

8 Eliciting Probabilities from Experts

*Steven Hora***ABSTRACT**

Decision models sometimes require the use of expert judgments to quantify uncertainties. Modes and methods for eliciting probabilities are presented in this chapter. Criteria for selecting issues and experts are discussed and protocols for acquiring these judgments derived from research and practice are described. These protocols are designed to eliminate problems that sometimes arise when using experts for uncertainty quantification. A range of considerations for organizing a probability elicitation is presented as a guide for developing a protocol for a given situation. The identification and avoidance of psychological biases arising from the application of heuristics is discussed as important background for probability elicitation. Probability training is presented as one method for reducing these biases. Measures of the goodness of assessed probabilities and probability distributions such as calibration measures and proper scores are given. Methods for completing a distribution from assessment fractiles are shown.

CONTENTS

Elicitation Modes

Probabilities of Events

Assessment for Continuous Quantities

Measuring the Quality of Assessed Probabilities

Heuristics and Biases in Forming Probability Judgments

Eliciting Probabilities

Selecting and Posing Questions

Selecting Experts

Training

Organizing an Expert Judgment Process

The ability of decision analysis to deal with significant uncertainties is one of its most attractive features. Probabilities or probability distributions provide the primary means for integrating uncertainties into a decision model. Often times, however, the acquisition of probabilistic information is, itself, a challenge. One method of developing probabilistic input for a decision model is to engage experts who have special knowledge about the likelihood of values or events in question. In order that the judgments of these experts be integrated with other types of information in the decision analysis, they should be codified as probabilities. Probabilities are the natural mathematical language of uncertainty and lend themselves to mathematical manipulations that cannot be accomplished with less rigorous expressions of uncertainty. The process of obtaining probabilities from knowledgeable individuals is sometimes called “expert judgment elicitation.”

Probabilities can be elicited in several different forms. The least complex situation is when the uncertainty is about the occurrence or non-occurrence of an event. In such a situation a single probability summarizes the expert’s judgment about the likelihood of the event. The next level of complexity involves an event that may resolve into more than two outcomes such as the winner of a horserace. Sometimes, the uncertainty is about a numerical quantity that has an infinite number of possible values. For example, the winning time in the before mentioned horse race. For such a quantity, we require the expert to provide a probability density function or its integral,

the distribution function. Because there are an infinite number of potential values, one must use other means than individual probabilities to express the judgment about uncertainty.

Eliciting judgments from experts is only one method for obtaining probabilities. When there is sufficient historical data and the stability of the process that generated the data can be guaranteed, these data should be used to develop probabilities or probability distributions rather than using experts. Expert judgment is most useful in situations where the evidence for mechanically estimating probabilities or distributions is incomplete. Expert judgment should not be used, however, when there is absence of a basis for the making judgments. This basis can be composed of data, models, analogues, theories, physical principles, etc., but without such a basis the judgments are mere guesses.

The impetus for the development of theories and methods for expert judgment has come from four directions. The development of Bayesian statistical methods and the subjectivist interpretation of probability have played an important role. For the Bayesians, it is the requirement of a prior distribution in order to produce a posterior distribution that has been the impetus. The prior distribution is inherently subjective in nature (meaning its philosophical interpretation) and thus is consistent with probabilities given by experts which are also inherently subjective. Subjectivists such as de Finetti (1974) and Savage (1954) provided theoretical underpinnings for the use of personal or subjective probabilities in decision making contexts.

Bayesian statistics has become a cornerstone of statistical decision theory (Raiffa & Schlaifer 1964), which in turn, has been major contributor to decision analysis. Decision researchers have made a major contributions to the techniques used to assess probabilities. Their motivations sprang from the practical need to represent uncertainties through probabilities in order to quantify decision models.

Cognitive psychologists have studied judgment formation about uncertain events. From their studies has emerged a large literature about cognitive biases in elicited probabilities. These biases are discussed later in the chapter.

A fourth area having made major contributions to probability elicitation is risk analysis. An important milestone here was WASH-1400, The Reactor Safety Study (United States Nuclear Regulatory Commission, 1975). In this major analysis of the safety of nuclear power generating stations, expert judgments were used extensively, and openly, to develop subjective probabilities of various events opening the way for greater use of expert judgment in public policy studies.

Elicitation Modes

Probability elicitation normally entails asking questions about events or about quantities.

Probabilities of Events

The simplest case to begin with is the uncertainty regarding an event that resolves into one of two states – the event occurs or does not occur. The required response is a single probability. It is important to distinguish this situation from a sequence of events where some events in the sequence resolve one way and others resolve another way. With a sequence of events there is a frequency of occurrence that is conceptually knowable and it is proper to create a probability distribution for this frequency. This is not the case for a non-repeating event – one that can occur only once. It is tempting to assign probabilities to probabilities or to use a range for a probability as if the probability had some physical or measurable value even if this is not the case. In any event, even if one was to assign probabilities to probabilities, expectations are linear in probabilities and neither expected values nor expected utilities would differ from those obtained using the mean of the probabilities. For a discussion of second order probabilities and their

meaningfulness see B. de Finetti (1977) and Skyrms, B. (1980). It is perhaps disconcerting that the expert's probability about a non-repeating event simultaneously carries information about the likelihood of the event and its uncertainty. These two aspects of knowledge about a non-repeating event are inseparable, however.

The most straightforward approach to obtaining a probability on a non-repeating event is to ask an expert for that numerical value. An expert who replies with "I don't know the value" is most likely thinking that there is some physical/measurable value that should be known but is not. The expert must be reminded that the probability is a degree of belief and does not have a true, knowable value. It is, instead, a reflection of the expert's knowledge about the likelihood of the event and will differ from expert to expert and over time as new information is acquired.

Sometimes indirect methods work better. Odds provide a simple re-expression of a probability and the two are easily calculated from one another. Odds require to a relative judgment about the likelihoods of an event and its complement rather than a direct judgment resulting in a numerical value. The judgment that an event is, for example, four times more likely to occur than not occur may be easier for the expert to feel comfortable with than is a direct judgment that the event has a .8 probability.

Another type of comparison is to a physical representation of the probability of an event and the probability of its complement. The probability wheel developed for use in probability elicitation by Stanford Research Institute requires such comparisons. This device provides a visual analogue to the probability of an event and its complement. The partial disks can be moved so that one segment is proportional to a number between .5 and 1.0 while the other segment is proportional to the complement. Which segment represents the event and which segment represents the complement is decided by which is more likely.

Beyond odds and the wheel, analysts have attempted to use verbal descriptors of likelihoods such as probable, rare, virtually certain, etc. Sherman Kent, head of the CIA's Office of National Estimates, proposed such a verbal scale (Kent 1964). But research has shown that these descriptors are interpreted quite differently by various individuals (Druzdel 1989).

Perhaps the best approach to events with multiple outcomes is to decompose the assessment into a number of assessments of events with binary outcomes. Judgments about such events are not as difficult to make. This is a "divide and conquer" approach and will result in coherent assessments. Decomposition can be accomplished through probability trees, influence diagrams (discussed in a later chapter), or even formulas. When probability trees are used, the assessed probabilities are conditional probabilities and marginal probabilities of the conditioning events or variables. Influence diagrams are discussed in Chapter 6. Expressing a target quantity as function of several other variables has been termed algorithmic decomposition.

There are many decompositions possible for a given problem. One should look for a decomposition that requires judgments that the expert is best prepared to make. Hora, Hora, and Dodd (1993) note that it is possible to over decompose a problem and make the assessment task more difficult. If time allows, one can look at several decompositions of a given problem and resolve inconsistencies among the recomposed probabilities. Ravinder, Kleinmuntz, & Dyer (1988) examine the propagation of error in subjective probability decompositions.

Many assessments are concerned with the value of a variable. Usually, these variables have a continuous range of potential values. Sometimes the range of a variable is bounded by physical consideration and sometimes it is unbounded. Also, one end of the range may be bounded and the other may be conceptually infinite. For example, the depth of snow at a given location is

bounded below by zero but has no well defined upper limitation. Unbounded variables are troublesome in that making judgments about the most extreme possible values is difficult.

In some instances, the decomposition of a probability assessment depends upon the source of uncertainties and the resolvability of those uncertainties. Knight (1921) made the distinction between "risk" (randomness with knowable probabilities) and "uncertainty" (randomness with unknowable probabilities). Today, these components of uncertainty are termed "aleatory" and "epistemic" uncertainties. Reliability Engineering & System Safety devoted a special issue to this subject (Helton & Burmaster 1996). Morgan and Henrion (1991) also discuss various components of uncertainty.

Assessment for Continuous Quantities

Assessment of continuous distributions is most often accomplished by having the expert provide a number of points on the distribution function (cumulative probability function). We denote such a point by the pair (p, v) where p is the probability that the quantity in question has a value no larger than v , that is $P(X \leq v) = p$ where X is the uncertain quantity. Of course, one is limited to assessing only a small number of such points, often less than ten. The pairs (p, v) may be denoted by v_p and are called fractiles or quantiles. We will use the term fractile although the two terms are used interchangeably. Fractiles can be assessed either by specifying p and asking for v or by specifying v and asking for p . Both techniques can be used in a probability elicitation although the first approach has less tendency to direct the expert towards specific values as explained later in the discussion of the anchoring bias. These two approaches, fixing p and asking for v and vice-versa, have been termed the p and v methods (Spetzler & von Holstein 1975).

A variation on the p method is called successively subdivision. It requires the expert to specify a value that breaks an interval into two equally likely subintervals. The process is

repeated several times until enough points on the distribution function are obtained to have a good idea of its shape. At the first subdivision, the expert is asked to provide a value such that the uncertainty quantity is equally likely to be above as below this value. This assessment yields $v_{.50}$, the median of the distribution. The second stage of successive subdivision entails dividing the range below the median into two equally likely sub-subintervals and similarly dividing the range above the median into two equally likely sub-subintervals. These subdivisions yield $v_{.25}$ and $v_{.75}$, the first and third quartiles respectively. There are now three values, $v_{.25}$, $v_{.50}$ and $v_{.75}$, that divide the range of possible values into four subintervals each having a probability of .25. Once again some or all of these intervals may be resubdivided until the level of fineness needed to represent the distribution is obtained.

An analyst may choose to use successive subdivision for the median and quartile values and then switch modes to direct of assessment using the p or v method for values needed to complete the distribution. Comparisons should also be made to check the consistency of results. For example, the analyst might ask whether it is more or less likely that the quantity is between $v_{.25}$ and $v_{.75}$ than outside that interval. This is a consistency check, and if the analyst finds that inside or outside the interval is more likely, it will be necessary to make some adjustments.

Once a sufficient number of points on a distribution function have been obtained, the distribution must be completed by interpolation or curve fitting, and if end points have not been obtained, extrapolation will be necessary. The simplest method of completing the distribution is to use linear segments between assessed values on the distribution function. This is shown in Figures 8.1a and 8.1b. The resulting density is a histogram as connecting the fractiles by linear segments spreads the probability evenly throughout an interval. Although the image may appear unpleasing compared to fitting some kind of curve through the assessments, it does have

advantage of maximizing the entropy of the distribution subject to satisfying the assessed fractiles. Entropy (Cover & Thomas 1991) is a measure on disorganization or uncertainty and is calculated from the density function:

$$I(f) = - \int_{-\infty}^{\infty} f(x) \ln f(x) dx \tag{1}$$

where $I(f)$ is the entropy of the density $f(x)$. By maximizing the entropy, the analyst has added as little information as possible into the completed distribution, that is, relative to any other distribution with the same fractiles, the distribution completed by linear segments has the least information in it.

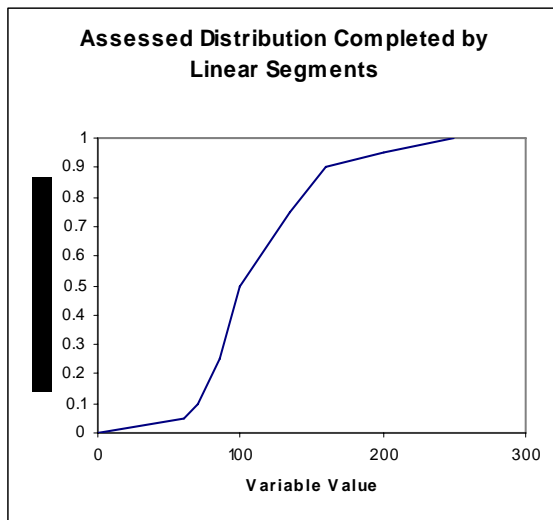


Figure 8.1a Distribution Function

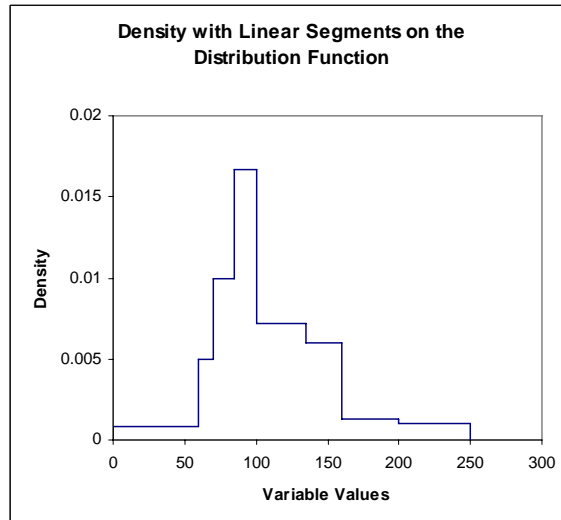


Figure 8.1b Density Function

The process of extrapolating to obtain endpoints is much more difficult to justify than interpolating to obtain interior values and there is no agreed upon method for doing so. In some decision analyses, the behavior of the distribution in the tail may be rather inconsequential while in others, the tail behavior may be critical. This is apt to be true in studies of low-probability, high consequence hazards such as terrorism and technological hazards. Sadly, the behavior in the tail may be both of greatest interest and most challenging to assess.

An alternative to the p and v methods for eliciting a subjective distribution is that of selecting a parametric family of distributions and asking the expert to provide judgments about the parameters, either directly or indirectly through fractiles or other assessments. This is a shortcut method in that the number of assessments required is usually greatly reduced. The costs of this method are that an arbitrary family of distributions is imposed on the expert's judgment and the subjective estimates of the parameters may require more difficult judgments and thus greater error introduced into the assessment process. For example, judgments about a mean may be much more difficult to make than judgments about a median or mode when a distribution is asymmetric. Likewise, judgments about standard deviations may be more difficult to make than judgments about interquartile ranges.

It is sometimes desirable to encode elicited fractiles into a specific parametric family of distributions such as the gamma, beta, or normal family of distributions. A general method for making this conversion is to minimize the L_2 -norm between the assessed density and the parametric density. Figure 8.2 displays a distribution assessed by fractiles and a parametric distribution, in this case a gamma distribution.

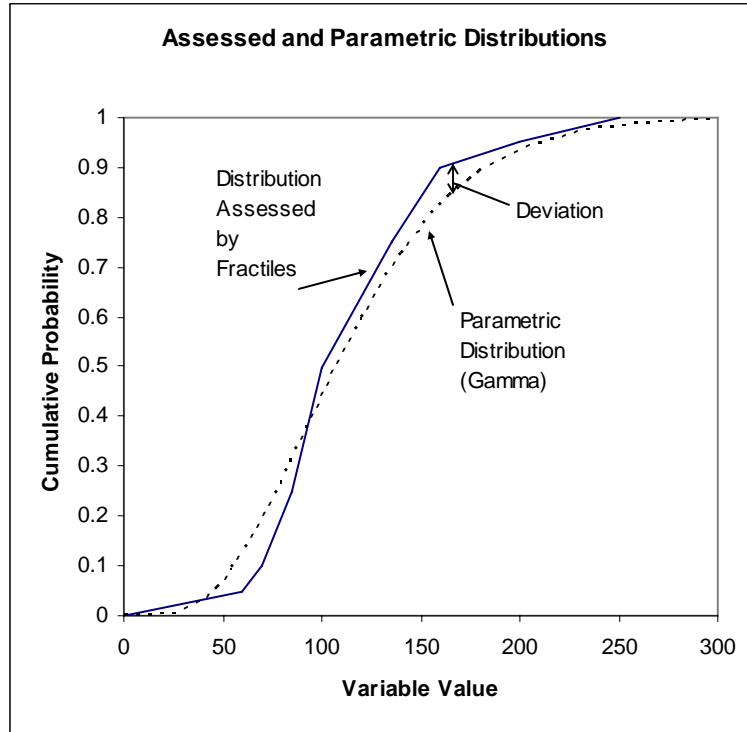


Figure 8.2 Assessed and Parametric Distributions

Denoting the assessed density function by $g(x)$ and the parametric density by $f(x|\theta)$, the problem is to find the value of the parameter vector θ that minimizes:

$$L_2[g(x), f(x|\theta)] = \left[\int_{-\infty}^{\infty} [g(x) - f(x|\theta)]^2 dx \right]^{\frac{1}{2}} \quad (2)$$

We denote the assessed fractiles by x_{p_i} , where p_i is the cumulative probability associated with the i^{th} fractile. When the assessed distribution is completed by the method of linear segments, the L_2 -norm can then be expressed as:

$$L_2[g(x), f(x|\theta)] = \left[\sum_{i=1}^{m-1} \frac{(p_{i+1} - p_i)^2}{(x_{p_{i+1}} - x_{p_i})} - 2 \sum_{i=1}^{m-1} \frac{(p_{i+1} - p_i)}{(x_{p_{i+1}} - x_{p_i})} [F(x_{p_{i+1}}|\theta) - F(x_{p_i}|\theta)] + \int_{-\infty}^{\infty} f^2(x|\theta) dx \right]^{\frac{1}{2}} \quad (3)$$

where m is the number of fractiles, by assumption $p_1 = 0$ and $p_m = 1$, and $F(x|\theta)$ is the distribution function associated with the density $f(x|\theta)$.

Minimizing the L_2 -norm with respect to θ is equivalent to minimizing the right hand side of:

$$L_2[g(x), f(x|\theta)]^2 - \sum_{i=1}^{m-1} \frac{(p_{i+1} - p_i)^2}{(x_{p_{i+1}} - x_{p_i})} = -2 \sum_{i=1}^{m-1} d_i [F(x_{p_{i+1}}|\theta) - F(x_{p_i}|\theta)] + \int_{-\infty}^{\infty} f^2(x|\theta) dx \quad (4)$$

where $d_i = \frac{(p_{i+1} - p_i)}{(x_{p_{i+1}} - x_{p_i})}$ is the density in the interval between successive fractiles. Solving the

minimization problem can be accomplished as a spreadsheet exercise if one has access to distribution function and knows the parametric form of the last term in the right of (4). This last term, the integral of the squared density, is the “expected density” and is given in Table 1 for some density functions.

Table 8.1. *Density Functions*

Name	Density	Expected Density
Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	$\frac{1}{2\sqrt{\pi}\sigma}$
Gamma	$\frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta}$	$\frac{\Gamma(2\alpha - 1)}{\beta 2^{2\alpha-1} [\Gamma(\alpha)]^2}$
Beta	$\frac{1}{\beta(a,b)} x^{a-1} (1-x)^{b-1}$	$\frac{\beta(2a - 1, 2b - 1)}{[\beta(a,b)]^2}$
Weibull	$\frac{\alpha}{\beta} x^{\alpha-1} e^{-(x/\beta)^\alpha}$	$\frac{\Gamma(2 - \frac{1}{\alpha})}{2^{\frac{1}{\alpha}}} \frac{\alpha}{\beta}$

We note that the expected density may not exist for some parameter values. Specifically, for the gamma and Weibull densities, α must be greater than .5 and, for the beta density, both a

and b must be greater than .5. The following graph shows a gamma distribution as fitted by the method to the fractiles shown in Figure 8.3.

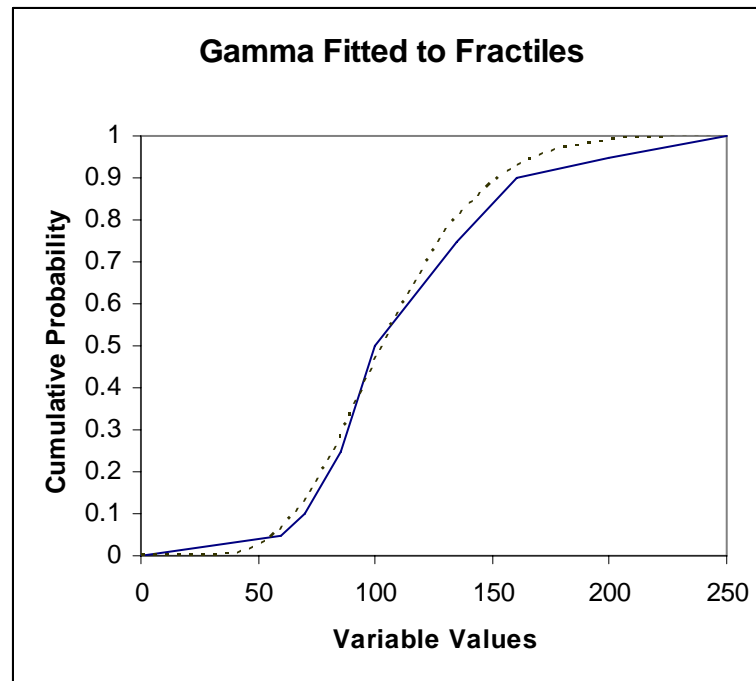


Figure 8.3 Gamma fitted to fractiles

A further complication arises when there are multiple variables that are probabilistically dependent. Dependence may be the product of physical relationship or it may result from shared knowledge about the uncertain quantities. For example, gold and copper ores are often found together so that knowing the amount of copper present in an ore may provide information about the amount of gold in a given ore. This is an example of a physical relation. Conversely, knowing the atomic weight of copper may provide some insight into the atomic weight of gold. Thus being informed of the atomic weight of copper, one would modify their uncertainty distribution for the atomic weight of gold.

Capturing dependence in probability assessments can be a daunting task. A general method for assessing the joint distribution of two or more variables is to break the assessment task into

stages with marginal and conditional distributions being assessed. It may be necessary, however, to make many conditional assessments as the distributions given the value of the conditioning variable. Influence diagrams can provide a mechanism for organizing this effort. There are several ways to reduce the assessment effort, however. The first is to assume a particular family of distributions such as the multivariate normal or Dirichlet can faithfully represent the judgments of the experts and then concentrate on assessments that will allow one to derive parameters of the joint distributions such as means, variances, and covariances. A more general approach that permits the marginal distributions to be general in form is to employ a copula (Clemen & Reilly 1999, Clemen, Fischer & Winkler 2000). Copulas are discussed in the next chapter in the context of capturing dependence among experts.

Measuring the Quality of Assessed Probabilities

Because subjective probabilities are personal and vary from individual to individual and from time to time, there is no “true” probability that one might use as a measure of the accuracy of an elicited probability. A weather forecaster who provides a probability of precipitation of .5 cannot be wholly right or wrong. Only probabilities of 1.0 and 0.0 can be held to such a standard. There are, however, two properties that are desirable to have in probabilities:

- Probabilities should be informative
- Probabilities should authentically represent uncertainty

The first property, being informative, means that probabilities closer to 0.0 or 1.0 should be preferred to those closer to .5 as the more extreme probabilities provide greater certainty about the outcome of an event. In a like manner, continuous probability distributions that are narrower or tighter convey more information than those that are diffuse. The second property, the appropriate representation of uncertainty, requires consideration of a set of assessed

probabilities. For those events that are given an assessed probability of p , the relative frequency of occurrence of those events should approach p . Perhaps this is most easily understood by looking at a graph of probabilistic precipitation forecasts (horizontal axis) and observed relative frequencies (vertical axis) as reported in Murphy and Winkler (1977), Figure 8.4a, and a graph of medical diagnoses report in Christensen-Szalanski and Bushyhead (1981), Figure 8.4b, where the axes are the probability of pneumonia assigned by a physician in an initial diagnosis and a latter diagnosis by x-ray.

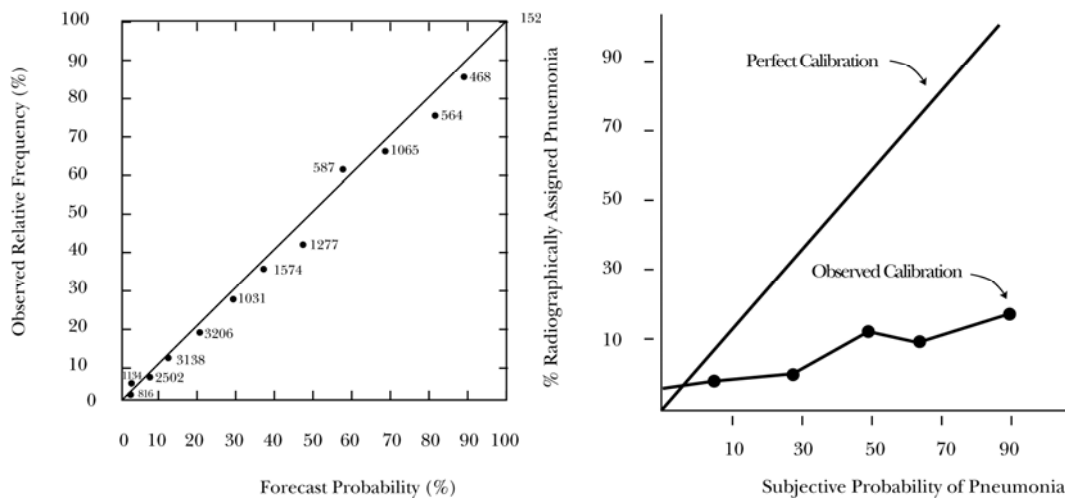


Figure 8.4. Two calibration graphs: The left panel (Fig.8.4a) shows the calibration of weather forecasters (Murphy and Winkler, 1977); the right panel (Fig. 8.4b) shows the calibration of medical doctors (Christensen -Szalanski and Bushyhead (1981)

Ideally, each graph would have a forty-five degree line indicating that the assessed probabilities are faithful in that they correctly represent the uncertainty about reality. The weather forecaster graph shows a nearly perfect relation while the graph for the physicians shows very poor correspondence between the assessed probabilities and relative frequencies. The graph is not even monotonic.

Graphs showing the relation between assessed probabilities and relative frequencies are called calibration graphs and the quality of the relationship is loosely called calibration which can be good or poor. Calibration graphs can also be constructed for continuous assessed distributions. Following Hora (2004), let $F_i(x)$ be a set of assessed continuous probability distribution functions and let x_i be the corresponding actual values of the variables. If an expert is perfectly calibrated, the cumulative probabilities of the actual values measured on each corresponding distributions function, $p_i = F_i(x)$, will be uniformly distributed on the interval (0,1). Figure 8.5 below shows a graph of responses of experts to almanac questions (see Hora, Hora and Dodd, 1992).

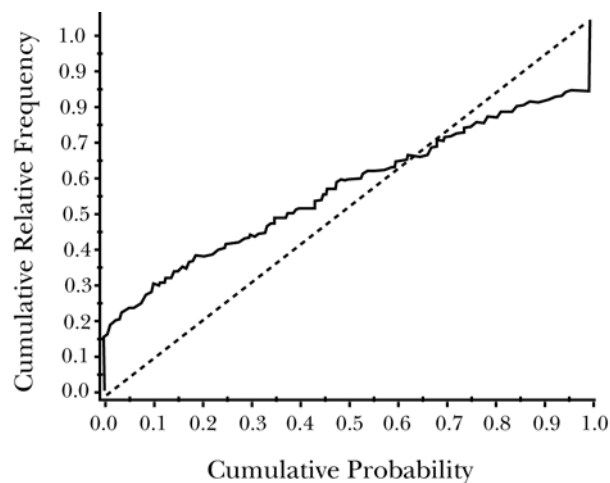


Figure 8.5 Assessment of probability

There is a clear departure from the ideal forty-five degree line. The steep rise of the graph at both extremes is indicative of distributions that were not spread widely enough to embrace the true quantities. Hora (2004) proposes using the area between the forty-five degree line of perfect calibration and the observed calibration curve as a measure of miscalibration.

Although calibration is an important property for a set of probabilities or probability distributions to possess, it is not sufficient as the probabilities or probability distributions may not be informative. For example, in an area where it rains on 25% of the days, a forecaster who always predicts a 25% chance of rain will be perfectly calibrated but provide no information from day-to-day about the likelihood of rain. But information and calibration are somewhat at odds. Increasing the information by making probabilities closer to zero or one or by making distributions tighter may reduce the level of calibration. Strictly proper scoring rules are functions of the assessed probabilities and the true outcome of the event or value of the variable that measure the goodness of the assessed distribution and incorporate both calibration and information into the score. The term “strictly proper” refers to the property that the expected value of the function is maximized when the probabilities or probability functions to which the function is applied are identical to the probabilities or probability functions that are used to take the expectation. An example will clarify.

A simple strictly proper scoring rule for the assessed probability of an event is the Brier or quadratic rule:

$$\begin{aligned} S(p) &= -(1-p)^2 \text{ if the event occurs} \\ &= -p^2 \text{ if the complement of the event occurs.} \end{aligned}$$

where p is the assessed probability. For any probability q , the expectation

$E_q(S(p)) = -q(1-p)2 - (1-q)p^2$ is maximized with respect to p by setting $p = q$. Thus, if an expert believes the probability is q , the expert will maximize the perceived expectation by responding with q . In contrast, the scoring rule $S(p) = p$ if the event occurs and $S(p) = 1-p$ while intuitively pleasing does not promote truthfulness. Instead, the expected score is maximized by providing a probability p of either 0.0 or 1.0 depending on whether q is less than or larger than 0.5. Winkler (1996) provides a discussion of this Brier rule and other strictly proper scoring rules. Also see Lichtenstein, Fischhoff, & Phillips (1982) and Cooke (1991).

The concept of a strictly proper scoring rule can be extended to continuous distributions (Matheson and Winkler 1976). For example, the counterpart to the quadratic scoring rule for continuous densities is:

$$S[f(x), w] = 2f(w) - \int_{-\infty}^{\infty} f^2(x)dx \quad (5)$$

Expected scores can sometimes be decomposed into recognizable components. The quadratic rule for continuous densities can be decomposed in the following manner. Suppose that an expert's uncertainty is correctly expressed through the density $g(x)$ but the expert responds with $f(x)$ either through inadvertence or intention. The expected score can be written as:

$$E_g \{S[f(x), w]\} = I(f) - C(f, g) \quad (6)$$

where $I(f) = \int_{-\infty}^{\infty} f^2(x)dx$ and $C(f, g) = 2 \int_{-\infty}^{\infty} f(x)[f(x) - g(x)]dx$

$I(f)$ is the expected density associated with the assessed distribution and is a measure of information. $C(f, g)$ is a strictly non-negative function that increases as $g(x)$ diverges from $f(x)$. Thus $C(f, g)$ is a measure of miscalibration. Further discussion of decomposition can be found in Lichtenstein, Fischhoff, and Phillips (1982), Murphy (1972, 1973).

Haim (1982) provides a theorem that shows how a strictly proper scoring rule can be generated from a convex function. See also Savage (1971).

Heuristics and Biases in Forming Probability Judgments

The process of expressing one's knowledge in terms of probabilities is not simple and has shown to be subject to some repeatable types of errors. Cognitive psychologists have detected, classified, and analyzed these errors in experimental settings. This work dates back to the 1970's and was led by Kahneman and Tversky (Bar-Hillel, 2001). Kahneman, Slovic, & Tversky (1982) provide a compilation of research in this area up to 1982 while Gilovich, Griffin & Kahneman (2002) contains contributions from the next two decades.

Judgmental errors are thought to be related to the way information is processed, that is, the heuristics used in forming the judgments. Two predominant heuristics have been labeled the representiveness heuristic and the availability heuristic.

Representativeness is the process of using some relevant cues to associate a target event or quantity with a similar set of targets. M. Bar-Hillel (2001) notes, however, that similarity judgments obey different rules than probability judgments. For example, subjects given a description of what could possibly be a location in Switzerland respond with a higher probability that the location is in Switzerland than the probability that the location is in Europe (Bar-Hillel & Neter, 1993) . Such a judgment is irrational as one event is entirely included within another (being in Switzerland implies being in Europe.)

Another manifestation of the representiveness heuristic is termed the base rate bias. This is illustrated by the following situation taken from a CIA publication. During the Viet-Nam War, a U.S. fighter pilot is strafed by a fighter plane that is either Cambodian or North Vietnamese. The

pilot can correctly identify a plane's nationality with 80% accuracy meaning that a Cambodian will be correctly identified as Cambodian with a probability of .8 and a North Vietnamese aircraft would be correctly identified with a similar .8 probability. There are six times as many Vietnamese aircraft in the fray as Cambodian aircraft. The pilot identifies the plane as Cambodian. What probability should be assigned to the strafing aircraft being Cambodian? The representiveness heuristic might lead one to assign a probability of .8, as this is "representative" of the pilot's accuracy. But this assessment fails to take into account the base rate or background frequency of the two nations' aircraft. A correct assessment would be based on Bayes' theorem and would result in a probability of the strafing aircraft being Cambodian of 4/9.

Other effects related to the representativeness heuristic include failure to regress, failure to consider sample size, and incorrect interpretations of randomness. Failure to regress occurs when one fails to consider that values greater or lesser than the mean will tend to return to the mean as a process continues. A baseball player who hits .400 for the first twenty games of the season will likely end the season with a lesser batting average. Using the current batting average as a best estimate of the future average would entail this failure. Problems with sample size occur because people do not correctly adjust for the reliability of data. People will often make judgments about the accuracy of a sample by examining the proportion of a population included in the sample rather than examining the variability within the population and the absolute size of the sample which are proved to be the determinants of the error in an estimate. Finally, people tend to find fictitious structure in random events. For example, when asked which is more likely to occur in the flip of five coins, many will answer that THTHTH is more likely than HHHHH although both have the same probability. People often conclude that someone is on a hot streak which is likely to continue when they have made several free throws or won in craps.

A second major class of biases arises from the availability heuristic. Availability refers to the ability to access or recall information. Cues that are easier to recall tend to be given more weight in forming probability judgments. An experiment that illustrates this effect entails showing a subject a list of names of celebrities. Manipulating the list so that the members of one sex are somewhat more famous than the members of the opposite sex will result in the subject overestimating the relative frequency of names on the list belonging to the sex having the more famous members (Tversky & Kahneman 1973). Another manifestation of availability occurs when subjects are asked to estimate the relative frequency of various causes of death in the United States. Subjects tend to overestimate the frequency of sensational causes and underestimate the frequency of mundane causes. For instance, the likelihood of death from a stroke will be underestimated while the likelihood of death from firearms will be overestimated. Information about death from firearms is more widely publicized – every such death will receive media attention while death from stroke is likely to be reported only in the case of well known individuals. Thus, the mind finds it easier to recall instances of death from firearms or, just as importantly, overestimates the frequency by overestimating the number of such instances that could be brought to mind if one really tried (Fischhoff & MacGregor 1982.).

A form of anchoring, particularly salient to probability elicitation, occurs when a subject gives too much credit to a reference point so that other possible references or the fallibility of the selected reference is ignored (Tversky & Kahneman 1974). For example, two experts who are equally qualified and have conducted studies to determine a particular quantity are likely to give more credit to their own work and less credit than is warranted to their colleague's work. Anchoring can also occur when there is sparse evidence. The expert may rely too heavily on evidence that is available and ignore how new evidence could change our judgments. Consider a

situation where there are two known sources of information, each leading to a somewhat different estimate of quantity, and there is a third source of information not known to the expert at this time. Let the three pieces of evidence be represented by dots on a line as shown below.

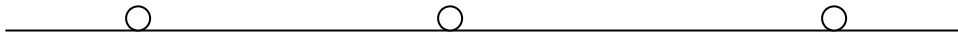


Figure 8.6. Illustration of the anchoring bias

Suppose that one of the dots is initially invisible. If the invisible dot is chosen randomly from the three dots, it is twice as likely that the invisible dot is outside the range of the visible dots. Thus, the appearance of the a third piece of information is twice as likely to spread the range of possibilities as to confirm the range if one “anchors” the range using the observable values. The paradox is that in judgment formation, information may result an increased expression of uncertainty so that the more we know, the less we say we know.

Anchoring can also occur as the result of a scale or instrument that suggests values for a variable. Subjects are uncomfortable going outside of the set of suggested values. Moreover, the subject is drawn to the center of the scale as a representative or central value. Hora, Hora, & Dodd (1972) and Winman, Hannson, & Juslin (2005) have shown that asking for probabilities of fixed intervals rather than asking for intervals with fixed probabilities can result in a loss of calibration.

Another scale effect can occur when there are discrete, mutually exclusive outcomes and one of the outcomes is a catchall, everything else category. For example, in a hospital study, a description of an admittee’s symptoms was given along with a list of possible diagnoses. Two

lists were used, one list with four named diagnoses and a catchall category to include all unnamed diagnoses. The second list had three possibilities, two of the four named diagnoses in the first list and a catchall category that implicitly contained the missing two diagnoses included on the first list. The result is that subjects given the second list gave a lower probability for the catchall category than corresponding probabilities given to the two missing diagnoses and the catchall category on the first list. They should be equivalent. The availability of the additional diagnoses has been attributed as the reason for this incongruity. This has been termed the packing effect. One of the first studies in this area involved an experiment using automobile fault trees (Fischhoff, Slovic, & Lichtenstein 1978). In this study the some potential causes of automobile failure were omitted from a diagram rather than being relegated to a catchall category. The subjects were asked to estimate the total probability of the missing causes and consistently underestimated this probability. In this study, the bias was termed the pruning effect. Fox & Clemen (2005) provide an alternative explanation of this bias.

The psychological bias that has received the greatest attention in probability elicitation is the tendency to assign probabilities that are too close to zero or one or to give uncertainty distributions that are too narrow. This bias is termed overconfidence or apparent overconfidence as the effect is to provide answers that are more certain than is warranted. Much of the research in this area has employed probabilities for binary events based on almanac data or sensory experiments. For example, a subject might be asked to name the higher mountain, Mt. Fuji or the Matterhorn, and then assign a probability to the subject's answer being correct (Lichtenstein & Fischhoff 1977). Keren (1988) conducts a similar experiment but asks the subjects to identify letters while manipulating the difficulty of the task. While subjects tend to assign probabilities

that are too high, the effect is much more pronounced when the questions are, in some sense, difficult rather than easy.

There have been many studies that have examined elicited distributions for continuous quantities. Lichtenstein, Fischhoff, & Phillips (1982) provide a table of the findings of such studies by concentrating on the faithfulness of the tail probabilities. In each of the studies, some extreme fractiles, such as the .01 and .99 fractiles were assessed and in most of the studies the interquartile range was assessed. The studies employed almanac data with known answers or values that would be known shortly, such as football scores (Winkler 1971). Almost uniformly, each study reported that the number of times the target values occurred in these extreme tails was greater than would be indicated by the probabilities – usually much greater. For example, in a study of stock price prediction by graduate business students, Stael von Holstein (1972) reports that 30% of the responses fell above the .99 fractile or below the .01 fractile while 27% fell in the interquartile range rather than the desired 50%. Klayman, et al. (1999) find that overconfidence is more pronounced in assessment of continuous quantities (interval assessments) than the assessment of simple questions such as two choice questions.

Overconfidence can be very troubling where a correct expression of risk is needed. In such circumstances it is wise to educate the expert about this bias and give some training and feedback of the expert's performance. While not every individual will exhibit this bias, it appears to exist in every group of subjects, whether they be experts or students. There is ample evidence (Alpert & Raiffa 1982) to show that probability exercises and feedback will reduce this bias, although it will not entirely eliminate it.

Overconfidence is associated with the hard-easy effect wherein subjects responding to more difficult questions exhibit more overconfidence than those responding to easy questions

(Lichtenstein & Fischhoff 1977). Apparent overconfidence observed in almanac tests has been explained as a manifestation of the hard-easy effect by Juslin (1994). See also Gigerenzer (1991) and Gigerenzer, Hoffrage, & Kleinbolting (1991). This area is, to date, controversial and is reviewed in Brenner, Koehler & Liberman (1996).

Another active area of research is support theory which was introduced by Tversky and Koehler (1994) to explain why elicited probabilities depend on the manner of presentation of the issue and apparently do not conform to the normative principle of additivity of probabilities across mutually exclusive events. The theory envisions probability judgment formation as consisting of three sets of elements: mutually exclusive hypotheses, evidence that supports the hypotheses, and expressed probabilities. The hypotheses are descriptions of events rather than the events themselves. The evidence or support is the perceived strength of the evidence for particular hypothesis. The judged probability is the weight in favor of a particular hypothesis relative to the weights of all hypotheses.

The weight of evidence, or support for a hypothesis A is given by $s(A)$. The real content of support theory is in the assumption that support is subadditive. Mathematically, if A_1 and A_2 are a partition of A , that is $A = A_1 \cup A_2$ and $A_1 \cap A_2$ is empty, then support theory requires that $s(A) \leq s(A_1) + s(A_2)$. If the inequality holds strictly, then one will find the judged probability of an event to be less than the sum of the probabilities of the constituents elements of the partition of that event. Of course, this is relevant to probability elicitation as the way the question is posed may influence the response given.

To make things more concrete, let A be homicide, A_1 be homicide by an acquaintance, and A_2 be homicide by a stranger. Subadditivity operates if $s(A) \leq s(A_1) + s(A_2)$ and this has been shown to be the case in numerous studies (Tversky & Koehler 1994, Fox & Tversky 1998). A more

subtle form of the subadditivity is found when the question is posed with the elements of the partition stated but an assessment made of the union of the elements of the partition. For example, asking for the probability of a homicide by an acquaintance or a stranger vis-à-vis the probability of a homicide. While the events are equivalent, the implicit “unpacking” of the hypothesis “homicide” into the constituents “homicide by an acquaintance” and “homicide by a stranger” may lead to a higher probability for the first questions than the second question. The impact of suggesting additional factors and observing an increase in the given probability is termed implicit subadditivity (Sloman, et al. 2004). These authors find that the evidence for implicit subadditivity is mixed and they demonstrate reversals of the effect when using a partition of atypical conditions.

Eliciting Probabilities

Subjective judgments permeate complex decisions (Bonano, et al. 1989). One has the choice of using these judgments in a loose unstructured manner or insisting on a more formal, more rigorous, and better documented approach. It is particularly important to employ more formal methods when the decision or risk analysis will be subjected to review. In dealing with public risk issues such as policies about the collection and distribution of medical blood supplies, there will be external reviews and perhaps challenges in the courts. Using a more structured, more formal approach will help satisfy reviewers of the study’s accuracy and meaningfulness. In this section, a series of steps are given that should be considered when putting together a study using expert judgments.

Selecting and Posing Questions

Not all uncertainties, however, are of equal consequence in making a decision. There will normally be a few major questions that drive the uncertainty about the optimal choice of actions.

These questions are candidates for a more structured probability assessment activity. Other issues – those that play a minor role – can often be treated less formally or through sensitivity analysis, saving the resources for the more important issues. A sensitivity analysis using initial estimates of probabilities and probability distributions is often performed after the decision has been structured. The sensitivity analysis identifies those questions deserving of a more penetrating study.

But not all issues lend themselves to quantification through expert judgment. In addition to being important contributors to uncertainty and risk, an issue should:

- Be resolvable in that given sufficient time and/or resources, one could conceivably learn whether the event has occurred or the value of the quantity in question
- Have a basis upon which judgments can be made and can be justified

The requirement of resolvability means that the event or quantity is knowable and physically measurable. We consider a counter example. In a study of risk from down wind radioactivity following a power plant failure, a simple Gaussian dispersion model of the form $y = ax^b$ was employed (Harper, et al. 1994). In this model, a and b are simply parameters that give good fit to the relation between x , downwind distance, and y , the horizontal width of the plume. But not all experts subscribe to this model. More complex alternatives have been proposed with different types of parameters. Asking an expert to provide judgments about a and b violates the first principle above. One cannot verify if the judgments are correct, experts may disagree on the definition of a and b , and experts who do not embrace the simple model will find the parameters not meaningful. It is very difficult to provide a value for something you don't believe exists.

The second requirement is that there is some knowledge that can be brought to bear on the event or quantity. For many issues, there is no directly applicable data so that data from analogues, models using social, medical or physical principles, etc., may form the basis for the

judgments. If the basis for judgments is incomplete or sketchy, the experts should reflect this by expressing greater uncertainty in their judgments.

Once issues have been identified, it is necessary to develop a statement that presents the issue to the experts in a manner that will not color the experts' responses. This is called framing the issue. Part of framing is creating an unbiased presentation that is free of preconceived notions, political overtones, and discussions of consequences that might affect the response. Framing also provides a background for the question. Sometimes there are choices about stating conditions for the issues or ignoring the conditions and asking the experts to integrate the uncertainty about the conditions into their responses. In a study of dry deposition of radioactivity, the experts were told that the deposition surface was northern European Grassland but they were not told the length of the grass which is thought to be an important determinant of the rate of deposition (Harper, et al. 1994). Instead, the experts were asked to treat the length of grass as an unknown and to incorporate any uncertainty that they might have into their responses. The experts should be informed about those factors that are considered to be known, those that are constrained in value, those that are uncertain, and, perhaps, those that should be excluded from their answers.

Finally, once an issue has been framed and put in the form of statement to be submitted to the experts, it should be tested. The best way to do this testing is through a dry-run with stand-in experts who have not been participants in the framing process. Although this seems like a lot of extra work, experience has shown (Hora and Jensen, 2002) that getting the issue right is both critical and difficult. All too often, the question that the expert is responding to differs from what was intended by the proposer. It is also possible that the question being asked appears to be resolvable to the person who framed the question but not to the expert who must respond.

Selecting Experts

The identification of experts requires that one develop some criteria by which expertise can be measured. Generally, an expert is one who “has or is alleged to have superior knowledge about data, models and rules in a specific area or field” (Bonano et al., 1990). But measuring against this definition requires one to look at indicators of knowledge rather than knowledge *per se*. The following list contains such indicators:

- Research in the area as identified by publications and grants
- Citations of work
- Degrees, awards or other types of recognition
- Recommendations and nominations from respected bodies and persons
- Positions held
- Membership or appointment to review boards, commissions, etc.

In addition to the above indicators, experts may need to meet some additional requirements. The expert should be free from motivational biases caused by the economic, political, or other interest in the decision. Experts should be willing to participate and they should be accountable for their judgments (Cooke, 1991). This means that they should be willing to have their names associated with their specific responses. Many times physical proximity or availability at certain times will be an important consideration.

How the experts are to be organized also impacts the selection. Often, when more than one expert is used, the experts will be redundant of one another meaning that they will perform the same tasks. In such a case, one should attempt to select experts with differing backgrounds, responsibilities, fields of study, etc., so as to gain a better appreciation of the differences among beliefs. In other instances, the experts will be complementary, each bringing unique expertise to the question. Here, they act more like a team and should be selected to cover the disciplines needed.

Some analyses undergo extreme scrutiny because of the public risks involved. This is certainly the case with radioactive waste disposal. In such instances, the process for selecting (and excluding) experts should be transparent and well documented. In addition to written criteria, it may be necessary to isolate the project staff from the selection process. This can be accomplished by appointing an independent selection committee to seek nominations and make recommendations to the staff (Trauth, Hora, Guzowski 1994).

How many experts should be selected? Experience has shown that the differences among experts can be very important in determining the total uncertainty about a question. Clemen and Winkler (1985) examine the impact of dependence among experts using a normal model and conclude that three to five experts are adequate. Hora (2004) created synthetic groups from the responses of real experts and found that three to six or seven experts are sufficient with little benefit from additional experts beyond that point. When experts are organized in groups and each group provides a single response, then this advice would apply to the number of groups. The optimal number of experts within a group has not been addressed and is certainly dependent on the complexity of issues being answered.

Training

Most experts, even well trained scientists, will not have a great deal of experience with probability elicitation. It is a very good idea to provide some initial training before asking for their judgments. There are different forms that this training can take but usually it will include some or all of the following items:

- Practice with forming probability judgments and feedback on results
- A discussion of personnel/subjective probability and its role in the analysis
- Background information about the elicitation questions
- Discussion of biases in judgment formation

Practice is usually accomplished through the use of a training quiz composed of almanac questions.

Organizing an Expert Judgment Process

In addition to defining issues and selecting and training the expert(s), there are a number of questions that must be addressed concerning the format for a probability elicitation. These include:

- The amount of interaction and exchange of information among experts
- The type and amount of preliminary information to be provided to the experts
- The time and resources that will be allocated to preparation of responses
- Venue – the expert’s place of work, the project’s home, or elsewhere
- Will there be training, what kind, and how will it be accomplished?
- Are the names of the experts to be associated with their judgments and will individual judgments be preserved and made available?

These choices result in the creation of a design for elicitation that has been termed a protocol. Some protocols are discussed in Morgan and Henrion (1990), Merkhofer (1987), and Keeney & von Winterfeldt (1991) and in Cooke (1991). We will briefly outline two different protocols that illustrate the range of options that have been employed in expert elicitation studies.

Morgan and Henrion (1990) identify the Stanford Research Institute (SRI) assessment protocol as, historically, the most influential in shaping structured probability elicitation. This protocol is summarized in Spetzler and von Holstein (1975). It is designed around a single expert (subject) and single analyst engaged in a five-stage process. The stages are:

- Motivating – Rapport with the subject is established and possible motivational biases explored
- Structuring – The structure of the uncertainty is defined
- Conditioning – The subject is conditioned to think fundamentally about his judgment and to avoid cognitive biases
- Encoding – This is the actual quantification in probabilistic terms
- Verifying – The responses obtained in the encoding are checked for consistency

The role of the analyst in the SRI protocol is primarily it to help the expert avoid psychological biases. The encoding of probabilities roughly follows a script. Stael von Holstein and Matheson (1979) provide an example of how an elicitation session might go forward.

A distinguishing feature of the SRI Protocol is the use of the probability wheel described earlier. The encoding stage for continuous variables is described in some detail in Spetzler and von Holstein (1975). It begins with assessment of the extreme values of the values of the variable. An interesting sidelight is that after assessing these values, the subject is asked to describe scenarios that might result in values of the variable outside of the interval and to provide a probability of being outside the interval. The process next goes to a set of intermediate values whose cumulative probabilities are assessed with the help of the probability wheel. Then an interval technique is used to obtain the median and quartiles. Finally, the judgments are verified by testing for coherence and conformance with the expert's beliefs.

While the SRI protocol was designed for solitary experts, a protocol developed by Sandia Laboratories for the U.S. Nuclear Regulatory Commission (Hora and Iman, 1989, Ortiz, et al. 1991) was designed to bring multiple experts together. The Sandia protocol consists of two meetings.

- First meeting
 - Presentation of the issues and background materials
 - Discussion by the experts of the issues and feedback on the questions
 - A training session including feedback on judgments

The first meeting is followed by a period of individual study of approximately one month.

- Second meeting
 - Discussion by the experts of the methods, models, and data sources used
 - Individual elicitation of the experts

The second meeting is followed by documentation of rationales and opportunity for feedback from the experts. The final individual judgments are then combined using simple averaging to the final probabilities or distribution functions.

There are a number of significant differences between the SRI and Sandia protocols. First, the SRI protocol is designed for isolated experts while the Sandia protocol brings multiple experts together and allows them to exchange information and viewpoints. They are not allowed, however, to view or participate in the individual encoding sessions nor comment on one another's judgments. Second, in the SRI protocol, it is assumed that the expert is fully prepared in that no additional study, data acquisition, or investigation is needed. Moreover, the SRI protocol places the analyst in the role of identifying biases and assisting the expert in counteracting these biases while the Sandia protocol employs a structured training session to help deal with these issues. In both protocols, the encoding is essentially the same although the probability wheel is today seldom employed by analysts. Third, the Sandia protocol places emphasis on obtaining and documenting multiple viewpoints which is consistent with the public policy issues addressed in those studies to which it had been applied.

References

Alpert, M. & Raiffa, H. (1982) A progress report on the training of probability assessors in judgment under uncertainty: Heuristics and biases, in D. Kahneman, P. Slovic, and A. Tversky, (Eds.) *Judgment under uncertainty: Heuristics and biases*, Cambridge: Cambridge University Press.

Bar-Hillel, M. & Neter, E. (1993), How alike is it versus how likely is it: A disjunction fallacy in stereotype judgments, *Journal of Personality and Social Psychology* 65, 1119-1132.

Bar-Hillel, M. (2001). *Subjective probability judgments*, in Smelser, N.J. & Baltes, D.B. (Eds.) *International Encyclopedia of the Social & Behavioral Sciences*, Amsterdam: Elsevier Science Ltd. 15248-15251.

Bonano, E.J., Hora S.C., Keeney, R.L., and von Winterfeldt, D.. (1989). *Elicitation and Use of Expert Judgment in Performance Assessment for High-Level Radioactive Waste Repositories*, NUREG/CR-5411. Washington: U.S. Nuclear Regulatory Commission.

Brenner, L.A., Koehler, D.J., Liberman, V. (1996) Overconfidence in probability and frequency judgments: A critical examination, *Organizational Behavior and Human Decision Processes*, 65, 212–219.

Christensen-Szalanski, J. & Bushyhead, J. (1981). “Physicians; Use of Probabilistic Information in a Real Clinical Setting,” *Journal of Experimental Psychology: Human Perception and Performance*, 7, 928-935.

Clemen, R.T., Fischer, G.W., & Winkler, R.L. (2000) Assessing dependence: Some experimental results, *Management Science*, 46, 1100-1115.

Clemen, R.T. & Winkler, R.L. (1985). limits for the precision and value of information from dependent sources, *Operations Research*, 33, 427-442.

Clemen, R.T. & Reilly, T. (1999) Correlations and copulas for decision and risk analysis, *Management Science*, 45, 208-224.

Cooke, R.M. (1991). *Experts in Uncertainty*, Oxford: Oxford University Press.

Cover, T.M. & J.A. Thomas (1991). *Elements of information theory*. New York: Wiley-Interscience.

de Finetti, B. (1974), *Theory of Probability*, Vol. 1. New York: John Wiley and Sons.

de Finetti, B. (1977). Probabilities of probabilities: a real problem or a misunderstanding? in A. Aykac and C. Brumat, (Eds.) *New Developments in the Application of Bayesian Methods*, Amsterdam: North Holland Publishing Company.

Druzdel, M.J. (1989). Verbal Uncertainty Expressions: Literature Review, Technical Report CMU-EPP-1990-02-02,. Pittsburgh: Department of Engineering, Carnegie Mellon University.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978), Fault trees: Sensibility of estimated failure probabilities to problem representation, *Journal of Experimental Psychology: Human Perception and Performance*, 4, 330–344.

Fischhoff, B. & MacGregor, D. (1982) Subjective Confidence in Forecasts, *Journal of Forecasting*, 1, 155-72.

Fox, C.R. & Clemen, R.T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, in press.

- Fox, C.R. & Tversky, A. (1998). A belief-based account of decision under uncertainty, *Management Science*, 44, 879-95
- Gigerenzer, G. (1991). How to make cognitive illusions to disappear: Beyond heuristics and biases. In W. Stroebe & M. Hewstone (Ed.), *European Review of Social Psychology*, vol. 2. Wiley.
- Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506-528.
- Haim, E. (1982). *Characterization and Construction of Proper Scoring Rules*. Unpublished doctoral dissertation, University of California, Berkeley.
- Harper, F.T., Hora, S.C., Young, M.L. Miller, L.A. Lui, C.H. McKay, M.D. Helton J.C., Goossens, L.H.J. Cooke, R.M. Pasler-Sauer, J. Kraan, B. & Jones, J.A. (1994). *Probability Accident Consequence Uncertainty Analysis, Vols. 1-3, (NUREG/ CR-6244, EUR 15855 EN)* Brussels: USNRC and CEC DG XII.
- Helton, J.C. & Burmaster, D.E.,Eds.(1996) Special issue on treatment of aleatory and epistemic uncertainty, *Reliability Engineering & System Safety*, 54, Nos. 2-3.
- Hora, S, Dodd, N.G., & Hora, J. (1993). The use of Decomposition in Probability Assessments of Continuous Variables, *The Journal of Behavioral Decision Making*, 6, 133-147.
- Hora, S.C. & Iman, R.L. (1989). Expert Opinion in Risk Analysis: The NUREG-1150 Experience, *Nuclear Science and Engineering*, 102, 323-331
- Hora, S.C. & Jensen, M. (2002). *Expert Judgement Elicitation*, Stockholm:. Swedish Radiation Protection Authority
- Hora, S.C. (2004). Probability Judgments for Continuous Quantities: Linear Combinations and Calibration, *Management Science*, 50, 597-604.
- Hora, S.C., Hora, J.A., & Dodd, N.G. (1992). Assessment of Probability Distributions for Continuous Random Variables: A Comparison of the Bisection and Fixed Value Methods, *Organizational Behavior and Human Decision Processes*, 51, 133-155.
- Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, 57, 226-246.
- Kahneman, D., Slovic, P., & Tversky, A. (1982), (Eds.) *Judgment under uncertainty: Heuristics and biases*, Cambridge: Cambridge University Press.
- Keeney, R., & von Winterfeldt, D. (1991). Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management*, 38, 191-201.

Kent, S. (1964). Words of Estimated Probability, in D. P. Steury, ed., *Sherman Kent and the Board of National Estimates: Collected Essays* Washington: Center for the Study of Intelligence, 1994.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertainty knowledge: some calibration studies. *Acta Psychologica*, 67, 95-119.

Klayman, J., Soll, J.B., Gonzalez-Vallejo, C., Barlas, S (1999). Overconfidence: It depends on how, whom you ask. *Organizational Behavior and Human Decision Processes*, 79, 216-2247.

Knight, F. H. (1921). *Risk, Uncertainty, and Profit* Boston: Houghton Mifflin Company.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159–183.

Lichtenstein, S., Fischhoff B., & Phillips, L.D.. (1982). Calibration of probabilities: The state of the art to 1980, in D. Kahneman, P. Slovic, and A. Tversky (1982), (Eds.) *Judgment under uncertainty: Heuristics and biases*, Cambridge: Cambridge University Press.

Lichtenstein, S., Fischhoff, B., & Phillips, L.D. (1982). Calibration of probabilities: The state of the art to 1980, in D. Kahneman, P. Slovic, and A. Tversky, (Eds.) *Judgment under uncertainty: Heuristics and biases*, Cambridge: Cambridge University Press.

Lichtenstein, S., Fischhoff. B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159-183.

Matheson, J.E. & Winkler, R.L. (1976). *Scoring Rules for Continuous Probability Distributions*, *Management Science*, 22, 1087-1096

Merkhofer, M. W. (1987). Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics*, 17, 741-752.

Morgan, M.G., Henrion, M (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*, New York: Cambridge University Press.

Murphy, A. H. (1973) A New Vector Partition of the Probability Score, *Journal of Applied Meteorology*, 12 595-6

Murphy, A.H. & Winkler, R.L. (1977). Reliability of Subjective Probability Forecasts of Precipitation and Temperature, *Applied Statistics*, 26, 41-47.

Murphy, A.H. (1972). Scalar and Vector Partitions of the Probability Score Part I) Two-state situation, *Journal of Applied Meteorology*, 11, 273-82.

Ortiz, N.R., Wheeler, T.A., Breeding, R.J., Hora, S., Meyer, M.A., & Keeney, R.L. (1991). The Use of Expert Judgment in the NUREG-1150, *Nuclear Engineering and Design*, 126, 313-331.

Raiffa, H. & Schlaifer, R. (1964), *Applied Statistical Decision Theory*, Cambridge: MIT Press.

Rasmussen, N. et al. (1975). *Reactor safety study: WASH-1400, NUREG-751014*, Washington: U.S. Nuclear Regulatory Commission.

Ravinder, H. V., Kleinmuntz, D. N., & Dyer, J. S. (1988). The reliability of subjective probability assessments obtained through decomposition. *Management Science*, 34, 186–199.

Savage, L.J. (1954), *The Foundations of Statistics*. New York: John Wiley and Sons (second edition, 1972, New York: Dover).

Savage, L.J. (1971). The elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66, 783—801.

Skyrms, B. (1980). "Higher order degrees of belief." in "Prospects for Pragmatism: Essays in Honor of F. P. Ramsey," D. H. Mellor (Ed.), Cambridge University Press, Cambridge, pp. 109-137.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C.R. (2004). Typical versus atypical unpacking and superadditive judgment, *Journal of Experimental Psychology*, 30, 573-582.

Spetzler, C.S. & Stael von Holstein, C-A. S. (1975). Probability Encoding in Decision Analysis, *Management Science*, 22, pp. 340-358.

Stael von Holstein, C.-A.S. & Matheson, J.E. (1979), *A manual for encoding probability distributions*, Menlo Park, California: SRI International.

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., Hora S.C. & Guzowski, R.V., (1994), *A formal expert judgment procedure for performance assessments of the waste isolation pilot plant, SAND93-2450*, Albuquerque: Sandia National Laboratories.

Tversky, A. & Kahneman D. (1973). Availability a Heuristic for Judging Probability, *Cognitive Psychology* 5, 207-232. and in an abbreviated form in D. Kahneman, P. Slovic, and A. Tversky, (Eds.) *Judgment under uncertainty: Heuristics and biases*, Cambridge: Cambridge University Press.

Tversky, A. & Kahneman D. (1974). Judgment under uncertainty: Heuristics and biases, *Science*, pp1124-1131 and in D. Kahneman, P. Slovic, and A. Tversky, (Eds.) *Judgment under uncertainty: Heuristics and biases*, Cambridge: Cambridge University Press.

Tversky, A. & Koehler, D.J. (1994) Support theory: A nonextensional representation of subjective probability, *Psychological Review*, 101, 547-567.

Winkler, R.L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, 66, 675-685.

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: How to cure overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30, 1167-1175.