignore. Moreover, it is just as important to design a process for the elicitation of expert judgment carefully as it is to design a scientific experiment carefully or to set up a model-building exercise carefully. Collecting expert judgments in an arbitrary manner or failing to give due consideration to what is known about the assessment of probabilities from experts makes the process vulnerable to attack. When expert judgment is used, a formal process developed by analysts experienced in eliciting expert judgments is important.

The discussion in this chapter has involved important aspects of a formal process for the elicitation of expert judgments. This suggests that evaluating a set of expert judgments can be accomplished to a large extent by looking at the process used to obtain the judgments. Aspects deserving careful attention include:

- The selection of experts;
- The selection of analysts to design and conduct the elicitation process;
- Attempts to motivate the experts in an impartial manner;
- The training of the experts;
- The opportunity for practice elicitations;

- The development of definitions of the issues, events, and variables of interest;
- The distribution of background materials to the experts;
- The allowing of ample time for experts to study the materials, think about the issues, and ask questions;
- Possibly the opportunity for experts to interact to exchange information;
- The use of accepted formal probability elicitation techniques;
- The recording of the rationales (informal arguments, theory, data models) underlying the expert judgments;
- Justification for any processing of judgments; and
- Possibly the opportunity for post-elicitation interaction among experts and/or feedback regarding the elicited probabilities and rationales, followed by the opportunity for revising judgments.

Documentation of each step of the process is essential. The process may proceed sequentially, with some steps repeated again after new evidence is obtained or the experts have had the opportunity to interact. Care and attention to all of these aspects of an expert judgment process should result in a successful, justifiable process that produces valuable information for uncertainty analysis and performance assessment.

3 PROBABILITY AND EXPERT JUDGMENT

3.1 INTRODUCTION

When most people are exposed to the concept of probability, it is within the framework of a general theory or mathematical system. In this context, probabilities are simply numbers between zero and one, inclusive, that satisfy certain rules. Given certain probabilities, the theory of probability can be used to determine certain other probabilities. Many books have been written and many courses are taught focusing strictly on mathematical properties of probabilities. Just viewing probability as a mathematical system, however, leaves a fundamental and crucial question unanswered: What do these probabilities really mean and how should they be interpreted? Clarity on this interpretation is important not just from the standpoint of the analyst or decision maker who must draw conclusions or make decisions based on the experts' judgments, but also from the viewpoint of the experts themselves. Part of the process of training experts to assess probabilities is a review of the appropriate interpretation of those probabilities.

The purpose of this section is to clarify the appropriate interpretation of experts' probability judgments in the context of uncertainty analysis and performance assessment. First, a review of the major interpretations of probability is presented. The question of the appropriate interpretation for experts' judgments is then considered, followed by some comments on the search for "true probabilities."

3.2 INTERPRETATIONS OF PROBABILITY

Philosophers, mathematicians, statisticians, and others have devoted a great deal of effort over the years to discussions of interpretations of probability. This has been an issue of great debate, with the concerns ranging from philosophical/foundational to very applied/practical. It has had a strong influence on the development of statistics and on the way people try to apply the mathematical theory of probability in the real world. Although there are different nuances and offshoots of theories, the three main interpretations of probability — objective (classical, frequency), and subjective — will be discussed here.

3.2.1 Objective Probabilities

The classical interpretation of probability originated in the study of games of chance in Europe in the eighteenth century. Consider, for example, the probability that when tossed, a six-sided die lands with the side having three dots facing up. Most people would say that this probability is one-sixth. The classical argument is that by virtue of the apparent symmetry of the die, each of the six possible outcomes is equally likely. If the possibility of the die landing on edge or disintegrating in mid-air is ignored, and there is no reason to believe any one side is more likely to occur than any other, then each side must have probability one-sixth. According to the classical view, then, the probability of an event is equal to the number of outcomes comprising that event divided by the total number of outcomes. For example, the probability that a die yields an even number when tossed is 3/6, since there are six possible numbers and three of them are even. Aside from questions of counting the number of outcomes in complex situations, the classical interpretation is straightforward and easy to apply. The underlying assumption of equally likely outcomes is, however, a very strong assumption that is reasonable only in a limited set of practical applications.

The frequency interpretation is empirical in spirit, considering not just the possible number of outcomes, but past evidence regarding the event of interest. If a person is contemplating making a bet on the toss of a thumbtack instead of the toss of a die, there are two possible outcomes, the thumbtack landing point up or point down (as with the die, the possibility of landing on edge with the point straight out horizontally from the top of the tack is ignored for the purposes of this discussion). However, arguments that these two outcomes are equally likely are not particularly compelling. The person contemplating a bet might find it useful to see the results of an experiment in which the thumbtack was tossed 1000 times. If, for instance, it landed point up 420 times, the relative frequency of occurrence of "point up" would be 0.42. In the frequency interpretation, the probability of an event is the limiting relative frequency of occurrence of that event in a series of independent, identical trials as the number of trials approaches infinity. The law of large numbers shows that this relative frequency will approach the probability of the event. As with the classical interpretation, the relative frequency interpretation is generally straightforward. In terms of actual implementation, the key issue is the availability of appropriate data. Many events of interest are unique in the sense that no past record of identical, independent trials exists. Sometimes no data are available, and other times the data are not from "identical" situations.

3.2.2 Subjective Probabilities

In the subjective interpretation of probability, a probability is viewed as an individual's degree of belief that an event will occur. If an engineer is interested in the probability of an accident at a particular nuclear power plant in the next year, it is difficult to think in terms of the classical or frequency interpretations of probability. There is no reason that the possible outcomes need be equally likely, and the evidence from repeated, independent, identical trials is not available. It is possible to look at the past relative frequency of accidents at this and similar installations, but each plant may have some unique features and things may change from year to year as equipment ages and personnel change. Here the probability appropriate for the engineer is his or her degree of belief, which is based on any relevant information that is available (historical evidence, knowledge about the reliability of equipment, changes in personnel, and so on).

In terms of application, the classical and frequency interpretations are more restricted than the subjective interpretation. The classical interpretation requires an assumption that the outcomes are equally likely, something that might be reasonable in certain games of chance, but is not very reasonable in most real-world situations. The frequency interpretation needs data from a reasonably large sample of independent, identically distributed trials to be of most use. The insurance industry is one place where such data are available, with actuarial tables being based on large samples and providing the basis for insurance rates. The subjective interpretation has no such restrictions. However, it is limited by the fact that different experts may (and no doubt will in many cases) have different probabilities for the same event.

In a very real sense, the subjective view subsumes the other two interpretations. Ultimately, the decision as to whether the assumptions of the classical or frequency interpretations are applicable in a given situation is a subjective choice. An individual contemplating a game of chance may feel that the game is fair, with equally likely outcomes, or may feel that it is not fair (e.g., a die may be "loaded"). Even with data, the choice of an appropriate "reference set" of data is subjective. If the probability that a given person dies in the coming year is of interest, what is the appropriate reference set of past data to use in trying to apply the frequency interpretation? Is it all people, or all people of the age of the

individual in question, or all people of the same age and sex, or all people of the same age and sex and general health (assuming this can be measured and is in the data base)? Also, are past data applicable if a new cure has just been found for a disease that has been a major cause of death for people in the age group in question? Ultimately, these questions must be answered subjectively. Furthermore, in many applications the available data are sparse and even further removed from the question of interest than in this example.

3.3 EXPERTS' PROBABILITIES

The fact that experts' probabilities are subjective, while perhaps unappealing to some, is undeniable. The types of events or variables of interest in a uncertainty analysis and performance assessment do not meet the assumption of equally likely outcomes, and frequency data are either not available or form only part of the set of available information. The very use of the expression "experts' probabilities" indicates that the probabilities are associated with the experts who are consulted.

Some individuals feel uncomfortable with probabilities that are subjective, or soft, as opposed to so-called "objective" probabilities based on "hard data." Scientists in particular have been trained to avoid subjectivity. More will be said on the question of "objective" probabilities in the next section. The point of interest here is that for those who would like to base their probability judgments on hard data, there is the problem of what to do in case of sparse data. In a performance assessment of a geologic repository, the probability of an earthquake exceeding magnitude six within a 10,000-year period might be of interest. Direct data regarding the occurrence of earthquakes in this or similar locations are too limited to be of much use. Taking into account general scientific knowledge about seismology and specific knowledge about the location, experts can form judgments about the possibility of future earthquakes. But these judgments are subjective. Should they be ignored? Those who think so are falling into the trap noted by Savage (1954), who drew an analogy with building on sand. Those disdaining subjective information are, in effect, saying "Take away the sand, we will build on the void."

Moving to the question of implementation, the measurement of probability and the fact that probabilities as so measured obey the mathematical rules of probability are consistent with the classical and frequency views. But subjective probability also has very solid mathematical foundations, with important contributions from, among others, Ramsey (1931), de Finetti (1937), and Savage (1954). Some key work is presented in a book of readings edited by Kyburg and Smokler (1964). The important message is that under certain axioms of coherence, the existence of objective probabilities that obey the usual mathematical rules of probability can be shown. These axioms are generally viewed as reasonable; for example, if an individual judges A to be more likely than B, and B to be more likely than C, then he or she must judge A to be more likely than C.

In terms of measurement, it is possible to simply ask experts for their probabilities for events of interest. Alternatively, several devices are available to help the experts think about and assess their probabilities. For example, meteorologists can be asked whether they would rather receive $100 contingent on rain tomorrow or contingent on a red ball being drawn in a random drawing from an urn with 50 red balls and 50 white balls. If the urn is chosen, the meteorologist's probability of rain is inferred to be less that .50 (50/100); if the payoff contingent on rain is chosen, P(rain) must be greater than .50; if the meteorologist is indifferent, then P(rain) is taken to be equal to .50. By varying the proportion of red balls in the urn until the meteorologist is indifferent between betting on rain and betting

on red, it is possible to assess the meteorologist's probability of rain. Similar devices such as probability wheels are also available.

Techniques for assessing or eliciting probabilities represent only part of an overall approach to obtaining experts' probability judgments. This approach, typically includes some training in probability, some practice probability assessments, careful definition of the events or variables of interest, assessments of probabilities for those events or variables, checks for consistency with the rules of probability, and reviews of the assessments and revisions where deemed appropriate. For a general discussion of probability assessment, see Spetzler and Stael von Holstein (1975).

3.4 THE NOTION OF "TRUE" PROBABILITIES

The desire for objective probabilities is fueled in part by the notion that for a given event, there is a "true" probability. This leads to the idea of trying to estimate this true probability. Think, however, about the probability of a serious accident at a nuclear power plant. It is reasonable to think that at a given point in time, there is some true probability that an accident will occur in the coming year? What would be the basis for that true probability? Thinking about all of the possible causes of accidents (including both human causes and natural causes), the idea that there is a true probability of an accident seems not scientific but rather almost a religious belief that everything is preordained. Is there a true probability that a group of terrorists will blow up the nuclear power plant? That a worker will turn the wrong dial and thereby cause an accident? That an earthquake nearby will occur and will be strong enough to cause an accident? That an amateur pilot will lose control of his plane and crash into the power plant? This list could go on and on, and for any of these scenarios, there is no true, objective probability there for the estimating.

De Finetti (1970) makes the point of this section very concisely when he says, "PROBABILITY DOES NOT EXIST." What he means is that objective probability does not exist. For a given event, each individual has a probability, but there is no such thing as the probability. An expert's probability represents the expert's judgment based on any available information. Since different experts may have access to different information or may interpret it differently, it is not surprising if they have different probabilities for the same event.

3.5 SUMMARY

When people learn about probability theory, it is generally in terms of mathematical manipulations of probabilities, with little attention paid to interpretation. When probabilities are used to represent uncertainty in practice, however, the interpretation is very important. In cases where there is general agreement that the outcomes are equally likely or that extensive past data are directly relevant, the classical and frequency approaches can be useful. This is seldom the case in uncertainty analyses and performance assessment of a high-level viable repository.

The appropriate interpretation of probability for the purposes of uncertainty analysis and performance assessment is the subjective interpretation. This may seem to fly in the face of the scientific desire for "objectivity," but that cannot be avoided. There are no "objective," or "true," probabilities for events such as the release of a certain amount of radioactivity from a waste storage facility in a given period of time, any more than there is a "true" probability that a Republican will be elected as President of the United States in 1996. (Those who would argue that the release of radioactivity is predictable

scientifically should think not just about uncertainties regarding the exact makeup of the waste, the condition of the containers in which the waste is stored, etc., but also about the impact of potential earthquakes, human intrusion, and so on.)

Experts have relevant information and can represent this information in the form of subjective probabilities. Fortunately, a solid scientific foundation exists in the form of a theory of subjective probability and a set of formal methods that have been developed to elicit subjective probabilities in practice. Thus, it is possible to justify the use of expert judgment in the form of probabilities and to obtain such probabilities for use in risk assessment, decision analysis, and other situations where experts' probabilities can provide valuable information.

4 QUALITY OF EXPERTS' PROBABILITIES

4.1 INTRODUCTION

This chapter reviews the published literature on the quality of probability assessments from individuals. While the emphasis is on probabilities obtained from experts for events or variables in their field of expertise, the number of studies meeting this emphasis is small compared to the number of studies using experts to give nonprobabilistic assessments and those using nonexperts to give probabilistic judgments. A portion of the review will include studies that do not involve experts. No attempt will be made to review the extensive literature on experts providing nonprobabilistic assessments.

4.2 NONEXPERTS PROVIDING PROBABILITIES

Many of the studies of the quality of subjective probabilities have been conducted in the fields of decision analysis and cognitive psychology. Most of these studies used student subjects. Reviews of this work are found in Lichtenstein et al., (1982) and Wallsten and Budescu (1982). These studies can be divided into those assessing probabilities of events and those assessing probability densities for unknown quantities. First, studies where probabilities of events were assessed will be reviewed.

4.2.1 Probabilities of Events

In one pair of early studies (Adams and Adams, 1958, 1961), small numbers of student subjects participated in a multiple-session experiment. In the first study, students were asked to assign probabilities to pairs of words being synonyms, antonyms, or unrelated words. In the second study, the students were asked to assess the percentage of blue dots in an array of red and blue dots, judge the truth of general knowledge statements, judge blindfolded the weight of lifted objects, and perform the same task as in the first study. Some participants were provided feedback between sessions about the calibration of their probabilities while others (the control group) were not. The calibration of subjects, as measured by the absolute differences between assessed probabilities and observed relative frequencies, improved significantly with feedback.

Other studies have been conducted using general knowledge or almanac questions. Often the assessment task requires the subject to select one of two alternatives as more likely to be true and to provide a probability that the statement is true. Thus, in response to the question "Is the White House or the Treasury Building shown on the twenty dollar bill?", the subject might respond "the White House" and provide a 0.7 probability that the judgment is correct. These studies, whether conducted using students as subjects or professionals from various fields, share one common finding — overconfidence, for example, probabilities too extreme in relation to the corresponding relative frequencies. Figure 4-1 is taken from Lichtenstein et al., (1982). This figure shows calibration plots from four studies. The studies by Hazard and Peterson (1973) used armed forces personnel studying at the Defense Intelligence School as subjects. The Phillips and Wright (1977) study used student subjects, while the type of subject in the Lichtenstein study is unknown.

In Fischhoff et al., (1977), various question formats and response modes (no alternative, one or two alternatives, half and full-range responses, and logs) for questions about events were tried, but overconfidence remained. They reported that only 72 to 83 percent of items assigned probabilities of 1.0 were actually true, for instance.

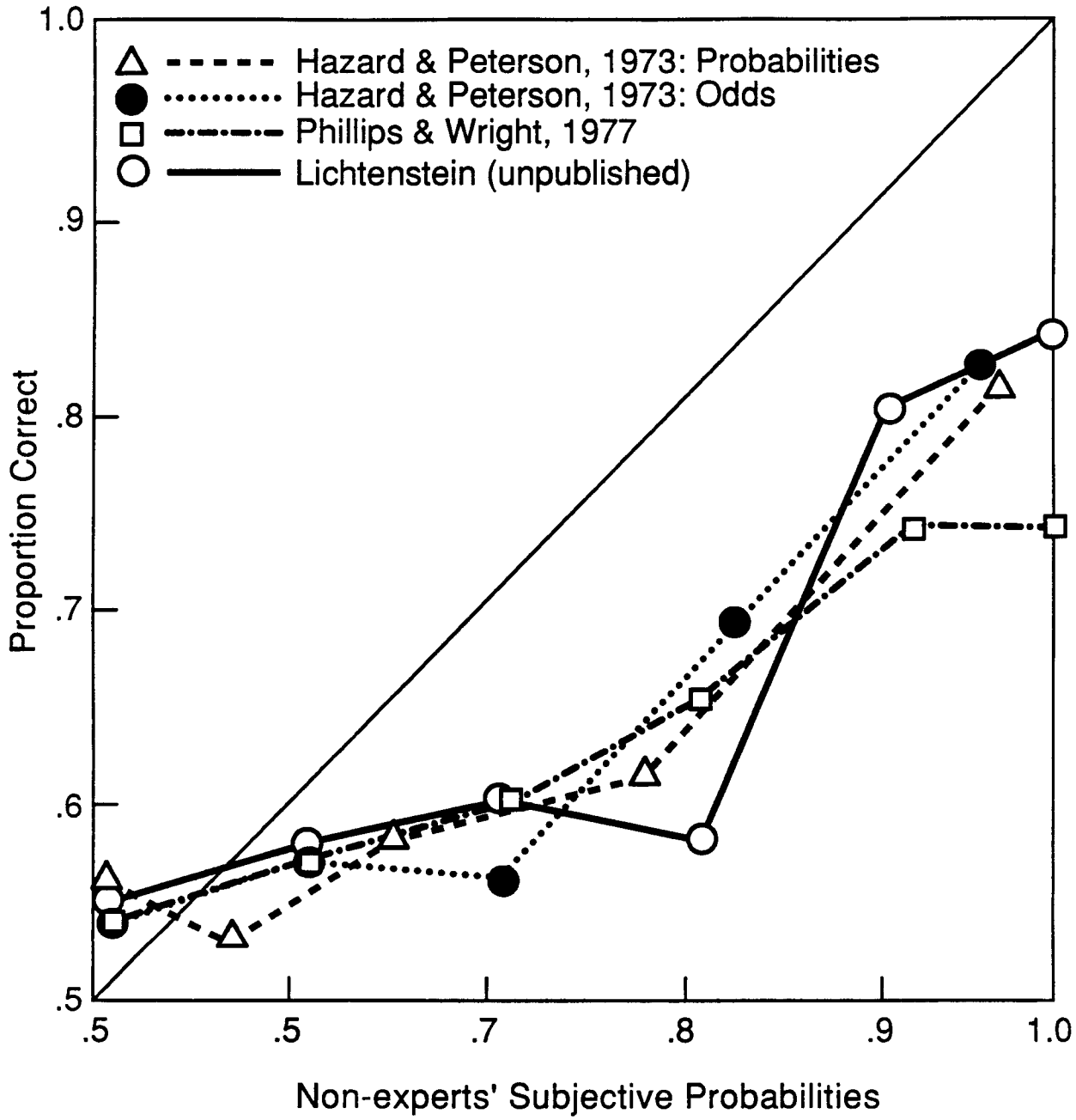


Figure 4-1. Calibration curves elicited from nonexperts using general knowledge questions

Several early studies indicated that the difficulty of the questions could influence calibration. Unfortunately, some of these studies were methodologically flawed in that the responses by the subjects were used to determine both the difficulty of the questions and the degree of calibration. Lichtenstein and Fischhoff (1980), however, employed a set of 500 general knowledge questions and a measure of difficulty derived from the ratio of values inherent in the two alternatives in each question to show that overconfidence increases with question difficulty. Although the evidence is modest, it appears that calibration becomes worse when the question difficulty increases.

Koriat et al., (1980) found that calibration could be improved by asking subjects to list reasons for and against their responses. Improvement in calibration was due almost exclusively to the listing of negative reasons. This suggests that persons fail to consider or give proper weight to negative indicators. In a study of point estimates, MacGregor et al., (1988) found that student subjects who listed reasons for their judgments performed better than those who did not.

4.2.2 Probability Distributions for Quantities

Overconfidence is also pervasive in studies employing continuous quantities. Table 4-1, adapted from Lichtenstein et al., (1982), shows the results of a number of such studies. Shown are the number of assessments (questions times subjects), the number of responses falling within the assessed interquartile ranges, and the expected and actual proportion of responses falling in the extreme tails of the assessed distributions, that is, the surprise index. The relative frequency within the interquartile range should be 50 percent when the subjects are well calibrated. The expected relative frequency of responses falling into the extreme tails depends upon the definition used in the study. Each of these studies points to fewer than 50 percent of the true values being within the quartiles. Almost invariably, however, the true values are found in the tails of the distribution more often than they should be.

The five results reported for the Alpert and Raiffa (1982) study were obtained using almanac questions. The first and second groups were asked to provide 0.01 and 0.99 fractiles and 0.001 and 0.999 fractiles, respectively. The third and fourth groups were asked to provide endpoints that were the judged minimum and maximum possible values and values that were astonishingly high and low, respectively. The fifth group is actually the combined result for three groups who received feedback after their first ten assessments. Two things are clear in these findings: the spread of the tails seems to be somewhat insensitive to the definition of the extreme fractiles, and while feedback provides substantial improvement in overconfidence, it does not do away with this fault.

The two results reported by Selvidge (1975) used assessments of five fractiles (0.01, 0.25, 0.50, 0.75, and 0.99) and seven fractiles (0.10 and 0.90, in addition to the five fractiles). Selvidge also permuted the order that the fractiles were assessed and found better calibration when the more central fractiles were assessed first. The Moskowitz and Bullers (1978) study employed two types of questions, self-generated proportions and the Dow-Jones average, and a three and a five fractile assessment protocol. Self-generated proportions were obtained from the subjects by asking questions such as "Do you prefer Bourbon or Scotch whiskey?"

Seaver et al., (1978) used five assessment techniques: fractiles, odds-fractiles, probabilities, odds, and log odds. The results in this study, however, may reflect the fact that the experimenters established values for some response modes (probabilities, odds, and log odds) which may have anchored the subjects to these values and thus, inadvertently, improved the calibration. Schaefer and Borcherding

Table 4-1. Calibration studies with continuous quantities; nonexperts

Study (Type of Subjects)	Number of Assessments	Observed Interquartile Frequency		Expected Surprise Index	Observed Surprise Index	
		Before Training/ Feedback	After Training/ Feedback		Before Training/ Feedback	After Training/ Feedback
Alpert & Raiffa (1969) (Graduate Business Students)	880 500 700 700 2270	0.33 0.33 0.33 0.33 0.34	 0.44	0.02 0.002 0.00 0.00 0.00	0.46 0.40 0.47 0.38 0.34	 0.19
Hession & McCarthy (1974) (Graduate Students)	2035	0.25		0.02	0.47	
Selvidge (1975)	400 520	0.56 0.50		0.02 0.02	0.10 0.07	
Moskowitz & Bullers (1978)	120 145 210 210	- 0.32 - 0.20		0.02 0.02 0.02 0.02	0.27 0.42 0.38 0.64	
Pickhardt & Wallace (1974) (Graduate Business Students)	? ?	0.39 0.30	0.49 0.45	0.02 0.02	0.32 0.46	0.20 0.24
Brown (1973)	414	0.29		0.02	0.42	
Lichtenstein & Fischhoff (1980)	924	0.32	0.37	0.02	0.41	0.40
Seaver, von Winterfeldt & Edwards (1978)	160 160 180 180 140	0.42 0.53 0.47 0.47 0.31		0.02 0.02 0.02 0.02 0.02	0.34 0.24 0.05 0.05 0.20	
Schaefer & Borcherding (1973) (Students)	396 396	0.23 0.16	0.38 0.48	0.02 0.02	0.39 0.50	0.12 0.06
Larson & Reenan (1979)	450	-		Reasonably Certain	0.42	

(1973) also experimented with two assessment methods, in this case fractile assessment and hypothetical samples. For both methods, the extensive feedback produced substantial improvements in calibration.

Finally, Larson and Reenan (1979) asked subjects to establish a best guess and then two bounding values such that they were "reasonably certain" that the true value would be found. The results are similar.

4.3 EXPERTS PROVIDING PROBABILITIES

In this section, published results of the quality of probabilities obtained from experts are reviewed. In the review, studies involving experts from similar fields will be reviewed together. Results from some of the studies are summarized in Table 4-2. The fields reviewed include engineering and risk analysis, weather forecasting, medicine and psychiatry, intelligence and military applications, and business.

4.3.1 Engineers, Scientists, and Risk Analysts

Hofer et al., (1985) relate an example originally reported by Hynes and VanMarcke (1976). The practical background was the construction of a highway through the tidal marshes north of Boston. The question concerned the additional height of fill at which stability failure of the embankment would occur due to the deformation behavior of the clay foundation. Comprehensive geological data, the evaluation of test drilling, and the results of various measurements performed prior to and during a certain period of the filling process were made available to experts for use in their computational models. Figure 4-2 shows the 95 percent probability intervals formed by the experts.

Mosleh et al., (1987) present an interesting comparison of judgmental distributions for component repairs to empirical rates obtained from review of power plant operating histories. The judgmental distributions had been obtained for use in nuclear reactor probability risk assessments (PRAs). This situation is not the same as when an expert provides a probability distribution for a single quantity, however, since the validating data also form a distribution. The authors conclude that the experts systematically underestimated the actual variation of the mean maintenance duration from one plant or component to another. However, in eleven of the twelve cases examined, the judgmental mean was within a factor of four of the empirical mean. Nine of the twelve maintenance time means were larger than the corresponding empirical means, perhaps indicating some conservatism. This result compares favorably to the data reported by Kidd (1970), which shows that engineers considerably underestimated the repair times for generators.

Some studies known as the European Benchmark Exercises have examined systems analysis of nuclear safety by using several independent groups to analyze the same problem. In Amendola (1986), the Auxiliary Feedwater system of the EdF 1300 MWe PWR was studied. The results of eight teams were assessed at different points in the study. The ratio of maximum to minimum failure estimates provided by the teams was 25 on first evaluation. The ratio increased to 36 after the teams made a comparison of qualitative analyses but used different fault trees, and it fell to 9 when the teams used a common fault tree but different data sets. When the teams used the fault tree and the same data set, the ratio of the maximum to minimum estimate became nearly one. This finding supports the contention expressed in Meyer and Booker (1991) and Hora et al., (1992) that the way decomposing or thinking about a problem, in this case the fault tree, is an important determinant of the judgments.

Table 4-2. Calibration studies for continuous variables; experts

Study (Type of Subjects)	Number of Assess-ments	Observed Interquartile Frequency		Expected Surprise Index	Observed Surprise Index	
		Before Training/ Feedback	After Training/ Feedback		Before Training/ Feedback	After Training/ Feedback
Pratt (1975) (One Expert)	175	0.37		0.00	0.05	
Murphy & Winkler (1974) (Weather Forecasters)	132	0.45		0.25	0.27	
Murphy & Winkler (1977b) (Weather Forecasters)	432	0.54		0.25	0.21	
Stael von Holstein (1971a) (Weather Forecasters)	1269	0.27		0.02	0.30	
Hora, Hora & Dodd (1992) (Scientists and Nuclear Engineers)	480 480			0.00	0.35	
Cooke (1991a) (Space Systems and Atmospheric Experts)	80 52 396 154	- - - -		0.10 0.10 0.10 0.10	0.44 0.37 0.37 0.10	
Henrion and Fischhoff (1986)	27 17 7 38 7 306	0.41 0.41 0.14 0.50 1.0 -		0.02	0.11 0.29 0.14 0.27 0.00 0.07	
Hynes and Van Marcke (1977)	7			0.00	0.57	
Tomassini et al., (1982)	6x	0.64		0.02 0.20	0.07 0.14	

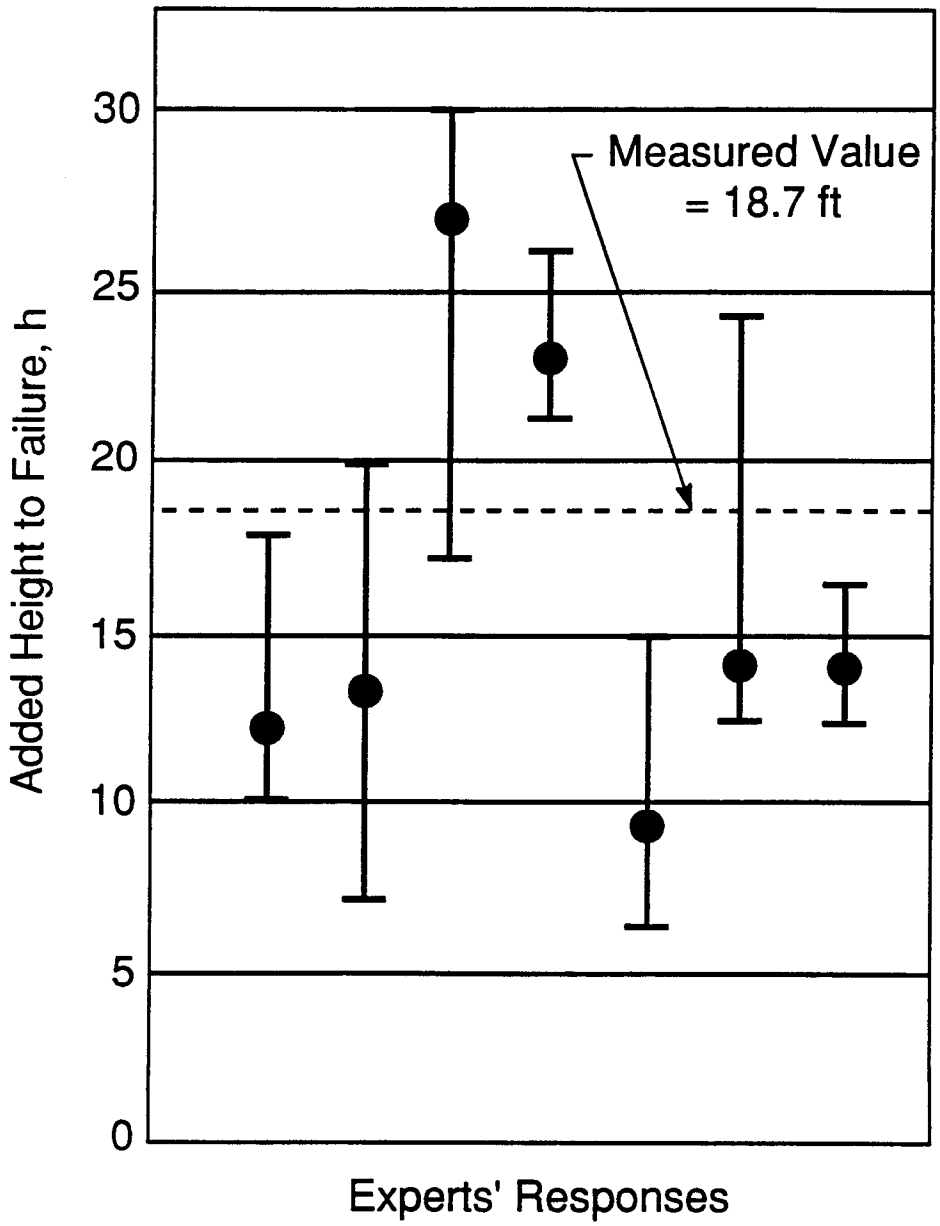


Figure 4-2. Results from a survey of expert opinion in geotechnics (Hynes and VanMarcke, 1976)

The benchmark studies do not, however, make comparisons to known values and thus they do not provide a true method of validation. In contrast, what they do provide is a comparison among analyses and a measure of agreement. However, agreement and accuracy are not synonymous. One can have good agreement on the wrong value.

Henrion and Fischhoff (1986) examine, retrospectively, values and uncertainties for a number of physical constants. They find that past assessments of uncertainties associated with estimates of constants such as the speed of light, Avogadro's number, Planck's constant, etc., tend to be underestimated. For example, Figure 4-3 shows estimates and uncertainties for the speed of light published between 1875 and 1958. The error bars in Figure 4-3 represent one standard deviation from the central estimate. These error ranges have been converted to 50 percent and 98 percent confidence intervals by assuming a normal error distribution. The resulting relative frequencies are shown in Table 4-2.

A study with space scientists is reported in Cooke (1991a). As part of the study, ten experts were asked to provide distributions for system reliabilities. Cooke developed eight questions about space systems reliabilities and reliabilities of similar systems in a manner such that the experts would not have had access to the data. The experts assessed the 0.05, 0.50, and 0.95 quantiles. Thirty-five of 80 responses did not contain the true value within the 0.05 and 0.95 quantiles.

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

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

Examination of Cooke's data also show that "surprises" (in the upper or lower 5 percent of the tails) occur 27 percent of the time. The poor performance of the surprise index can be traced to gross overconfidence exhibited by several of the scientists. Very little training in probability assessment was provided to the scientists, however.

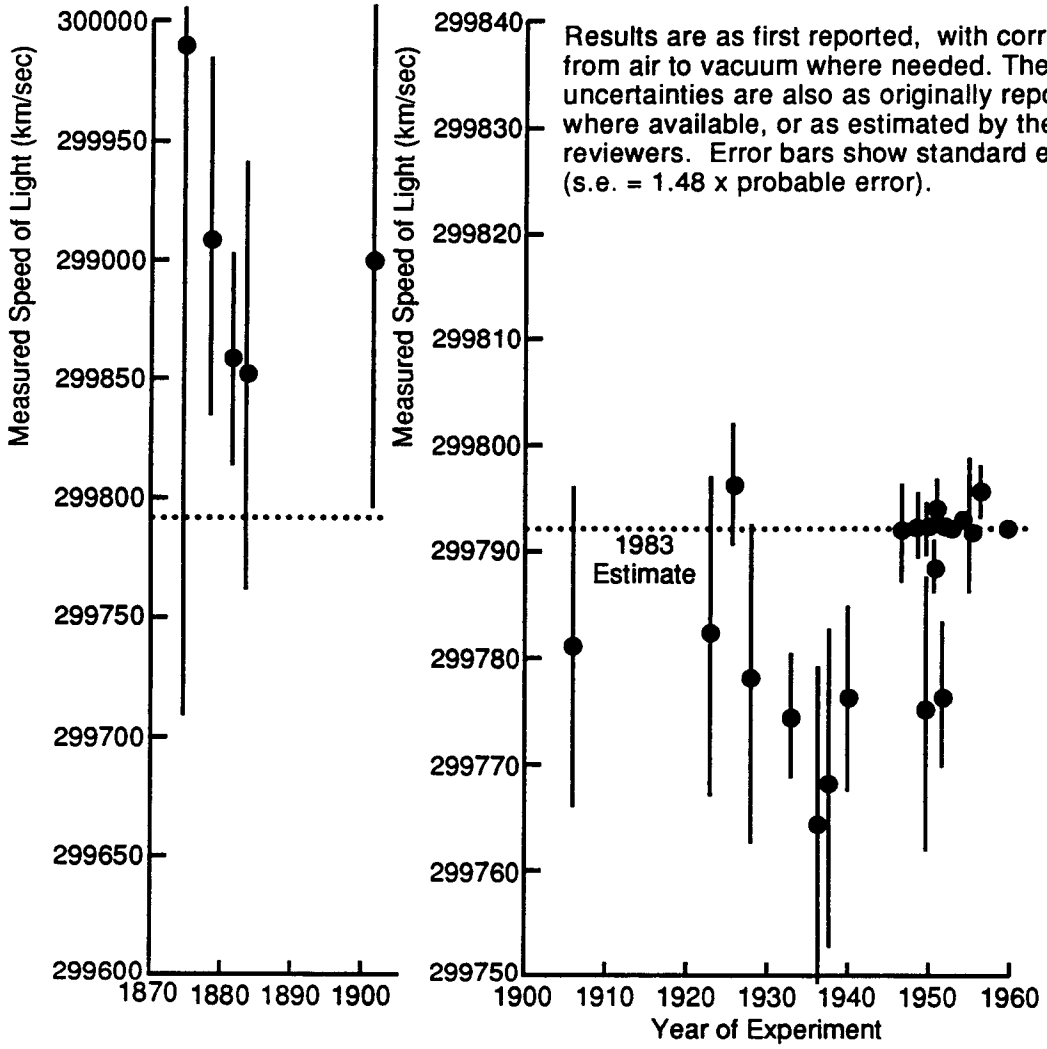


Figure 4-3. Experts' estimates and uncertainties for speed of light published between 1875 and 1958 (Henrion and Fischhoff, 1986)

4.3.2 Weather Forecasting

The most studied, and perhaps best calibrated, experts are weather forecasters. Beginning in 1965, U.S. weather forecasters have routinely made probabilistic predictions of precipitation. Murphy and Winkler (1977a) studied precipitation forecasts given by professional weather forecasters in Chicago. Some 17,514 forecasts were analyzed. Figure 4-4 shows the calibration plot, which is nearly perfect, for these forecasts. Weather forecasters have several advantages: the task is repetitious, there is an excellent base of information, feedback is provided, and they are rewarded for good performance. In a recent study, Winkler (1991) reports precipitation forecast performance in terms of scoring rules at twenty National Weather Service stations. Performance was found to vary with location because of difficulties of forecasting and from real differences among expertise at various stations. An asymmetric scoring rule was employed to measure the skill of forecasters.

Murphy and Winkler (1977b) report on the performance of forecasters making probabilistic predictions of minimum and maximum temperature. In contrast to precipitation forecasting, which requires the assignment of a single probability to an event, this task involves assessing a distribution for a continuous quantity. The calibration of forecasters making assessments by giving the fractiles was very good in contrast to the rather poor calibration of those who provided probabilities for fixed intervals centered around the median.

In contrast, Stael von Holstein (1971) asked weather forecasters to provide probability distributions for the next day's average temperature, average temperature 4 and 5 days into the future, and total rain for the next five days. The distributions were obtained using assessments of probabilities for fixed intervals. The resulting distributions were poorly calibrated; in terms of the interquartile range and the surprise index, they were comparable to distributions obtained from nonexperts using almanac questions. The question of refinement, as opposed to calibration, unfortunately, was not addressed in this study.

4.3.3 Medicine

Physicians have also been expert subjects in studies of probability assessment. Wallsten and Budescu (1982) report on a study by Lusted (1977) involving physicians assigning probabilities to the most important and most likely diagnosis. The diagnoses were then evaluated via an x-ray. Ludke et al., (1977) presented calibration curves for three classes or problems -- skull fracture, extremity fractures, and pneumonia. The calibration for extremity fractures is nearly perfect, while probabilities are uniformally overestimated for skull fractures. For pneumonia, low probabilities are overestimated while the converse is true for large probabilities. DeSmet et al., (1979) also present calibration data for physicians which also suggest over estimation of probabilities associated with head traumas.

Overestimation of probabilities of illness by physicians was observed in Centor et al., (1984). In this study, physicians assigned probabilities of a streptococcus inflection to patients complaining of sore throats. The researchers compared the probabilities assigned by the physicians to those estimated through a regression model. The comparison was based on Brier scores. The results showed that while the physicians had better resolution than the regression model, the regression model predictions were better calibrated. A similar comparison was made by Lee et al., (1986) with the result that a proportional hazards regression model gave better Brier scores than four out of five doctors.

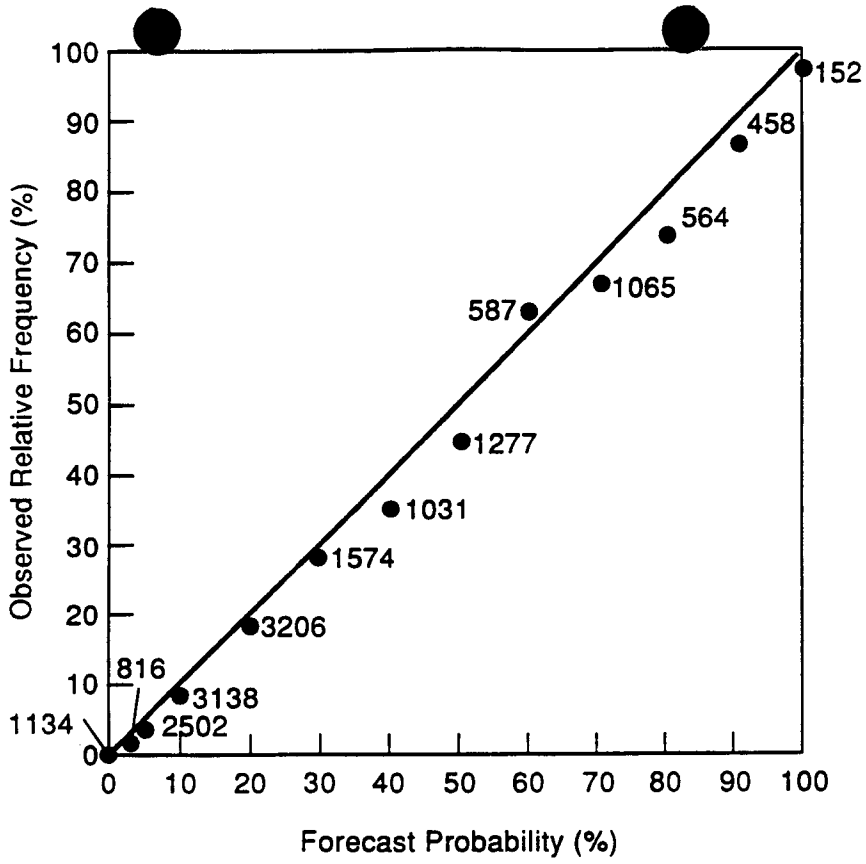


Figure 4-4. Calibration plot for professional weather forcasters making probabilistic predictions of precipitation. The number of forecasts is shown for each point (Murphy and Winkler, 1977a).

In a study by Christensen-Szalanski and Bushyhead (1981), nine physicians examined 1,531 first-time patients for pneumonia. Each patient was examined by one physician, and a pneumonia probability was assigned. Verification was made by x-rays. The physicians assigned probabilities that were much too high. The authors suggest that patients may be assumed to have pneumonia until it is proven otherwise. They also suggest that a check list of symptoms may be helpful, as well as base rate information. Physicians may rely too much on the presence of cues and ignore the absence of cues. Of course, motivational bias, caused by the asymmetry of the costs of misdiagnosis (declaring a healthy patient ill versus declaring an ill patient healthy) may also explain the tendency to assign probabilities that are too large.

In contrast, McClish and Powell (1989) and Poses et al., (1989) present findings that show very good calibration among physicians. The McClish and Powell study deals with patient mortality in an intensive care unit. Once again, the physicians' performance is compared to the performance of a model with the result that the physicians have more resolution but the model is better calibrated. The setting for the Poses et al. study is a critical care unit. Three calibration curves are shown in Figure 4-5. These curves show the assigned probability of survival (abscissa) plotted against the empirical probability of survival in each forecast bin. The first plot, A, is for house officers who show some underestimation of survival probabilities, while the second plot, B, is for critical care attendings who show remarkably good calibration.

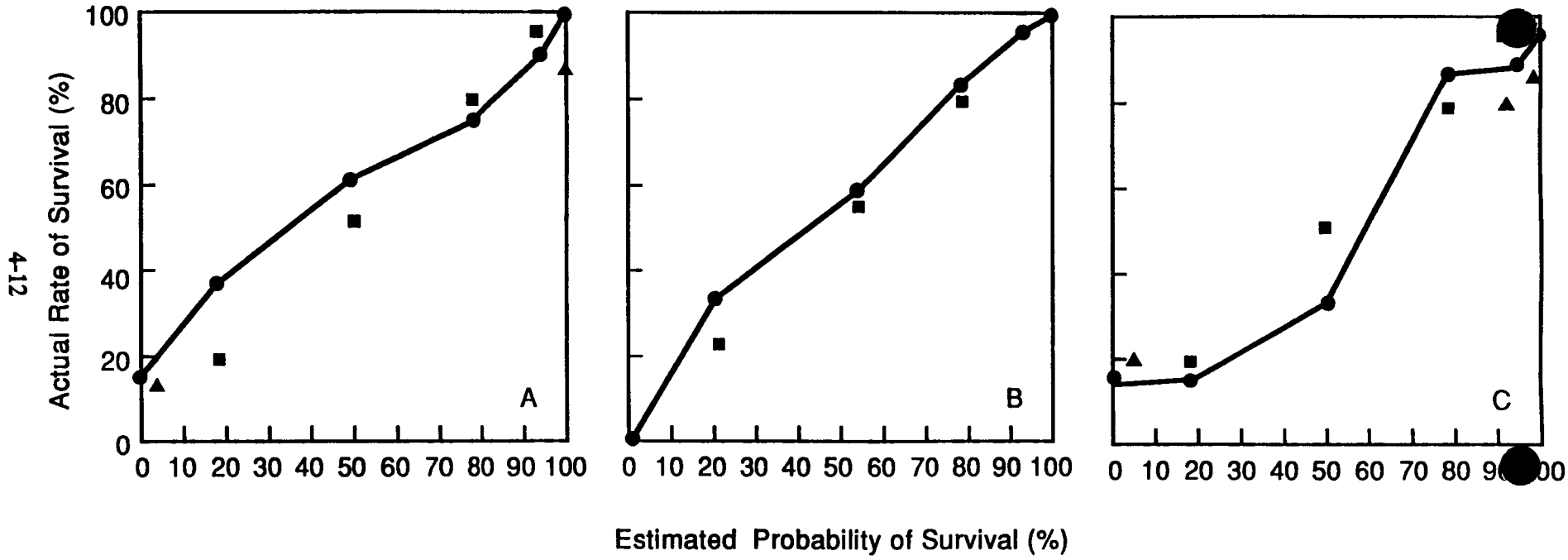


Figure 4-5. Calibration curves for physicians estimating patient mortality in intensive care units (Poses et al., 1989)

Similarly, Winkler and Poses (1991) examined probabilities of patient survival given by physicians in an intensive care unit. The physicians were classified into four levels of experience by title. While all four groups of physicians were found to be reasonably well calibrated, those with the most experience and expertise performed better overall in terms of average scores. This study found that the key factors in performance involved discrimination and resolution. The more experienced physicians demonstrated the ability to group together patients with similar survival chances.

4.3.4 Business

In another study by Stael von Holstein (1972), a total of 72 participants were used to assess discrete probability distributions for changes in stock prices. Ten of the participants worked in the stocks and bonds department of a Stockholm Bank, 10 were connected with the stock exchange, 11 were statisticians in academic positions, 13 were business administration faculty, and 28 were students. The target quantities were share prices two weeks hence for twelve stocks. The assessments involved assigning probabilities to five classes of price changes. Ten sessions were conducted with feedback on performance provided before each new assessment.

The evaluation of the resulting probability distributions was based on a quadratic scoring rule. Comparisons to log and spherical rules were also made. Stael von Holstein was surprised to find that the bankers performed most poorly. From best to worst, the groups of the participants were statisticians, stock market employees, students, business teachers, and, far behind, bankers. At the beginning of the study, the statisticians were found to give more spread-out distributions. As the sessions progressed the distributions averaged across all experts became more spread out. Thus, the experts learned through feedback to spread their distributions. The response categories (classes of price changes) were a priori chosen to be of about equal probability. Assigning equal probabilities to all classes in every session would have resulted in better scores for 69 out of 72 participants.

The task assigned to the experts in this study is, indeed, most difficult: the efficient market hypothesis popular in finance suggests that the subject can do no better than random chance without inside information.

Auditors served as subjects in a study by Tomassini et al., (1982). They were asked to review six scenarios and develop subjective probability distributions for quantities using the 7-fractile method of assessment. The results were compared to actual audit values. The auditors appeared to be better calibrated than most subjects; they exhibited less overconfidence and perhaps some underconfidence in their distributions, as shown in Table 4-2.

4.3.5 Experts Answering General Knowledge Questions

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

Hora et al., (1992) report calibration data obtained from scientists and engineers who participated in the expert judgment elicitation associated with NUREG-1150. During the training for

probability elicitation, the experts were asked to complete an almanac question training exercise to demonstrate the overconfidence bias. Each participant responded to one-half the questions by providing probabilities for intervals (direct assessment), and to the other half of the questions by providing values for cumulative probabilities (successive bisection.) Figure 4-6 shows calibration plots for these participants. These graphs are created by plotting the cumulative probability of the true answer against the empirical distribution function of that probability among all assessed probabilities. The plot should describe a 45 degree line when the experts are well calibrated. The calibration plot for this study, however, displays the typical degree of overconfidence obtained with other groups responding to general knowledge questions.

4.4 SUMMARY

The quality of individual probability assessments has been most often examined in terms of calibration. A few studies have used scoring rules which measure overall performance, incorporating refinement as well as calibration. The evidence about the quality of probabilities, however, is for the most part confined to evidence about calibration.

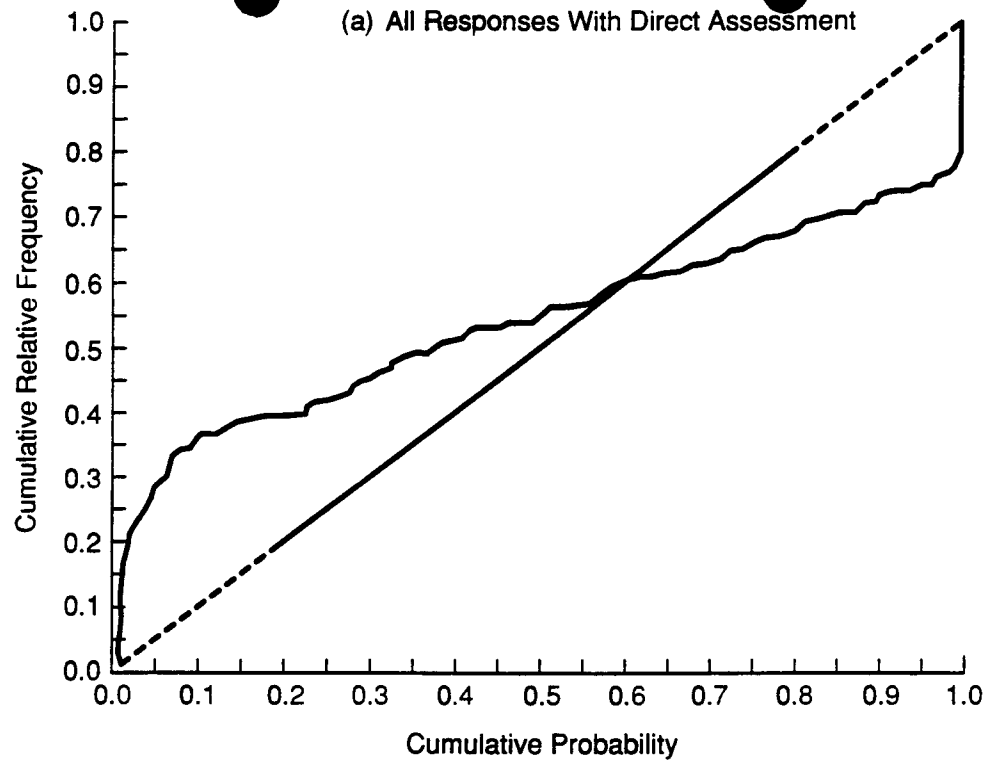
As Wallsten and Budescu (1983) note, experts involved in assessing probabilities for events with which they are familiar can be very well calibrated. The evidence indicates that continuous distributions from experts can also be well calibrated. A similar quality of goodness has rarely been demonstrated in laboratory settings with nonexpert subjects. Other studies with experts, however, show that when the task is less familiar, so that there is little opportunity for practice and feedback, experts may succumb to the same limitations as nonexperts.

The repetitive nature of weather forecasting and medical diagnosis provide opportunities for practice and feedback. Clearly, in these fields, probability assessors can, and should, perform well. Unfortunately, many of the assessments that must be made in risk analysis and performance assessment are not repetitive, nor can feedback be provided.

The majority of those studies in which training has been employed show modest to substantial improvement in calibration. Training and practice in making assessments, then, should be considered as an important step in improving the quality of assessments. The evidence also indicates that the quality of assessments improves with simpler questions. This suggests that a strategy of decomposition, such as that used in the EPRI and WIPP studies, may also help improve the quality of the elicited probabilities.

Because the most essential flaw in many studies of elicited probabilities is the tendency toward an understatement of uncertainty, methods should be considered to counteract this tendency. Asking assessors to consider reasons why an event does not occur as well as why it might occur is an example of an elicitation technique that can be used to improve the quality of probabilities. The choice of scales (e.g., log probabilities or log odds) may also help to counteract this bias.

In summary, in some circumstances experts may provide excellent probabilities. One should not conclude, however, that expertise alone is sufficient to guarantee that probabilities are of high quality. Practice and evaluation seem to be key ingredients in producing high quality probability assessments, and careful design of the overall assessment process is also important.



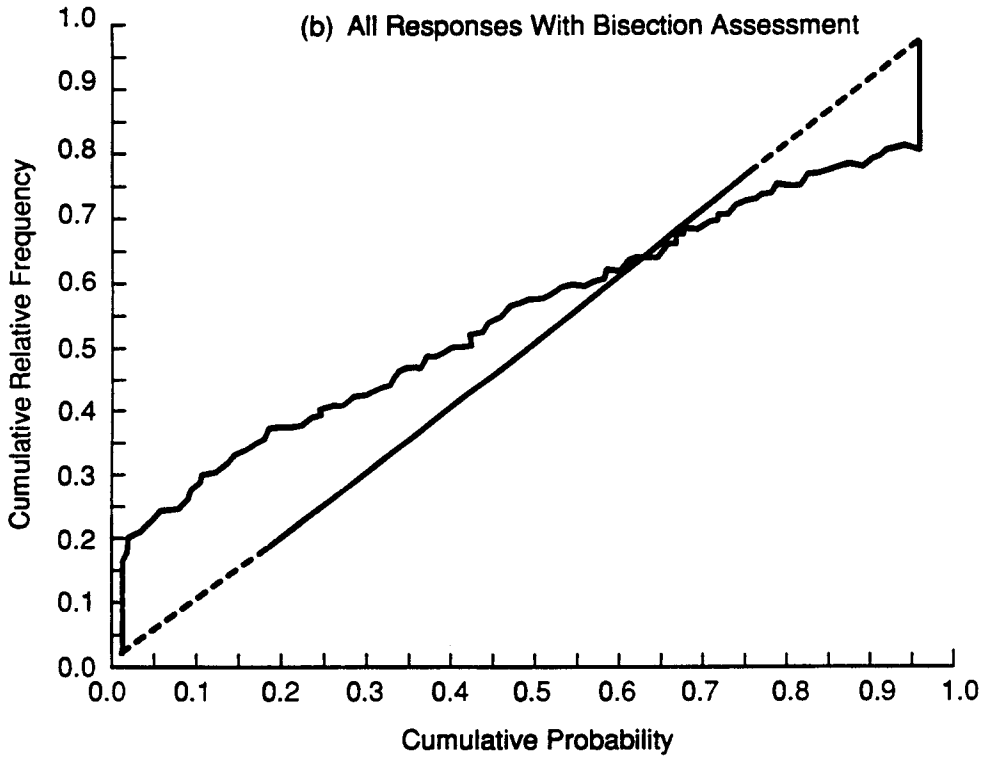


Figure 4-6. Calibration plots for scientists and engineers responding to almanac questions (Hora et al., 1992)

5 REFERENCES

Adams, J.K., and P.A. Adams. 1958. Training in confidence judgments. *American Journal of Psychology* 71: 747-751.

Adams, J.K., and P.A. Adams. 1961. Realism of confidence judgments. *Psychological Review* 68: 33-45.

Alpert, M., and H. Raiffa. 1982. A progress report on the training of probability assessors. D. Kahneman, P. Slovic, and A. Tversky, eds. *Judgment Under Uncertainty: Heuristics and Biases* Cambridge University Press: Cambridge, United Kingdom: 294-305.

Amendola, A. 1986. Uncertainties in systems reliability modeling: Insight gained through European benchmark exercises. *Nuclear Engineering and Design* 93: 215-225.

Bonano, E.J., Hora, S.C., Keeney, R.L., and D. von Winterfeldt. 1989. *Elicitation and Use of Expert Judgment in Performance Assessment for High-Level Radioactive Waste Repositories*. SAND89-1821. NUREG/CR-5411. Sandia National Laboratories (SNL): Albuquerque, New Mexico.

Brier, G.W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78: 1-3.

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

Camp, A.L., Kunsman, D.M., Miller, L.A., Sprung, J.L., Wheeler, T.A., and G.D. Wyss. 1990. *Level III Probabilistic Risk Assessment for N Reactor*. Vol. 2: WHC-MR-0045 and SAND 89-2102. Westinghouse Richland Co.: Richland, Washington.

Centor, R.M., Dalton, H.I., and J.F. Yates. 1984. Are physicians' probability estimates better or worse than regression model estimates? *Medical Decision Making* 4: 538.

Christensen-Szalanski, J.J.J., and J. Bushyhead. 1981. Physicians; Use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance* 7: 928-935.

Cooke, R.M. 1991a. *Experts in Uncertainty: Expert Opinion and Subjective Probability in Science*. Oxford University Press: Oxford, United Kingdom.

Cooke, R.M. 1991b. *Expert Judgment Study on Atmospheric Dispersion and Deposition*. Report 91-81: Faculty of Technical Mathematics: Delft University of Technology: Delft, The Netherlands.

de Finetti, B. 1937. La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré* 7: 1-68.

de Finetti, B. 1970. Teoria Delle Probabilita. Torino: Giulio Einaudi. Translation by A. Machi and A. Smith. *Theory of Probability*. Vols. 1 and 2. 1974 and 1975. Wiley: London, United Kingdom.

Electric Power Research Institute (EPRI). 1986. *Seismic Hazard Methodoloogy for the Central and Eastern United States, Vol. 1: Methodology*. NP-4/26. EPRI: Palo Alto, California.

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

Fischhoff, B., Slovic, P. and S. Lichtenstein. 1977. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance* 3: 552-564.

Green, D.M., and J.A. Swets. 1974. *Signal Detection Theory and Psychophysics*. 2nd Ed. Wiley: New York, New York.

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

Henrion, M., and B. Fischhoff. 1986. Assessing uncertainty in physical constants. *American Journal of Physics* 54: 791-798.

Hofner, E., Javeri, V., and H. Loffler. 1985. A survey of expert opinion and its probabilistic evaluation for specific aspects of the SNR-300 Risk Study. *Nuclear Technology* 68: 180-225.

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

Hora, S.C., Hora, J.A., and N. Dodd. 1992. Assessment of probability distributions for continuous random variables: A comparison of the bisection and fixed value methods. *Organizational Behavior and Human Decision Processes*. In press.

Hynes, H.E., and E.M. VanMarcke. 1976. Reliability of embankment performance predictions. *Mechanics in Engineering*. University of Waterloo Press: Waterloo, Ontario.

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

Keeney, R.L., and D. von Winterfeldt. 1989. On the uses of expert judgment on complex technical problems. *Institute of Electrical and Electronics Engineers, Inc. (IEEE) Transactions on Engineering Management* 36: 83-86.

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

Kidd, J.B. 1970. The utilization of subjective probabilities in production planning. *Acta Psychologica* 34: 338-47.

Koriat, A., Lichtenstein, S. and B. Fischhoff. 1980. Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory* 6: 107-118.

Kyburg, H.E., and Smokler, H.E., Eds. 1964. *Studies in Subjective Probability*. Wiley: New York, New York.

Larson, J.R., and A.M. Reenan. 1979. The equivalence interval as a measure of uncertainty. *Organizational Behavior and Human Performance* 23: 49-55.

Lee, K.L., Pryor, D.B., Harrell, F.E., Califf, R.M., Behar, V.S., Floyd, W.L., Morris, J.J., Waugh, R.A., Whalen, R.E., and R.S. Rosati. 1986. Predicting outcome in coronary disease. *The American Journal of Medicine* 80: 553-560.

Levi, K. 1985. A signal detection framework for the evaluation of probabilistic forecasts. *Organizational Behavior and Human Decision Processes* 36: 143-166.

Lichtenstein, S. and B. Fishhoff. 1980. Training for Calibration. *Organizational Behavior and Human Performance* 26: 149-71.

Lichtenstein, S., Fischhoff, B. and L.D. Phillips. 1982. Calibration of probabilities: The state of the art to 1980. H. Jungermann and G. de Zeeuw, eds. *Decision Making and Change in Human Affairs*. Reidel: Dordrecht, The Netherlands: 275-324.

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

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

MacGregor, D., Lichtenstein, S., and P. Slovic. 1988. Structuring knowledge retrieval: An analysis of decomposing quantitative judgments. *Organizational Behavior and Human Decision Processes* 42: 303-23.

Matheson, J.E. and R.L. Winkler. 1976. Scoring rules for continuous probability distributions. *Management Science* 22: 1087-1096.

McClish, D.K., and S.H. Powell. 1989. How well can physicians estimate mortality in a medical intensive care unit? *Medical Decision Making* 9: 125-132.

Merkhofer, M.W., and A.K. Runchal. 1989. Probability encoding: Quantifying judgmental uncertainty over hydrological parameters for basalt. *Proceedings of the Conference on Geostatistical, Sensitivity and Uncertainty Methods for Ground-Water Flow and Radionuclide Transport Modeling*. Battelle Press: Columbus, Ohio: 629-648.

Meyer, M.A., and J.M. Booker. 1991. *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Academic Press: New York, New York.

Morgan, M.G., Amaral, D., Henrion, M., and S. Morris. 1984. Technological uncertainty in quantitative policy analysis: A sulfur polution example. *Risk Analysis* 3: 201-220.

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

Moskowitz, H., and W.I. Bullers. 1978. *Modified PERT Versus Fractile Assessment of Subjective Probability Distributions*. Paper No. 675. Purdue University: Lafayette, Indiana.

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

Murphy, A.H. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* 12: 595-600.

Murphy, A.H., and R.L. Winkler. 1977a. Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics* 26: 41-47.

Murphy, A.H., and R.L. Winkler. 1977b. The use of credible intervals in temperature forecasting: Some experimental results. H. Jungermann and G. deZeeuw, eds. *Decision Making and Change in Human Affairs*. Reidel: Dordrecht, The Netherlands: 45-56.

Murphy, A.H., and R.L. Winkler. 1977. Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest* 2: 2-9.

Murphy, A.H., and R.L. Winkler. 1984. Probability forecasting in meteorology. *Journal of the American Statistical Association* 79: 489-500.

Murphy, A.H., and R.L. Winkler. 1987. A general framework for forecast verification. *Monthly Weather Review* 115: 1330-1338.

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

Otway, H., and D. von Winterfeldt. 1992. Expert judgment in risk analysis and management: Process, context, and pitfalls. *Risk Analysis* 12: 83-93.

Phillips, L.D., and G.N. Wright. 1977. Cultural differences in viewing uncertainty and assessing probabilities. H. Jungermann and G. deZeeuw, eds. *Decision Making and Change in Human Affairs*. D. Reidel: Amsterdam, The Netherlands.

Poses, R.M., Bekes, C., Copare, F.J., and W.E. Scott. 1989. The answer to "What Are My Chances, Doctor?" depends on whom is asked: Prognostic disagreement and inaccuracy for critically ill patients. *Critical Care Medicine* 17: 827-833.

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

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

Richmond, H.M. 1987. *Development of Probabilistic Health Risk Assessment for National Ambient Air Quality Standards*. Air Pollution Control Association (APCA) International Specialty Conference on Regulatory Approaches for Control of Air Pollutants: Atlanta, Georgia.

Savage, L.J. 1954. *The Foundations of Statistics*. Wiley: New York, New York.

Savage, L.J. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* 66: 783-801.

Schaefer, R.E., and K. Borcherding. 1973. The assessment of subjective-probability distributions: A training experiment. *Acta Psychologica* 37: 117-129.

Seaver, D.A., von Winterfeldt, D., and W. Edwards. 1978. Eliciting subjective probability distributions on continuous variables. *Organizational Behavior and Human Performance* 21: 379-391.

Selvidge, J. 1975. *Experimental Comparison of Different Methods for Assessing the Extremes of Probability Distributions by the Fractile Method*. Report 75-13. Graduate School of Business Administration: University of Colorado: Boulder, Colorado.

Spetzler, C.S., and C.-A.S. Stael von Holstein. 1975. Probability encoding in decision analysis. *Management Science* 22: 340-358.

Stael von Holstein, C.-A.S. 1971. An experiment in probabilistic weather forecasting. *Journal of Applied Meteorology* 10: 635-645.

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

Trauth, K.M., Rechard, R.P., and S.C. Hora. 1991. Expert judgment as input to waste isolation pilot plant performance assessment calculations: Probability distributions of significant parameters. A.A. Moghissi and G.A. Benda, eds. Mixed waste. *Proceedings of the First International Symposium*. Baltimore, Maryland.

Tomassini, L.A., Solomon, I., Romney, M.B., and J.L. Krogstad. 1982. Calibration of auditors' probabilistic judgments: Some empirical evidence. *Organizational Behavior and Human Performance*. 30: 391-406.

U. S. Nuclear Regulatory Commission. 1989. *Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants*. NUREG 75/014. Washington, D.C.: Nuclear Regulatory Commission.

von Winterfeldt, D., and W. Edwards. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press: Cambridge, United Kingdom.

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

Winkler, R.L. 1967. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association* 62: 1105-1120.

Winkler, R.L., Smith, W.S., and R.B. Kulkarni. 1978. Adaptive forecasting models based on predictive distributions. *Management Science* 24: 977-986.

Winkler, R.L. 1985. Measuring skill and utility of probability forecasts. *Proceedings of the Ninth Conference on Probability and Statistics in Atmospheric Sciences*. American Meteorological Society: Boston, Massachusetts: 180-186.

Winkler, R.L. 1991. Evaluating Probabilities: Asymmetric Scoring Rules. Unpublished manuscript. Duke University: Durham, North Carolina.

Winkler, R.L., and R.M. Poses. 1991. Evaluating and Combining Physicians' Probabilities of Survival in an Intensive Care Unit. Unpublished manuscript. Duke University: Durham, North Carolina.

Wheeler, T.A., Hora, S.C., Cramond, W.R., and S.D. Unwin. 1989. *Analysis of Core Damage Frequency*. Vol. 2: Expert Judgment Elicitation. NUREG/CR-4550. SAND86-2084. SNL: Albuquerque, New Mexico.

Whitfield, R.G., and T.S. Wallsten. 1989. A risk assessment for selected lead-induced health effects: An example of a general methodology. *Risk Analysis* 9: 197-207.

Whitfield, R.G., Wallsten, R.L., Winkler, R.L., Richmond, H.M., Hayes, S.R., and A.S. Rosenbaum. 1991. *Assessing the Risk of Chronic Lung Injury Attributable to Long-Term Ozone Exposure*. ANL/EAIS-2. Argonne National Laboratory (ANL), Argonne, Illinois.

APPENDIX A

EVALUATION OF EXPERTS' PROBABILITIES

EVALUATION OF EXPERTS' PROBABILITIES

A.1 INTRODUCTION

It is common practice to judge the quality of a forecast retrospectively by comparing the forecast to what actually happens. For example, if the forecast is that an event will occur, then the forecast is evaluated after it is learned whether the event does in fact occur or not. Similarly, when the forecast is about a variable, such as tomorrow's high temperature or next year's growth in gross national product (GNP), the goodness of the forecast is judged retrospectively by the closeness of the forecast to the true value. Often a measure of the deviation of the forecast value from the true value, such as the squared error, serves to evaluate this closeness. More generally, if a series of forecasts and actual values have been observed, the characteristics of the forecasts can be studied by looking at the joint distribution of forecasts and actual values and at various measures based on that distribution.

The goodness or quality of probabilities can be investigated in the same overall manner as any other forecasts. Typically, however, people find the evaluation of probabilities more difficult to think about than the evaluation of point forecasts. The underlying concepts are similar, but people are not used to thinking carefully about probabilities and the concepts are a little more difficult intuitively at first glance when applied to probabilities.

In evaluating a probability of an event retrospectively, it must be understood that the realization (what actually happens) is, on a numerical scale, either zero or one. If the event occurs, the actual value is one; if not, it is zero. Obviously a probability forecast is "perfect" if the probability is one and the event occurs or if the probability is zero and the event does not occur. Just as obviously, perfect forecasts are seldom attainable. The expert is usually uncertain about whether an event will occur, so the expert's probability is somewhere between zero and one.

The purpose of this chapter is to present and discuss methods that have been developed for evaluating the quality of experts' probabilities. The joint distribution of probabilities and actual values is the basic starting point for an evaluation of the probabilities, and various summary measures of that distribution can provide useful information. The joint distribution will be presented first, followed by discussions of some of the most important characteristics of interest and summary measures that provide information about those characteristics.

A.2 JOINT DISTRIBUTION OF PROBABILITIES AND OBSERVATIONS

The evaluation of an experts' probabilities involves the correspondence of those probabilities and the corresponding observations (what actually occurred). If an expert's probability for an event is denoted by r and the corresponding observation by x (where $x=1$ if the event occurs and $x=0$ if the event does not occur), the correspondence between the probabilities and observations can be investigated by looking at the joint distribution $g(r,x)$ of probabilities and observations. Ideally, this joint distribution would consist of only two points, $(r,x) = (1,1)$ and $(r,x) = (0,0)$; that is, all of the forecasts would be "perfect". In practice, however, that is not feasible.

The joint distribution can be factored into conditional and marginal distributions two ways (Murphy and Winkler, 1987, 1992):

$$g(r,x) = g\,(x \mid r)\, g(r), \tag{A-1}$$

and

$$g(r,x) = g(r \mid x)\, g(x). \tag{A-2}$$

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

- $g(x \mid r)$ is the conditional distribution of the observations, given a probability value, and this conditional distribution can be studied for all possible probability values r. The mean of this distribution is the relative frequency with which the event occurs when a probability value of r is given. If this relative frequency always equals r, then the probabilities are said to be perfectly calibrated, and large differences between the relative frequency and r imply poor calibration. Calibration is an important characteristic of probabilities that will be discussed in greater detail later in this chapter.

- $g(r)$ is the marginal distribution of the probabilities. This indicates, for instance, how often the expert uses extreme probabilities (probabilities near zero and one) and how often the expert uses less extreme probabilities. By itself, $g(r)$ only says something about r, not about the observations. Experts who give extreme probabilities more frequently are said to be more refined; refinement will be discussed at greater length later in the chapter.

- $g(r \mid x)$ is called a likelihood; for a given x (zero or one), it shows how likely different values of r are. This relates to discrimination. The discrimination of a set of probabilities is high if the expert tends to give high probabilities on occasions when the event actually occurs and low probabilities on occasions when the event does not occur. In other words, considering a weather forecaster, the forecaster with high discrimination does a good job of discriminating the rainy days from the nonrainy days (in advance, of course). Discrimination will be discussed further later in the chapter.

- $g(x)$ is simply the base rate, indicating how frequently the event occurs (i.e., $x=1$) in the data set. This base rate is sometimes used as a standard of comparison in the sense that the performance of an expert's probabilities are compared with the performance of a scheme that just takes the base rate as the probability every time. The base rate itself is not very informative, so it would be hoped that the expert would easily outperform the base rate.

The four components in the two factorizations involve different characteristics of the experts' probabilities and the situations for which the probabilities are assessed. Although each characteristic can be studied by itself, the entire "package" is important. For example, if an expert always gives probabilities of zero or one (high refinement), this is not very helpful if the relative frequency of occurrence of the event is 0.5 when the probability is zero and 0.5 when the probability is one (poor calibration). On the other hand, if the expert always says 0.5 and is perfectly calibrated, the poor refinement means that the probabilities are not at all informative despite the perfect calibration. Some summary measures based on $g(r,x)$ provide overall measures of accuracy that take into account all of the characteristics. These summary measures, called scoring rules, will be discussed after more detail is given regarding the individual characteristics.

A.3 CALIBRATION

Probabilities may be said to be good in the sense that they correctly reflect uncertainty. Much work has been accomplished on the quality of probabilities obtained from weather forecasters (Murphy and Winkler, 1977, 1984). One of the activities of a weather forecaster is to provide predictions of precipitation. Since 1965, U.S. Weather Service forecasters have provided these forecasts in the form of probabilities. Calibration refers to the extent to which probabilities assessed for events conform to the relative frequencies with which these events occur. Thus, the stated probabilities of precipitation should correctly reflect the relative frequency with which rain occurs. On those days that the weather forecaster announces a 60 percent chance of rain, for example, it should rain about 60 percent of the time. If for every forecast value (10 percent chance of rain, 20 percent chance of rain, etc.) the observed relative frequency of rain corresponds to the probability, then the forecaster is well calibrated.

One property of a set of well-calibrated probabilities is that a plot of the observed relative frequencies against the probabilities will depict a 45 degree line. Such a plot is called a calibration diagram. Figure A-1 is a calibration diagram for weather forecasters (Murphy and Winkler, 1977) that displays a remarkable degree of calibration. The companion plot, Figure A-2 (Christensen-Szalanski and Bushyhead, 1981), shows very poor calibration for probabilities assigned by physicians for the presence of pneumonia in patients.

The calibration of probabilities given for discrete events has traditionally been measured by "binning" similar probabilities and then comparing the relative frequency of events in the bin to the probability associated with the bin (Lichtenstein, Fischhoff, and Phillips, 1982). Thus, if several events are each assigned probabilities between 0.2 and 0.3, a bin for these events should have a relative frequency of occurrence somewhere between 0.2 and 0.3. In Figure A-2 the bins are defined by the probabilities 0.1, 0.2, etc.

The difference between the observed relative frequency of events in a bin and the expected relative frequency (i.e., the bin probability) quantifies the degree of calibration. The greater the discrepancy between the observed and expected frequencies, the poorer the calibration. The discrepancy between observed and expected frequencies in the bins has been modeled using a multi-nominal distribution (Murphy, 1973). Denoting the assessed probabilities by r_i, the observed relative frequencies by f_i, and the number of events assigned to the ith bin (assigned the probability r_i) by n_i, the lack of calibration can be measured by

$$C = \sum_{i=1}^{k} (r_i - f_i)^2 n_i / n, \qquad \text{(A-3)}$$

where k is the number of bins and n is the total number of events assigned probabilities.

C takes on the value zero when the assessed probabilities and the relative frequencies are equal, and can be expressed as a summary measure of the joint distribution of probabilities and outcomes: the expected squared difference between r and the relative frequency given r. As will be seen later, the expected score from a quadratic scoring rule can be decomposed into a series of terms, of which C is one.

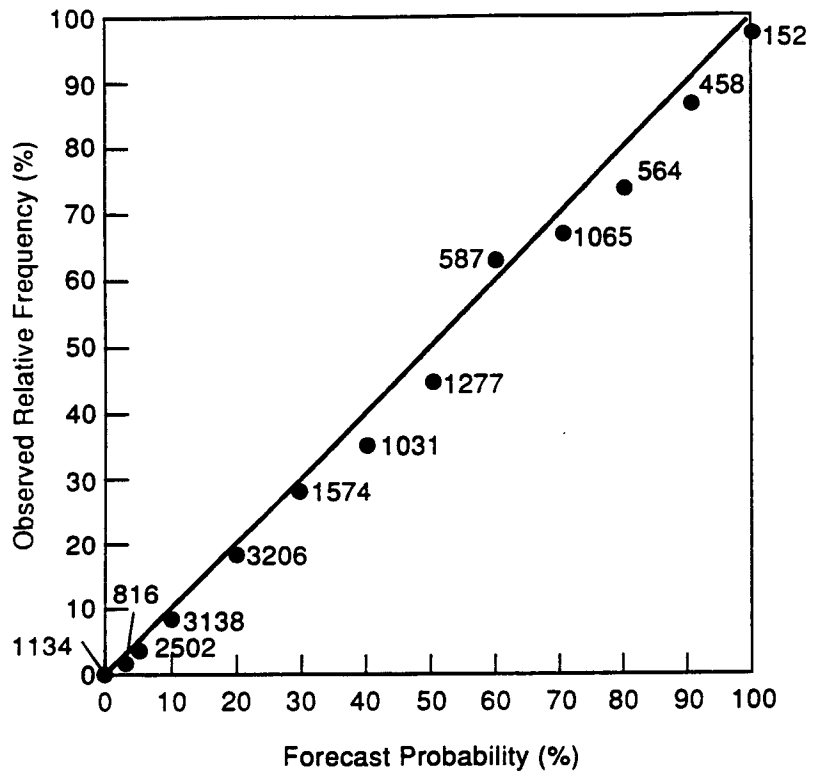


Figure A-1. Calibration plot for professional weather forecasters making probabilistic predictions of precipitation (Murphy and Winkler, 1977)

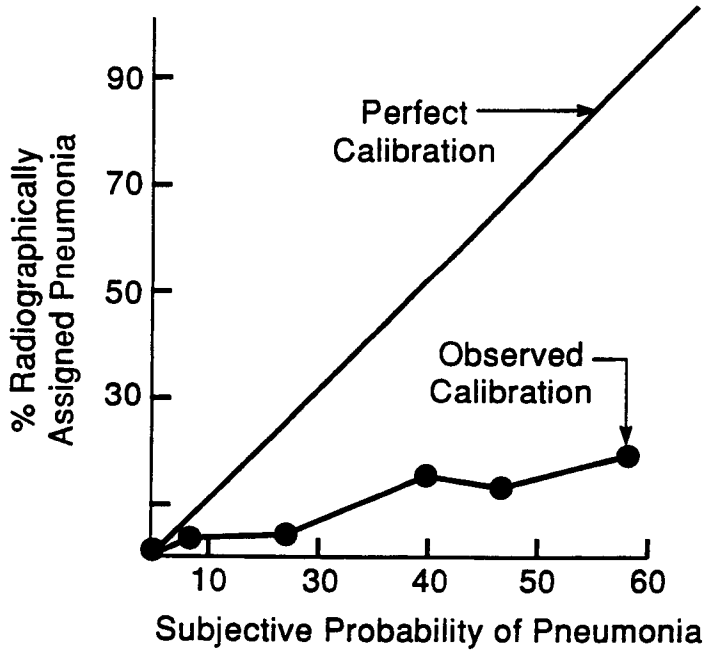


Figure A-2. Calibration plot for physicians probabilities for presence of pneumonia in patients (Christensen-Szalanski and Bushyhead, 1981)

Bins require a coarseness of probabilities that may be artificial in a given situation. That is, it may be necessary to bin somewhat dissimilar probabilities (0.5 and 0.7, for example) in order to have a sufficient number of events in each bin. The following measure of calibration avoids the necessity to bin or to have the same probability assigned to multiple events. Let E_j, for $j=1, \ldots$, be potential events, and let r_j be the assessed probability of E_j occurring. Let $I(z) = 1$ whenever $z \geq 0$ and $I(z)=0$ otherwise. Let $x(E)=1$ when E occurs and $x(E)=0$ otherwise. A measure of calibration is then

$$T=\max_{r\in[0,1]} \frac{\sum_{i=1}^{n} [x(E_i) - r_i]\, I(r-r_i)}{n}. \qquad \text{(A-4)}$$

If, for every r, the relative frequency of events assigned probabilities equal to or less than r equals the average of the probabilities assigned to these events, then $T = 0$.

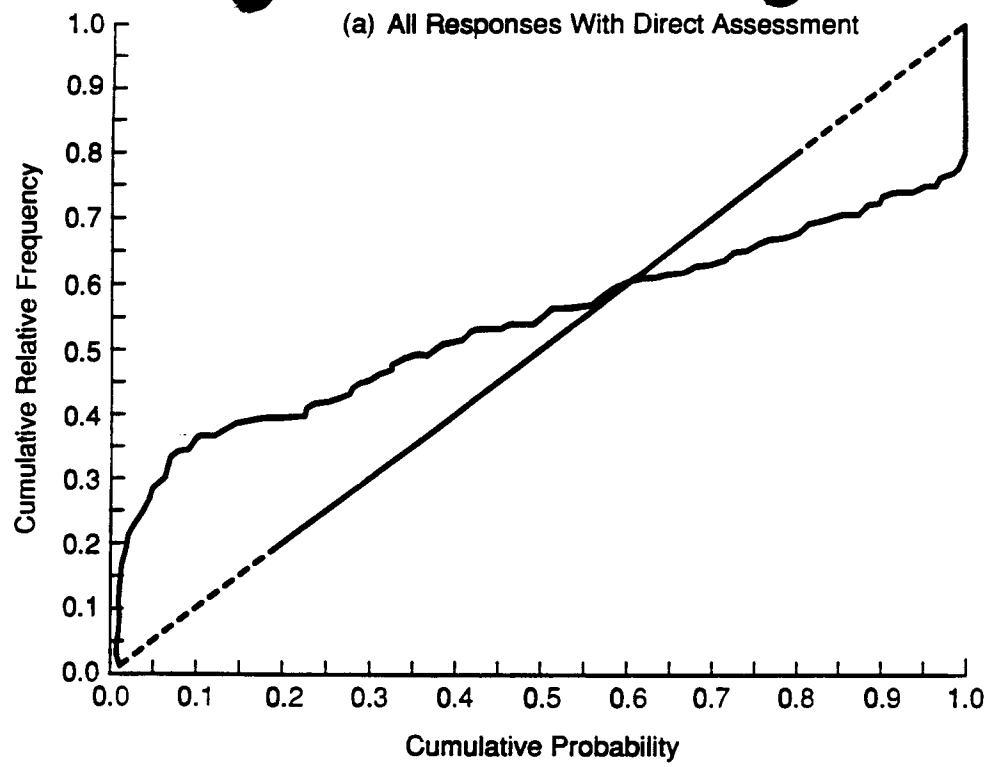
The assessment of probability distributions for continuous quantities is most often accomplished by assessing certain fractiles of the distributions such as the median, 0.25, 0.75, 0.05, and 0.95 fractiles. The remainder of the cumulative distribution function (CDF) is inferred from these points. For continuous quantities the underlying probability function is a density; thus, one cannot compute probabilities of individual values. Instead, probabilities for intervals of values can be computed.

The concept of calibration for continuous quantities, then, involves probabilities for intervals of values. One such approach is to develop bins corresponding to certain probability intervals. For example, continuous probability functions (densities) could be subdivided into intervals with equal probability, say 10 percent. Thus, equally probable bins delimited by the 0.0, 0.1, 0.2, etc., fractiles would be created. Calibration would be judged by the relative frequency of the true values in the bins. Ten percent of the uncertain quantities should turn out to be in each of the bins. Alternatively, assessed fractiles such as the 0.05, 0.25, 0.50, 0.75, and 0.95 fractiles could be used to define bins with probabilities 0.05, 0.20, 0.25, 0.25, 0.20, and 0.05.

Researchers in probability elicitation often chose to report only data on the extreme bins such as the lower and upper 0.05 bins. Values that turn out to be in these extreme bins are called "surprises." The literature abounds with examples where the number of surprises greatly exceeds the expected number of surprises (see Morgan and Henrion, 1990).

An approach to measuring and displaying calibration that avoids binning (Morgan and Henrion, 1990; Hora et al., 1992) is to use the cumulative probabilities of the observed values. For, example, if the CDFs F_1, F_2, ... are assessed for a series of variables and the resulting observed values are x_1, x_2, ..., then the quantities $p_i = F_i(x_i)$ should be uniformly distributed between zero and one. This is the same as saying that 5 percent of the quantities should appear in the lower 5 percent tails, 10 percent in the lower 10 percent tails, and so on.

A plot of cumulative probabilities obtained from scientists and engineers responding to almanac questions in the NUREG-1150 project (Hora et al., 1992) is given in Figure A-3. The horizontal axis is the cumulative probability of the true result while the vertical axis is the empirical cumulative relative frequency of the cumulative probabilities. The plot shows some lack of calibration, particularly in the tails. The lack of calibration exhibits itself as a deviation from the ideal 45 degree line — the line of



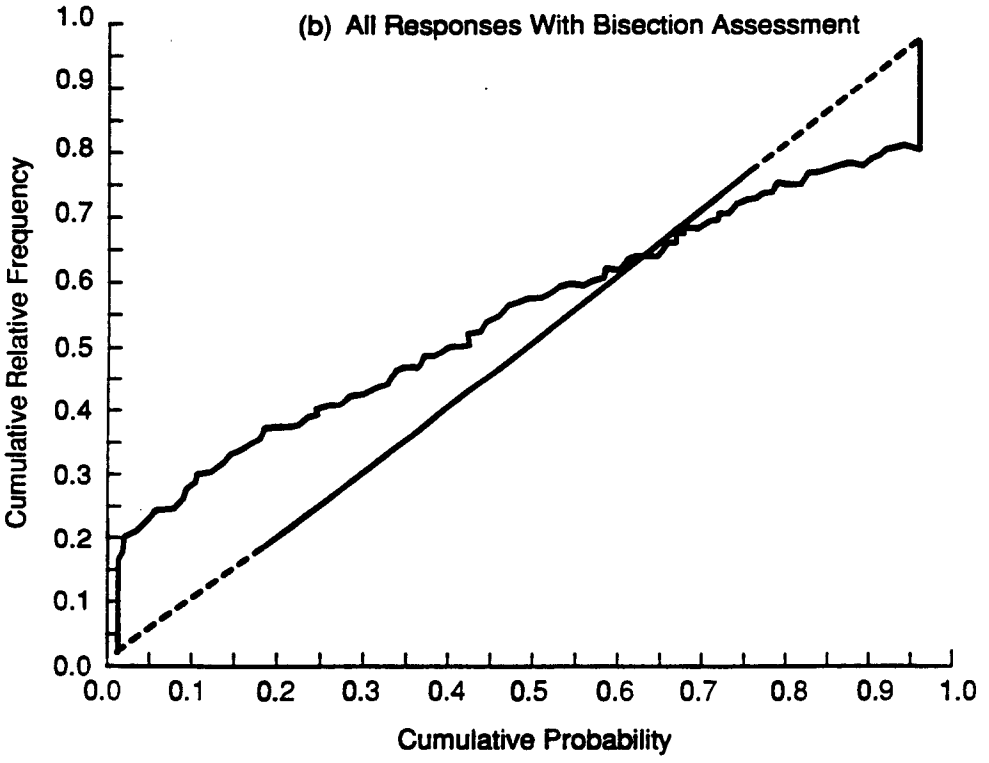


Figure A-3. Calibration plots for scientists and engineers responding to almanac questions (Hora et al., 1992)

uniformly distributed cumulative probabilities. A measure of the deviation from the 45 degree line, such as the maximum vertical distance from the 45 degree line, can be taken as a measure of calibration.

In summary, calibration is associated with the faithfulness of probabilities — how well they predict the relative frequencies of events. Calibration is not an end unto itself however, because calibration does not measure the information in distributions.

A.4 REFINEMENT

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

While both good calibration and high refinement are desirable properties, they are often in conflict as noted by von Winterfeldt and Edwards (1986). High refinement sometimes may be achieved only at the expense of good calibration, because the state of knowledge does not permit very extreme probabilities. In fact, it is possible to be well calibrated but have no refinement. A weather forecaster who gives the same forecast on every day may be well calibrated. If it rains on 20 percent of the days and the prediction each day is for a 20 percent chance of rain, the forecaster is well calibrated but has no refinement. One learns nothing from the forecasts that is not readily available from historical data. In contrast, if the forecaster predicts either a 100 percent chance or zero percent chance of rain, and the predictions are always correct, then the forecaster is both well calibrated and has perfect refinement — a clairvoyant.

Measures of refinement depend only on the assessed probabilities and not on the realizations of the events or random variables. Unlike calibration, refinement can be measured without reference to true values. Moreover, the refinement of a single probability or density can be measured while calibration as an empirical concept applies only to sets of probabilities or sets of probability distributions.

A.5 DISCRIMINATION

Calibration and refinement both focus on the probability values that are provided by the experts. Calibration involves how well these probability values agree with empirical reality in the form of relative frequencies, and refinement deals with the closeness of the probabilities to zero or one. In contrast, discrimination asks how different probability values discriminate between the occurrence and nonoccurrence of the events of interest, regardless of the actual numerical values themselves. For instance, suppose an expert always assigns probabilities of either 0.60 or 0.40; whenever the probability is 0.60, the event occurs, and whenever the probability is 0.40, the event does not occur. Although the numerical values of 0.60 and 0.40 do not seem very extreme, the expert in fact exhibits perfect discrimination. On the other hand, consider an expert who always assigns probabilities of 0.90 or 0.10; when the event occurs, it turns out that it follows a probability of 0.90 half the time and a probability of 0.10 half the time, which is identical to what happens when the event does not occur. This expert would seem, based only on the probability values of 0.90 and 0.10, to be a good candidate for high discrimination, but in fact there is no discrimination at all.

One way of investigating discrimination is through likelihoods. For a given value of a probability r for a single event, the two likelihoods of interest are $g(r \mid x=1)$ and $g(r \mid x=0)$, the likelihood of the expert providing the probability r if the event will occur and the likelihood of r if the event will not occur. If these likelihoods are equal, as in the second example in the preceding paragraph, then nothing is learned when the expert gives the probability r. The further apart the likelihoods are, the more discriminatory the probability r is.

Frequently the similarity of the likelihoods is measured via the likelihood ratio,

$$L(r) = g(r \mid x=1)/g(r \mid x=0). \tag{A-5}$$

A likelihood ratio of one (equal likelihoods) indicates no discrimination, while likelihoods less than or greater than one are more discriminatory as they move away from one. To avoid the asymmetry of this measure (e.g., likelihood ratios of 1/2 and 2 are equally discriminatory, even though the latter is twice as far from one), the log likelihood ratio is sometimes used instead:

$$\log L(r) = \log g(r \mid x=1) - \log g(r \mid x=0). \tag{A-6}$$

The overall discrimination of an expert can be displayed graphically in terms of a plot of L(r) or log L(r) as a function of r. The more this plot deviates from $L(r) = 1$ or $\log L(r) = 0$, the more discriminatory the expert's probabilities are.

Another graphical approach to discrimination is to plot $g(r \mid x=1)$ and $g(r \mid x=0)$ on the same graph. If the two distributions do not overlap at all (e.g., the probability is always over 0.40 when $x=1$ and always below 0.40 when $x=0$), the probabilities are perfectly discriminatory. At the other extreme, if the two distributions are identical, there is no discrimination. For nonoverlapping distributions, the likelihood ratio is always infinity or zero; for identical distributions, the likelihood ratio is always one. In terms of this graph, then, the degree of discrimination increases as the amount of overlap between the two distributions decreases.

Signal detection theory (Green and Swets, 1974) provides another way of measuring discrimination. A curve called the receiver operating characteristic curve is generated, and the area under the curve is taken as a measure of discrimination. This area can be interpreted as the probability that an expert's judgment is correctly able to discriminate between the event of interest occurring and not occurring. For details, see Green and Swets (1974) and Levi (1985).

Discrimination is a very appealing concept because it gets right to the heart of the matter, the ability of the expert to distinguish between occasions on which the event of interest will occur and occasions on which it will not occur. However, an underlying issue is that in order to take full advantage of this discrimination, the individual using the experts' probabilities will generally have to, in effect, take the probability value provided by the expert and treat it as though it were a different value. This is perfectly consistent with a Bayesian view of the world: an expert's probabilities provide information, and the individual using that information revises his or her own probabilities based on what the expert says. This can be thought of, in a way, as calibrating the expert to correct for miscalibration.

A.6 SCORING RULES

Scoring rules, (Winkler, 1967) are measures of the overall accuracy of assessed probabilities that are functions of calibration, refinement, and discrimination. Consider the instance of an expert providing probabilities for the occurrence of binary events. As before, suppose that the events are binned according to their probabilities. Let n_i be the number of events assigned the probability r_i and let f_i be the observed proportion of times events in this bin actually do occur. The first scoring rule proposed for such a situation is the Brier (1950) score. The Brier score for a response r_i when the associated event occurs is $(1-r_i)^2$ while the score when the event does not occur is r_i^2. Smaller scores are better. Averaging the scores over $n=n_1+n_2+\ldots+n_k$ responses binned into k categories gives the average score

$$S = \frac{1}{n}\sum_{i=1}^{k} [n_i f_i(1-r_i)^2 + n_i(1-f_i)r_i^2]. \qquad \text{(A-7)}$$

The average Brier score can be partitioned into three parts, corresponding to the difficulty of the assessments, the calibration of the expert providing the probabilities calibration, and the resolution of the assessments (Murphy, 1973). The partition is given by

$$S = \bar{f}(1-\bar{f}) + \frac{1}{n}\sum_{i=1}^{k} n_i(r_i-f_i)^2 - \frac{1}{n}\sum_{i=1}^{k} n_i(f_i-\bar{f})^2, \qquad \text{(A-8)}$$

where $\bar{f} = \sum n_i f_i/n$ is the overall proportion of times the events occur. The first term in the partition depends only on the proportion of events that occur (the base rate). Therefore, this term cannot measure any aspect of "goodness" of the probabilities; instead, it measures the empirical uncertainty about the occurrence and nonoccurrence of the events in question. If all the events do occur (or all do not occur), then this term is zero. If half the events occur, then the term reaches its maximum value of 1/4. The second term measures calibration through the squared difference between the assigned probabilities and observed relative frequencies. This term, which is just C from (A-3), is zero for a perfectly calibrated expert.

The third term measures a characteristic called the resolution of the probabilities, which relates to the differentiation of frequencies of events assigned to the various bins. For example, if events in each of the bins are found to have the same relative frequency of occurring, then the binning provides no resolution between more and less likely events. If, in contrast, the events are grouped so that some groups have high relative frequencies and some have low relative frequencies, then the resolution is high, implying that the binning (or assignment of probabilities) successfully distinguishes between more and less likely events. Note that the resolution term does not depend on the actual r_i values, but only on the grouping of situations into bins with different r_i values. In this sense, it is similar to discrimination.

While the Brier rule is perhaps the best known scoring rule, any strictly concave or strictly convex expected score function can be used to generate a "strictly proper" scoring rule (Savage, 1971). The term "strictly proper" is applied to such functions because of the property that an individual's expected score is optimized (minimized in the case of concave functions such as the expected Brier score, and maximized in the case of convex functions such as the logarithmic and spherical rules given below) whenever that individual responds with his or her honest probabilities. In contrast, consider the linear

scoring rule that gives a score 1-r when the event occurs and r when the event does not occur, where r is the assessed probability. (Compare to the Brier score, which gives scores $(1-r)^2$ and r^2.) If the expert actually believes the probability of the event occurring to be p, the expected score is maximized by responding $r=1$ whenever $p>.5$ and $r=0$ whenever $p<.5$. The expert is thereby "rewarded" for responding untruthfully, and the rule is called improper.

Some other proper scoring rules include the logarithmic rule,

$$S = \begin{cases} \log r \text{ if the event occurs,} \\ \log (1-r) \text{ if the event does not occur,} \end{cases} \tag{A-9}$$

and the spherical rule,

$$S = \begin{cases} \dfrac{r}{[r^2+(1-r)^2]^{1/2}} \text{ if the event occurs,} \\ \dfrac{1-r}{[r^2+(1-r)^2]^{1/2}} \text{ if the event does not occur.} \end{cases} \tag{A-10}$$

For these logarithmic and spherical rules, a higher score is better.

Each of these rules, including the Brier rule, which is quadratic, extends to events having several possible outcomes, and, by passing to a limit, extends to a corresponding rule for continuous variables (here the quadratic rule, like the others, is expressed so that a higher score is better). If f is the assessed density and x is the revealed or actual value of the variable, the rules for continuous variables can be expressed as follows:

- Quadratic: $$S = 2f(x) - \int_{-\infty}^{\infty} [f(u)]^2 du, \tag{A-11}$$

- Logarithmic: $$S = \log[f(x)], \tag{A-12}$$

- Spherical: $$S = \frac{f(x)}{\{\int_{-\infty}^{\infty} [f(u)]^2 du\}^{1/2}}. \tag{A-13}$$

The logarithmic rule has the property that the score depends only on the density at the realized value and not on the density at other values. While this is an advantage of the logarithmic scoring rule, the rule also assigns an infinite score if the realized value is assigned zero density, thus overwhelming any other assessments. Clearly, this is a drawback.

The above three scoring rules can be developed by direct extension from scoring rules for discrete events. Matheson and Winkler (1976) developed scoring rules for continuous distributions from another approach. They consider scoring rules for the binary events $x \le c$ and $x > c$. By allowing c to

vary over the entire range of x and integrating the scoring function with respect to c, they obtain another class of strictly proper scoring rules. One member of this class is another form of quadratic scoring rule,

$$S = \int_{-\infty}^{x} [F(u)]^2 du + \int_{x}^{\infty} [1-F(u)]^2 du, \qquad \text{(A-14)}$$

where F(u) is the assessed cumulative distribution function for x.

Some care must be used with scoring rules. It must be recognized that an average score developed with one set of questions is not comparable to an average score developed from another set of questions. An attempt to avoid this problem is represented by the asymmetric scoring rules developed in Winkler (1985, 1991). These strictly proper rules attempt to make scores from different situations comparable by scaling them appropriately. For instance, if the long-term relative frequency of precipitation at a given location at a particular time of year is 0.20, then a weather forecaster who gives a precipitation probability of 0.20 each day should get the worst average score. With the typical symmetric scoring rules discussed above, the worst score is obtained for a probability of 0.50, which is midway between zero and one, and the expected scores are identical for probabilities of, say, 0.20 and 0.80. For the asymmetric scoring rules, the worst score can occur at a probability other than 0.50, and the expected score improves as the probability moves from that value toward zero or one.

A.7 QUALITATIVE EVALUATION OF VALIDITY

Characteristics such as calibration, refinement, discrimination, and resolution are useful concepts for appraising the quality of probability assessments. They are useful in understanding the properties of "good" probability assessments. Empirical measures such as average scores and calibration measures are limited, however, to situations where the true outcomes are known. This will often not be the case in the assessment of radioactive waste disposal issues. More often than not, the quality of assessed probabilities must be judged by the quality of the experts themselves and the process used to acquire the probabilities.

Perhaps the foremost concern is the quality of the experts themselves. The quality of the experts may be judged along several dimensions.

- The experts should have possession or access to exceptional knowledge that sets them apart from others. In issues involving public safety, which are apt to receive strict review, the experts should be identifiable through their work and contributions in the subject area.

- The experts should be free from motivational biases. The outcome of the probability elicitation should not affect them. Individuals choosing experts for probability elicitation should be aware of possible economic ties to the issues under question. Sometimes these ties may be difficult to see. For example, a person's experimental program conceivably could be affected by the outcome of an elicitation exercise. Strong political positions may also carry with them a preference for certain outcomes, and, thus, the potential for bias.

- Those engaged in a probability elicitation exercise should be willing to provide the time and effort needed for a competent evaluation of the issues. This means studying the problem and documenting the rationales for the conclusions reached.

- When there are multiple scientific viewpoints or approaches to a problem, multiple experts should be used to ensure that diverse viewpoints are included. This will help capture the true range of uncertainty about the question.

The quality of probabilities depends, in part, on the way questions are asked. For example, questions should pass the "clairvoyance test." This test requires that the statement of the question be sufficiently complete so that a clairvoyant would be able to answer the question without further explanation (e.g., Spetzler and Stael von Holstein, 1975). Properly structuring and presenting questions helps to ensure that the expert is responding to the same question that is being asked. Moreover, questions should be asked in a manner free from suggesting or promoting certain answers; the judgments gathered should be influenced by analysts working with the experts only in the sense that assistance is provided in converting beliefs into probabilities. In general, a formal, well-structured elicitation process is important [see Morgan and Henrion (1990), Keeney and von Winterfeldt (1991)].

The quality of probabilities is therefore evaluated, at least in part, through the experts and the process used to collect the judgments. Here, documentation of all aspects of the process are vital. It is particularly important that the rationales and sources of information used by the experts be well documented. Numbers, by themselves, have little credibility. The methodology used to develop those numbers provides the support for their quality.

A.8 SUMMARY

The quality of probabilities can be investigated in terms of the joint distribution of probabilities and actual outcomes and characteristics such as calibration, refinement, discrimination, and resolution. Empirical measures of these characteristics are available but require the availability of appropriate data. Scoring rules provide overall measures of the goodness of probabilities through functions that incorporate the various characteristics of goodness. An investigation of all of the aspects of goodness can be useful in a diagnostic sense (Murphy and Winkler, 1992) to help experts improve their judgments, as well as in an evaluative sense.

Qualitative evaluation of expert probability distributions will most often be made on the basis of the process that generates the probabilities. The expert selection process, the training process, the statement of questions, the general elicitation procedures and the documentation, including the rationale behind the assessments all provide indirect indicators of the quality of the assessed probabilities.