Cognitive Factors Affecting Subjective Probability Assessment

Alyson G. Wilson

Institute of Statistics & Decision Sciences, Duke University

ISDS Discussion Paper #94–02

February 1, 1994

Cognitive Factors Affecting Subjective Probability Assessment

Alyson G. Wilson

Institute of Statistics and Decision Sciences

Duke University

Durham, NC 27708–0251, USA

Summary

Prior probabilities are central to Bayesian statistics. The Bayesian paradigm assumes that people can express uncertainty in terms of subjective probability distributions. This article will consider Hogarth's 1975 assessment that "man is a selective, sequential information processing system with limited capacity, . . . ill-suited for assessing probability distributions." Particular attention will be paid to when people make normatively "good" or "poor" probability assessments, what techniques are effective in eliciting "good," coherent probability assessments, and on how these ideas are relevant to the practicing Bayesian statistician. While there are situations where experts can make well-calibrated judgments, it will be argued that more research needs to be done into the effects of expertise, training, and feedback.

1 Introduction

Prior probabilities are central to Bayesian statistics. The Bayesian paradigm provides a systematic way to update prior beliefs to posterior beliefs based upon observed data. An important, though often unstated, assumption of this methodology is that people can express their prior beliefs in the form of probability assessments. The work in the statistics community has focused on formal theories of subjective probabilities that are primarily concerned with the conditions imposed on individual judgments to permit the construction of a proper, or normatively correct, probability measure. For the most part, these theories ignore the questions of how people actually assess the probability of an uncertain event (Tversky, 1974). Indeed, Winkler (1967, p. 777) states, "Despite the importance of prior distributions in Bayesian analysis, little previous work has been done on the practical problems of the assessment of non-diffuse distributions."

Most statisticians hold views similar to those discussed by Lindley et al. (1979) and assume that "the subject has, in some sense, a set of coherent probabilities that are distorted in the elicitation process" (p. 149). This belief has led to a focus in the statistical literature on how to make coherent any incoherent elicited probabilities. Other statisticians hold a view more influenced by the psychological literature on subjective probability assessment. They believe that there is no built-in prior distribution there for the taking. Elicitation techniques help to draw out an assessment of a prior distribution based on prior knowledge, but different techniques may produce different results because the method of questioning may affect how the problem is viewed (Winkler, 1967). Winkler (1967) states:

> Although there is no "true" prior distribution, there is a "satisficing" prior distribution—one which the assessor is content to live with at a particular moment of time. But this has no objective existence outside the assessor's head, and its choice may in fact be influenced by such things as ease in calculation. (p. 778)

It is interesting to note that the use of "computationally convenient" priors is widespread in statistics. Only rarely do Bayesian statisticians try to elicit and incorporate

meaningful prior distributions into analyses. In 1975, Hogarth published a review of the psychological literature pertaining to subjective probability assessment in the *Journal of the American Statistical Association.* The primary conclusion of his paper is that "since man is a selective, sequential information processing system with limited capacity, he is ill-suited for assessing probability distributions" (p. 271). Bayesian statisticians are left with few options if they accept these conclusions. They can decide that the Bayesian framework, while philosophically attractive, is practically untenable; they can shift the focus away from probability assessment toward robustness and sensitivity issues; or they can, as Hogarth suggests, focus on "assessment techniques ...designed both to be compatible with man's abilities and to counteract his deficiencies" (p. 279). Statisticians have chosen all three options.

This paper focuses on the last option, summarizing Hogarth's paper and examining the approaches to subjective probability assessment that researchers have explored in subsequent years. Emphasis will be placed on when people make normatively "poor" or "good" probability assessments, what techniques are effective in eliciting "good," coherent probability assessments, and on how these ideas are relevant to the practicing Bayesian statistician.

2 Criteria For "Good" Assessments

There are many criteria available for evaluating the "goodness" of probability assessments. It is important to remember, as Hogarth points out, that subjective probability assessments cannot be "wrong," provided they are coherent and that they capture all of the information available to the assessor. However, "probability assessments might seem to be more or less accurate in the light of subsequent events" (p. 271). Hogarth discusses three ways that assessments are judged: by comparison with the "objective" probabilities of an experimenter in a laboratory setting, by evaluation of a penalty or scoring function that depends on the elicited probabilities and the events that actually occur, and by evaluation over a series of trials the degree to which the empirical relative frequencies of predicted events correspond the assessed probabilities (i.e., are *calibrated*). Although each of these methods is widely used, they all have theoretical

shortcomings. In a laboratory setting, all an experimenter can do is compare his subjective probabilities with those of his subject, but assuming a personalistic definition of probability, the experimenter has no basis for calling his probabilities "objective" and those of his subject "subjective." There are also difficulties with evaluating assessments in terms of penalty functions or calibration. Penalty functions and empirical relative frequencies are useful only when evaluated over several trials, but how many trials to use is an inherently subjective specification.

Despite its shortcomings, calibration is the most widely used criterion for the evaluation of probability assessments. However, several other criteria have been used in the recent literature. Winkler and Murphy (1968) propose two criteria: *normative goodness*, which is the degree to which probability assessments reflect the assessor's beliefs (i.e., are *valid*) and conform to the axioms of probability theory (i.e., are *coherent*), and *substantive goodness*, which is the amount of knowledge the assessor has about the domain of the probability assessment. Winkler (1986) discusses the criterion *expertise*, which "relates to the degree to which the probability appraiser can approach perfect forecasts" (p. 267), or the degree to which a probability forecaster can correctly use probabilities close to zero and one. Sharp et al. (1988) identify *resolution* as "the judge's ability to discriminate correct from incorrect judgments by differentially assigning confidence judgments to accurate and inaccurate judgments" (p. 272), while Winkler and Poses (1991) define *discrimination* as the ability to give different probabilities for different conditions (e.g., to assign different survival probabilities to patients who ultimately live or die). *Reliability* measures whether the assessment is repeatable, stable, and consistent across elicitations (Wallsten and Budescu, 1983).

A person wishing to assess the goodness of a probability judgment has a multitude of options—many more than had been proposed when Hogarth wrote. These new options capture more fully the many features of a good probability assessment.

3 Implications From The Psychological Literature

Hogarth proposes that the literature on judgmental processes has produced at least two firm conclusions: that man has only limited information processing capacity, and that

the nature of the task has great impact on the strategies that are chosen to deal with the task. He derives three major consequences of man's limited information processing capability:

- Man's perception of information is not comprehensive but selective. Since man is only capable of apprehending a small part of his environment, his anticipations of what he will perceive determine to a large extent what he does perceive.
- As he does not have the capacity to make what one might call "optimal" calculations, man makes much use of heuristics and cognitive simplification mechanisms.
- Since he cannot simultaneously integrate a great deal of information man is forced to process information in a sequential fashion. (p. 273)

These consequences have been widely explored in the psychological literature.

3.1 Limited Information Processing Capacity (Pre-1975)

3.1.1 Denial of Uncertainty

One cognitive simplification mechanism that Hogarth identifies is the psychological reduction or denial of uncertainty. People seem to have a great deal of trouble accepting that uncertainty lies within themselves rather than being an intrinsic property of events in the environment. "The world is perceived by us as being probabilistic since we are unable to see and comprehend the myriad factors that cause events to occur" (Hogarth, 1987, p. 12). For example, it has been found that business managers try to avoid uncertainty by seeing it as subject to control and by believing that appropriate and skillful action can reduce risk (March and Shapira, 1987). Savage (1971) suggests that Western culture does not train people to think appropriately about uncertainty, stating that:

> The usual tests and the language habits of our culture tend to promote confusion between certainty and belief. They encourage both the vice of acting and speaking as though we were certain when we are only fairly sure and that of acting and speaking as though the opinions we do have were worthless when they are not very strong. (p. 800)

Since people are so reluctant to attribute uncertainty to themselves, one must question how effectively untrained people (and non-expert experimental subjects) can quantify subjective probabilities about uncertain events.

3.1.2 Representativeness Heuristic

Hogarth also addresses the work of Tversky and Kahneman (1974), who identify three cognitive simplification mechanisms: the representativeness, availability, and adjustment and anchoring heuristics. The representativeness heuristic is often applied to answer questions like, "What is the probability that event A originates from process B?" The more similar A is to B, the more likely it is to be judged to originate from B. For example, consider "Joe," who is described as active, aggressive, and enthusiastic. If asked to choose, most people would indicate that Joe's hobby is playing football, not writing poetry. When using the representativeness heuristic, the probability that Joe plays football is assessed by judging his similarity to the stereotype of football players.

Several judgment errors result from the application of this heuristic. The first is a tendency to ignore base-rate information in the presence of case-specific information. This clearly has a direct impact on the assessment of subjective probability distributions, as the assessor may take into account only information unique to the situation at hand and ignore previous data.

The second judgment error that results from application of the representativeness heuristic is an insensitivity to sample size. People apparently do not have an intuitive Law of Large Numbers, as "intuitive judgments are dominated by the sample proportion and are essentially unaffected by the size of the sample, which plays a crucial role in the determination of the actual posterior" (Tversky and Kahneman, 1974, p. 1125). A third resultant error is the misperception of chance and randomness. Many people believe in "local" randomness, or "that the essential characteristics of the process will be represented, not only globally in the entire sequence, but also locally in each of its parts" (Tversky and Kahneman, 1974, p. 1125). A common example of this error is the "gambler's fallacy." When flipping a coin, if several heads appear, people expect that the next flip will show tails. It is interesting to note that misperceptions of chance

are not confined to non-expert subjects. Tversky (1974) notices a belief, even among experienced researchers in the "law of small numbers." Investigators believe that a strong effect will appear, and be statistically significant, in a sample of ten subjects, just as it would in a sample of one thousand subjects. Consequently, researchers put too much faith in the results of small samples and often overestimate the replicability of small-sample results.

3.1.3 Availability and Adjustment/Anchoring Heuristics

A second cognitive simplification mechanism identified by Tversky and Kahneman (1974) is availability. When using this heuristic, people judge the frequency of occurrence of a class by the relative ease of recalling instances from the class. For example, one might judge the probability of having a car accident by recalling instances when friends have had wrecks. If your spouse has recently had an accident, your subjective probability of having a car wreck will temporarily rise due to the salience of the event. A common error associated with this heuristic is illusory correlation, where the co-occurrence of two events is judged on their strength of association. This impacts directly on subjective probability assessment, as assessors are often asked to estimate joint or conditional probabilities that depend on the correlations between events.

A third cognitive simplification mechanism is adjustment and anchoring. When people are asked to make predictions, they often select a salient (not necessarily relevant) starting point and adjust their guesses from there. In many situations it has been found that the adjustment is insufficient. This has a direct impact on subjective probability elicitation: several investigators (Winkler, 1967; Alpert and Raiffa, 1982) have found that when asked to choose a range which should contain 98 percent of the observed data, subjects tend to anchor the judgments on the median and insufficiently adjust outward, resulting in intervals which contain only about 70 percent of the data.

3.2 Limited Information Processing Capacity (Post-1975)

3.2.1 Conjunction Fallacy

Since Hogarth's 1975 paper, researchers have continued to probe the limits of man's information processing capacity. They have discovered many other errors common to probability judgments. Tversky and Kahneman (1983) discuss examples of the "conjunction fallacy." From elementary probability theory, it must be the case that the probability of a conjunction, P(A & B), cannot exceed the probability of either of its constituents, P(A) or P(B). However, it is often the case that the conjunction is either more representative of its class than either of its constituents, or is more available, and hence judgments of its probability are made using the representativeness or availability heuristic. For example, although the class of seven letter words ending in "ing" is contained in the set of seven letter words with an "n" in the sixth place, the former class is judged to be more common because it is easier to generate words ending in "ing" than words with an "n" as the sixth letter. This particular error calls into question whether subjective probability judgments are inherently coherent, as any judgment satisfying conjunction fallacy is inconsistent with the fundamental laws of probability.

3.2.2 Hindsight Bias

"The hindsight bias is the tendency of people with outcome knowledge to believe falsely that they would have predicted the reported outcome of an event" (Hawkins and Hastie, 1990, p. 311). This bias affects probability elicitation, because once outcomes are observed, the assessor may assume that they are the only outcomes that could have happened, and thus underestimate the uncertainty inherent in the outcomes that could have happened, but didn't. Edwards and von Winterfeldt (1986) suggest that proper questioning can mitigate the effects of hindsight bias:

> A formulation like "Twenty-four hours before the battle started, how probable was it that Napoleon would lose at Waterloo?" is at worst nonsense and at best incomplete, because it specifies neither whose the probability is nor the information on which it is based. The question "Twenty-four hours before

> the battle started, how probable did Napoleon consider it to be that he would lose at Waterloo?" is at least well framed. (pp. 656–657)

From a Bayesian perspective, the hindsight bias is difficult to correct, because the purpose of Bayesian statistics is to update prior beliefs based on observed data, and the effects of the hindsight bias are difficult to separate from actual expert learning.

3.2.3 Assessing Variance, Covariance, and Correlation

Jennings et al. (1982) conduct experiments where they look at both data-based correlation estimates, where the data consists of pairs of numbers or sounds, and at theory-based estimates, where no data is presented by the experimenter. In their data-based experiment, they find that subjects have a great deal of difficulty recognizing positive relationships with correlations of less than 0.6–0.7. They find that correlations in the range of 0.2–0.4 are barely detectable and that correlations of 0.6–0.8 are substantially underestimated. Only correlations over 0.85 are consistently rated as strongly positive. However, the subjects' theory-based estimates, while variable, do tend to estimate positive empirical correlations as positive, negative as negative, and to correctly capture the relative magnitude of the correlations. The "most striking" feature of the theory-based experiments is that subjects lose their "conservatism" when "freed from the constraints of immediately available data" (p. 223). Indeed, subjects are "apt to expect and predict covariations of considerable magnitude—often of far greater magnitude than are likely to have been presented by past experience or to be borne out by future experience" (p. 224).

Other studies have found that subjects also have considerable difficulty estimating statistical variance. Estimates tend to be influenced by the mean value of the stimuli: instead of estimating variance, subjects tend to estimate the coefficient of variation (the standard deviation divided by the mean). Peterson and Beach (1967) give the following explanation:

> Think of the top of a forest. The tree tops seem to form a fairly smooth surface, considering that the trees may be 60 or 70 feet tall. Now, look at

> your desk top. In all probability it is littered with many objects and if a cloth were thrown over it the surface would seem very bumpy and variable. The forest top is far more variable than the surface of your desk, but not relative to the sizes of the objects being considered. (p. 31)

Just as with the illusory correlation caused by the availability heuristic, misperceptions of variance and correlation can cause difficulties in joint and conditional probability assessment.

3.2.4 Conservatism

Another robust observation found in the literature is that man tends to be a "conservative" processor of information. "Book bag experiments" demonstrate this effect. Subjects are presented with two bags, one that contains 70 red and 30 blue poker chips, and one that contains 30 red and 70 blue. A bag is chosen at random, and a single chip is drawn. Based upon the color of this chip, the subject is asked to report his probability that the chip came from the bag containing 70 red chips. Typically, subjects revise their probabilities less than Bayes rule suggests that they should. It would be convenient if one could play off the over-shrinkage of the hindsight bias with the under-shrinkage of conservatism, but methods for achieving updating consistent with Bayes rule are difficult to find and highly dependent on task characteristics.

3.2.5 Discussion

While these errors paint a dismal picture of man's ability to make probabilistic judgments, there is some evidence that people perform rather well at certain probabilistic tasks. Peterson and Beach (1967) report that subjects can accurately describe proportions (although they do have difficulties with values near zero and one) and estimate measures of central tendency (e.g., means, medians, and modes). There are researchers, however, who criticize the entire research paradigm. Edwards and von Winterfeldt (1986, p. 670) ask, "What is the difference between the experimenter and the subject, other than that the former gets the answers right and the latter gets them wrong?" They

argue that the primary difference is that subjects are prevented from using appropriate physical and intellectual tools, saying:

> The topic of intellectual tools relates to expertise. Experts become expert in the use of intellectual tools as well as acquiring factual knowledge. They may use physical tools to implement the intellectual ones; experts on Bayesian statistics, though they have no difficulty recognizing a Bayesian problem, may need a hand calculator or even a computer to get the right answer. (p. 670)

They suggest that many experimental probability judgment tasks are equivalent to asking the subject to pound a nail into a board without a hammer. The subject is not likely to be very successful, as he does not have the tool he needs. Their criticisms make it clear that the pragmatic Bayesian must place sufficient resources—references, computers, task-specific information, etc.—at the disposal of his expert assessor.

3.3 Task Characteristics

As Hogarth points out in his 1975 article, there is also considerable evidence that task effects can cause different evaluations of subjective probability distributions. The literature since then supports this assertion. Order effects are discussed in great detail in Hogarth and Einhorn (1992), who point out that response mode, task complexity, and length of series can influence the primacy and recency effects of information presentation. It has also been shown that response mode influences elicited probabilities—people give different values if asked to assess fractiles, odds, or how to bisect an interval (Winkler, 1967; Hora et al., 1992). Johnson et al. (1991) discuss list-length effects: "Previous research indicates that the probabilities that people attach to various events can be influenced substantially by the extent to which all the possible events are explicitly listed for consideration" (p. 325). If a list contains N alternatives, non-expert subjects anchor on probabilities of $\frac{1}{N}$, and then adjust insufficiently. Subjective probabilities can also be affected by the measurement scale; for example, engineers may be more comfortable thinking on a logarithmic scale, and may therefore assess probabilities more easily in those terms (Hora et al., 1992; Johnson et al., 1991).

The effect of financial incentives has been mixed. Wright and Anderson (1989) find that performance-contingent incentives can overcome anchoring and adjustment biases in situations where "the task is significantly, but not excessively, demanding on a person's cognitive resources, and the person is motivated to allocate sufficient cognitive effort" (p. 79). Arkes et al. (1986), on the other hand, find that payoffs may motivate subjects to abandon decision rules that they are specifically told will improve their performance. Incentives are believed to work by encouraging people to work harder and devote more cognitive resources to a task. If, however, increased attention does not lead to more appropriate strategies, the subject may simply be encouraged to devote more resources to a flawed strategy, thus leading to poorer performance when incentives are present.

3.4 Calibration Studies

Many recent studies of probability elicitation have focused on calibration as a measure of goodness. For non-expert subjects, these studies can be divided into those assessing the probabilities of events and those assessing the probability densities of unknown quantities. Reviews of these studies can be found in Lichtenstein et al. (1982), Wallsten and Budescu (1983), Keren (1991), and Winkler et al. (1992).

3.4.1 Non-expert Assessors

When eliciting the probabilities of events, investigators often use general knowledge or almanac questions. These questions usually present subjects with two alternatives and ask them to assess which one is more likely and what its probability is. For example, the subject might be asked, "Which president appears on the front of a two dollar bill, Jefferson or Lincoln?" The subject might choose Jefferson with 80 percent confidence in being correct. These studies have found that subjects, whether naive or expert, tend to be consistently overconfident—the proportions correct are less than the assessed probabilities. Fischhoff et al. (1977) experiment with various question formats and response modes, but find overconfidence to be quite robust. They report that only 72 to 83 percent of items assigned probabilities of 1.0 were actually true. Koriat et al.

(1980) find that asking subjects to write down all of the arguments that support or contradict their choices significantly improves calibration. The specific technique that improves calibration most is listing negative evidence—they speculate that subjects may fail to give proper weight to negative indicators when assessing probabilities.

Overconfidence is also commonly found in the assessment of probability distributions. Alpert and Raiffa (1982) report results from five groups of students given almanac questions. All groups gave 25th, 50th, and 75th percentiles. In addition, Group 1 gave 1st and 99th percentiles; Group 2 gave 0.1th and 99.9th percentiles; Group 3 gave "minimum" and "maximum" values; and Group 4 gave "astonishingly low" and "astonishingly high" values. Group 5 received feedback after a first set of assessments. In every case, the spread of the tails of the distributions was too small, regardless of the definition of the extremes, and although feedback did improve the spread, it did not completely eliminate the overconfidence bias. Likewise, O'Connor and Lawrence (1989) find that feedback improves the calibration of confidence intervals. Their task involves time series predictions, and they find that calibration is influenced by the degree of forecasting difficulty. For simple series, subjects were underconfident, while for medium to high difficulty series, the subjects were overconfident.

3.4.2 Expert Assessors

Among experts, subjective probability assessment has been most studied in weather forecasters. Since 1965, weather forecasters in the United States have made their daily forecasts in terms of subjective probabilities (e.g., "I believe there is a 70 percent chance of rain tomorrow"). Murphy and Winkler (1977) present a calibration diagram of 154,799 probability of precipitation forecasts that shows that weather forecasters are almost perfectly calibrated. Their success can be attributed to several advantages they have when making their assessments: their task is repetitious, there is excellent supporting information, feedback is provided, and rewards are given for good performance.

Another group of experts that has been widely studied is physicians. Christensen-Szalanski and Bushyhead (1981) study nine physicians who examine 1531 patients with coughs. Each patient was assigned a probability of having pneumonia by one of the

physicians. The researchers find that physicians' calibration was quite poor: for the highest level of probability (0.88), the actual proportion of patients who had pneumonia was 0.20. Admittedly, this poor calibration may be due to the fact that misdiagnosing a healthy patient as having pneumonia is much less costly than misdiagnosing a patient with pneumonia as being healthy. In contrast, Winkler and Poses (1991), evaluate four groups of physicians with varying degrees of experience and expertise on their estimations of the survival probability of patients in intensive care. While all four groups were well-calibrated, the group with the most experience and expertise performed best. This study also found that more experienced physicians had better discrimination and resolution—they were better able to group together patients with similar survival chances.

Wallsten and Budescu (1983) note that experts estimating probabilities in areas with which they are familiar can be quite well calibrated. However, if the elicitation moves outside the area of expertise, experts fall prey to the same mistakes made by non-expert subjects. As Winkler et al. (1992) say:

> 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. (pp. 4–14)

The evidence shows that whether expert or naive, many factors affect the calibration and goodness of probability assessments.

4 Meaningfulness: When Does It Make Sense To Assess Probabilities?

Since there is a great deal of evidence to suggest that people make poor probability assessments, Hogarth (1975) suggests three criteria for evaluating when it is meaningful to ask people to make these judgments.

- The task should be meaningful to the assessor in that it concerns a domain with which he is reasonably familiar.
- The probability assessment should add something to the predictive accuracy over and above that which could be achieved by the best available statistical model.
- The judgments expressed in probabilistic form should be both more accurate (as evaluated by subsequent events) and useful than those expressed normally (deterministically).

The first criterion suggests that, while almanac questions asked of non-expert subjects may be enlightening as to the difficulties people face in making probability assessments, they should not be considered, evaluated, or used as meaningful judgments. The second criterion implies that the assessments of substantive experts, for example in meteorology, medicine, and waste management (Murphy and Winkler, 1984; Winkler and Poses, 1991; Winkler et al., 1992) are useful, but that care must be taken to insure that the events to be predicted do not exceed the limits of their expertise. As mentioned earlier, experts may have difficulty producing "good" forecasts outside their area of immediate expertise (Keren, 1991). However, since many probability assessments are made for one-time predictions, where statistical modeling is impossible, this criterion may be difficult to apply. Furthermore, this criterion is almost meaningless within the Bayesian paradigm, where probability assessment is integral to the statistical model. Bayesian models require the assessment of prior belief, even the prior belief of having "no information."

In support of the third criterion, Hogarth cites Martin and Gettys (1969):

> Originally it was thought that the use of the probability response mode would cause Ss to exhibit a more exacting type of inference and thereby improve their inferred nominal response performance. The results indicate that the opposite was true .... A possible explanation ...would be that probability responses require different information processing behavior than nominal responses require. When nominal responses are made, for example, perhaps only the few hypotheses judged to be most likely need to be considered since

> S's task is simply to choose the most likely hypothesis. If S's response is a probability, he should be concerned with the likelihood of all the hypotheses. (Hogarth, p. 278)

Subjects may need to be convinced of the need for subjective probability assessment. This points out the need for training the probability assessor, a subject that will be dealt with in more depth later.

5 Elicitation

Hogarth's primary conclusion about the actual elicitation process in his 1975 article is:

> Given man's limited information processing capacity, my own inclination is to favor assessment procedures which decompose the task into small, manageable units .... However, the issue which still has to be addressed is the definition of "manageable units." This may very well vary as a function of the judgmental task and the experience of the assessor. (pp. 279–280)

This issue has been addressed in recent years by both decision analysts and statisticians.

5.1 Three Phases Of Probability Encoding

5.1.1 Deterministic Phase

Spetzler and Staël von Holstein (1975) identify three phases of probability encoding: the *deterministic* phase, where relevant variables are identified and values are assigned to possible outcomes, the *probabilistic* phase, where the subjective probability assessment is made, and the *informational* phase, where the economic value of further reducing uncertainty is considered. During the deterministic phase, Merckhofer (1987) suggests that the analyst ask the assessor to state all relevant knowledge relating to the uncertain variable (recall the work of Koriat et al. 1980). Often this knowledge has a problem specific component and a "distributional" or base-rate component. He suggests that subjects be specifically reminded to use base-rate information to combat the effects of the representativeness bias. Kahneman and Tversky (1982) even propose a specific

corrective procedure to incorporate distributional information into an assessment. The deterministic phase is the time to make sure that the assessor has all the tools necessary to conduct the assessment.

5.1.2 Probabilistic Phase

Spetzler and Staël von Holstein (1975) make these suggestions about the probabilistic phase of encoding:

- Choose only uncertain quantities that are important to the decision. (Elicitation is a time-consuming and difficult process.)
- Be sure that the decision maker does not feel that the outcome of the quantity can be affected by his decision.
- If the decision maker feels that the values of the quantity are conditional on some other variable, explicitly incorporate the conditionality into the problem.
- Clearly define the variable. It should be able to pass the "clairvoyant test"—a clairvoyant should be able to reveal the value of the quantity by specifying a single number without requesting clarification.
- Describe the quantity using a scale that is meaningful to the subject.

Winkler et al. (1992) further suggest that questions be asked only about observable, or at least theoretically observable, quantities.

In addition, the analyst must be aware of motivational biases. "For example, a sales manager may consciously give a low prediction of sales because he thinks he will look better if the actual sales exceed his forecast" (Spetzler and Staël von Holstein, 1975, p. 345). Merckhofer (1987) identifies two types of motivational biases: "management" bias and "expert" bias. Management bias occurs when the subject views an uncertain variable as a goal rather than as an uncertainty. His assessment reflects what he thinks ought to happen instead of his actual uncertainty: "Well if that's the variable that the boss wants minimized, we'll minimize it." Expert bias occurs when a subject learns that

he has been designated an expert and decides that experts are expected to be certain, thus severely underestimating his actual uncertainty.

With these suggestions and potential biases in mind, Spetzler and Staël von Holstein (1975) suggest three encoding methods to use during the probabilistic phase. *P-methods* ask the assessor to give probabilities to fixed values; *V-methods* ask the assessor to give the values corresponding to fixed probabilities; and *PV-methods* ask questions that must be answered on both scales jointly. A second level of complexity is added when the two common response modes are considered. In direct response mode, the assessor is asked questions that require numbers (such as values, probabilities, or odds) as answers. In indirect response mode, the assessor chooses between two or more alternatives, typically bets. Indirect responses can be further characterized as to whether they depend on an external reference process (a familiar reference event), or an internal reference process, where the assessor chooses between events defined on the value scale for the uncertain quantity (e.g., the event of attendance being less than or equal to 10,000 people or that of attendance being greater than 10,000 people). Thus, probability encoding techniques can be classified according to the encoding method and the response mode used.

P- and V-Method, Indirect Response, External Reference. A P-method encoding technique commonly used by decision analysts is the probability wheel (Spetzler and Staël von Holstein, 1975; Merckhofer, 1987). A probability wheel has two sections of different colors and adjustable sizes, and a fixed pointer in the center. The assessor is asked questions like, "Which event do you consider more likely, that next year's sales will exceed 2500 units, or that the pointer will land in the red section of the wheel?" The size of the red section is adjusted until the assessor considers the two events to have equal probabilities. One advantage of the wheel is that it can evaluate probabilities between zero and one, although "because it is difficult for a subject to discriminate between the sizes of very small sectors, the wheel is most useful for evaluating probabilities in the range from 0.1 to 0.9" (Spetzler and Staël von Holstein, 1975, p. 349). To assess low probability events, an assessor could be asked to state a value that would occur with the same probability as tossing ten heads in a row on a fair coin (approximately 1/1000). This is an example of V-method encoding.

V-Method, Indirect Response, Internal Reference. An example of this assessment combination is the interval technique, which is often used to elicit the median and quantiles of a distribution. One starts with an interval containing all possible values of the uncertain quantity. An arbitrary first split is made, and the assessor is asked which interval he considers most likely. The dividing point is then moved to make that interval smaller. The process is repeated until the assessor is indifferent between the two intervals. This dividing point is the median. Each of the intervals is then split again to obtain the quartiles. Spetzler and Staël von Holstein (1975, p. 350) suggest that "it is usually not meaningful to continue the interval technique after the quartiles have been obtained, because each question depends on earlier responses, and the errors are thus compounded." The interval technique must be used with care, as it is clearly susceptible to the anchoring and adjustment bias.

P-Method, Indirect Response, Internal Reference. To assess the probabilities of quantities with only a few outcomes, the method of relative likelihoods is used. Assessors are asked to assign relative likelihoods, or odds, to two well-defined events, and then to judge how many times more likely the more common event is.

Direct Response Model. P-method, direct response techniques involve assigning cumulative probabilities by answering questions like, "What is the probability that next year's sales will be less than or equal to 2000 units?" V-method techniques assign values, asking questions such as, "What is the level of sales that corresponds to a 75 percent probability?" A common example of a V-method technique is the *fractile method*, which asks, "What is the value of sales such that there is a 0.25 probability that the true value is equal to or less than this value?" A typical PV-method would ask an assessor to draw a probability density or cumulative distribution function. PV-methods are most useful when subjects have a great deal of prior experience thinking in terms of distributions, for example, when eliciting priors from statisticians.

Hora et al. (1992) make comparisons between the calibration of assessments made using direct elicitation procedures, which ask assessors to provide probabilities for intervals of values, and those obtained using bisection methods (the interval technique), which require that intervals of values be subdivided into equally likely subintervals. Their sub-

jects, fifty scientists and engineers participating in probability elicitation training, were given sixteen almanac questions, eight to be assessed using the direct elicitation procedures, and eight to be assessed using bisections methods. For both methods, subjects were required to assess endpoints. Subjects found the bisection method more difficult to work with than the direct assessment method, having particular trouble making a second division to assess the quartiles. There was little difference, however, in the calibration of responses between the two methods, although the subjects performed better if they were able to bound their distributions.

It has also been shown that many assessors have a great deal of trouble establishing end points, the largest and smallest values a variable can attain, since end points are very susceptible to the overconfidence bias (Hora et al., 1992; Alpert and Raiffa, 1982). In one study (Alpert and Raiffa, 1982), 98 percent credible intervals contain the true value only about half the time. Hora et al. (1992, p. 135) present evidence that "the tendency to understate the spread of the distribution is not an artifact of a particular elicitation scheme, but is due instead to a persistent bias in judgment formulation."

5.1.3 Informational Phase

Following the assessment of a probability distribution, it is useful to make verifying checks to see if the assessor agrees with everything that his elicited values imply. This is done by eliciting the same quantity using different methods. If the elicited probabilities do not agree, the accepted reconciliation technique is discussing the differences with the expert and allowing him to revise his opinion (Winkler, 1967; Spetzler and Staël von Holstein, 1975; Merckhofer, 1987). However, Winkler (1967) notes a tendency for subjects to "split-the-difference" between discrepant assessments. "This may be because the subjects wanted to do this after careful thought; but it seems more likely that it is because it was an easy way to make the reconciliations" (p. 791). Hogarth (1975) makes the point that assessors should be encouraged to try to incorporate conflicting information, not to simply ignore it.

6 Feedback And Training

6.1 Feedback

Hogarth makes the point that one needs to capitalize on possible gains of consistency as a function of feedback and experience. He cites studies that suggest that outcome feedback is less effective than feedback that emphasizes the structure of the task (i.e., the relationships between cues in the environment and the variable to be predicted). Sharp et al. (1988) find that while outcome feedback does not improve calibration or overconfidence, it does improve resolution. Alpert and Raiffa (1982) find that task-oriented feedback, for example, pointing out to subjects that they tend to be overconfident, results in more spread distributions, but still does not result in good calibration or the avoidance of "surprises" (answers that fall below the 1st percentile or above the 99th percentile). Winkler (1986) discusses feedback in the form of scoring rules, which are a formal means of evaluating probabilities based on the elicited probabilities and the values of the uncertain quantity that actually occurs. While scoring rules are useful in assessing the quality of judgments, it is still questionable whether they provide effective feedback to non-expert assessors. The effects of various types of feedback on the goodness of subjective probability assessments is an area that requires more research.

6.2 Training

Most decision analysts and statisticians agree that probability elicitation should involve both an interviewer and a subject (Spetzler and Staël von Holstein, 1975; Merckhofer, 1987; Winkler et al., 1992). The interviewer's expertise involves probability elicitation and how to avoid the biases inherent in the process, while the subject should have substantive knowledge of the quantities or variables of interest. It is also widely held (Winkler et al., 1992) that assessors should undergo training before having their probabilities elicited. (A sample of simple training materials appears in Hogarth (1987). Materials targeted at statisticians appear in Berger (1980).) Winkler et al. (1992) identify multiple objectives in training assessors. The first is to motivate the subjects and provide an overview of the process, including how the elicited probabilities will be used.

They suggest that:

> 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 knowledge, but synthesizing disparate and often conflicting sources of information to produce an integrated picture. (pp. 2–3)

The second objective of training is to develop the assessor's confidence in his ability to express his judgments as probabilities. The assessors must also be made aware of possible cognitive and motivational biases. For example, to avoid the anchoring and adjustment heuristic, it makes sense to probe the extreme areas of the probability distribution before looking at the middle. In his 1975 article, Hogarth suggests that the effects of task characteristics should be discussed in training (i.e., response modes, payoffs, and order of information presentation). The third motivation is to insure that assessors have access to relevant background information and evidence specific to the questions of interest, and an opportunity to review these materials.

A few studies have considered the effects of prior training, or "expert knowledge," on the assessment of probability distributions. It is interesting to note that Winkler (1967) compares three different levels of statistical sophistication (normative goodness): Ph.D. level statisticians, business students with a knowledge of "introductory" statistics, and subjects with no statistical knowledge. Only the statisticians were able to consistently assess distributions across elicitation techniques. Winkler states:

> ...it seems that relatively limited prior experience has little effect with regard to the assessment of probability distributions; sophistication in these areas apparently does have quite an effect. (p. 789)

Wright and Anderson (1989, p. 68) find that "the anchoring effect is so dominant that increasing situational familiarity did not result in decreased anchoring." However, Johnson et al. (1991) find that in a task where naive and expert subjects were asked to estimate the probabilities of various ways of making an out in baseball, expertise was

able to overcome anchoring. They hypothesize that a strong mental representation of events can overcome the anchoring bias.

7 Group Assessment

The assessment of subjective probability distributions is not always confined to the individual. Often, it is useful to obtain the views of a group. Winkler et al. (1992) suggest three reasons why it is useful to aggregate the judgments of multiple experts:

- 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.
- It is easier to use a single distribution for further analysis. (pp. 2–6)

There are two types of methods for combining the elicited probabilities of a group of assessors: behavioral and mechanical. Behavioral aggregation involves some degree of contact and interaction among the members of the group, while mechanical approaches are primarily mathematical or statistical, ranging from simple averaging to complex Bayesian techniques (Lindley et al., 1979; Genest and Schervish, 1985; West, 1988). Hogarth (1975) discusses some of the formal statistical methods for aggregating group probabilities, complaining that "unfortunately, group probability assessment studies done to date have only paid lip service to the existing body of social-psychological knowledge" (p. 283). He calls for more work on the differences between "ad hoc" and "traditional" groups, and says:

> Perhaps the only firm recommendation one can currently make to groups of individuals seeking to assess single distributions is to use sensitivity analysis to identify the crucial aspects of the assessment task. If individuals disagree, how important is such disagreement relative to the problem at hand? (p. 283)

While group assessment does have relevance to the individual assessor, there are still sizable problems to be solved outside the group framework.

8 What Do Statisticians Need?

After this discussion of probability elicitation, it is interesting to consider the kinds of distributions that statisticians are interested in eliciting. There are certainly situations in which an expert's opinion on the probability of a specific event are of interest. Formal statistical methods for elicitation of these types of quantities are discussed in Gavasakar (1988). This work concentrates on precisely what quantities need to be elicited to provide enough information to reconstruct the underlying distribution. A more complicated problem is presented in Kadane et al. (1980). In this problem, they are interested in performing a Bayesian linear regression analysis. To do this properly, one needs to elicit priors on the beta coefficients and on the variance of the random error. These are not quantities about which experts have much prior knowledge. In order to frame the problem in a way that can provide a meaningful assessment, the elicitation must be performed in terms of predictive distributions. A predictive distribution is the assessor's best guess of the value of the dependent variable conditional on the independent variables. By eliciting facts about predictive distributions at various levels of the independent variables, the parameterized form of the predictive model can be derived. Even after the problem is reframed in this light, the questions that must be answered are non-trivial and non-intuitive. How to reframe elicitation problems into a form that is meaningful to the expert and yet provides enough information to reconstruct a useful statistical model is an active area of research in Bayesian statistics.

9 Conclusions

Hogarth's primary conclusion about probability assessment is that "explicit attention should be given to the conclusion that man does have difficulty in acting as an 'intuitive' statistician" (p. 284). Much work has been done from this perspective. The psychological literature is full of studies that demonstrate both man's limited informa-

tion processing capacity and the importance of task effects in the choice of problem solving strategy. Studies of non-expert subjects answering almanac questions, while not providing meaningful probability assessments, do provide insight into the cognitive strategies used in making subjective probability assessments. Decision analysts have used this research to structure elicitation procedures that ask questions that take into account cognitive simplification methods, and yet provide meaningful information to their clients. Statisticians are beginning to look at ways of asking questions that can produce both meaningful and statistically useful responses.

The effects of feedback and training need to be more carefully studied. While most researchers agree that feedback and training are necessary, there is little systematic evidence on what types of feedback improve calibration, discrimination, and other measures of goodness. Few studies assess how effective training is at overcoming the biases caused by cognitive simplification mechanisms. There is also little work addressing what types of elicitation procedures are effective in what situations. As Hogarth mentions, "The success of any judgmental strategy will necessarily depend on the extent to which it is suited to the characteristics of the task" (p. 284). He suggests the development of a taxonomy of assessment task characteristics that could be used to select appropriate elicitation techniques. I believe that subsequent research has shown that such a procedure is doomed to fail, because task effects are too pervasive to be easily categorized.

Ginossar and Trope (1987) propose that people use both statistical and non-statistical rules to make probabilistic judgments. They find that prior activation of rules, their relation to the goals of the task, and their applicability to the particular problem influence which problem-solving strategy people choose. They suggest that instead of asking whether people are inherently good or bad statisticians, the focus should be placed on the cognitive factors that determine the application of inferential rules. This line of research should lead to new insight into probability assessment and effective training methods.

Tversky (1974) puts things in perspective, when he notes that:

> The judgments must be compatible with the entire web of beliefs held by the individual, and not only consistent among themselves. Compatibility among

beliefs is the essence of rational judgment. (p. 158)

This points out how complementary the work of statisticians and psychologists can be in the development of subjective probability assessment.

References

[1] Alpert, M. and Raiffa, H. (1982). A progress report on the training of probability assessors. In *Judgment Under Uncertainty: Heuristics and Biases*, Eds. D. Kahneman, P. Slovic, and A. Tversky, pp. 294–305. Cambridge: Cambridge University Press.

[2] Arkes, H. R., Dawes, R. M., and Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, 37, 93–110.

[3] Berger, J. O. (1980). *Statistical Decision Theory: Foundations, Concepts, and Methods*, 2nd ed. New York: Springer-Verlag.

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

[5] Edwards, W. and von Winterfeldt, D. (1986). On cognitive illusions and their implications. In *Judgment and Decision Making: An Interdisciplinary Reader*, Eds. H. R. Arkes and K. R. Hammond, pp. 642–679. Cambridge: Cambridge University Press.

[6] Edwards, W. von Winterfeldt, D., and Moody, D. L. (1988). Simplicity in decision analysis: An example and a discussion. In *Decision Making: Descriptive, Normative, and Prescriptive Interactions*, Eds. D. Bell, H. Raiffa, and A. Tversky, pp. 443–464. Cambridge: Cambridge University Press.

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

[8] Gavasakar, U. (1988). A comparison of two elicitation methods for a prior distribution for a binomial parameter. *Management Science*, 34, 784–790.

[9] Genest, C. and Schervish, M. J. (1985). Modeling expert judgments for Bayesian updating. *Annals of Statistics*, 13, 1198–1212.

[10] Ginossar, Z. and Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology*, 52, 464–474.

[11] Hawkins, S. A. and Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, 107, 311–327.

[12] Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions (with discussion). *Journal of the American Statistical Association*, 70, 271–294.

[13] Hogarth, R. M. (1987). *Judgment and Choice: The Psychology of Decision*, 2nd ed. Chichester, England: John Wiley & Sons.

[14] Hogarth, R. M. and Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1–55.

[15] Hora, S. C., Hora, J. A., and 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.

[16] Jennings, D. L., Amabile, T. M., and Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In *Judgment under Uncertainty: Heuristics and Biases*, Eds. D. Kahneman, P. Slovic, and A. Tversky, pp. 211–230. Cambridge: Cambridge University Press.

[17] Johnson, R. D., Rennie, R. D. and Wells, G. L. (1991). Outcome trees and baseball: A study of expertise and list-length effects. *Organizational Behavior and Human Decision Processes*, 50, 324–340.

[18] Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., and Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75, 845–854.

[19] Kahneman, D. and Tversky, A. (1982). Intuitive prediction: Biases and corrective procedures. In *Judgment under Uncertainty: Heuristics and Biases*, Eds. D. Kahneman, P. Slovic, and A. Tversky, pp. 414–421. Cambridge: Cambridge University Press.

[20] Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77, 217–273.

[21] Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107–118.

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

[23] Lindley, D. V., Tversky, A., and Brown, R. V. (1979). On the reconciliation of probability assessments. *Journal of the Royal Statistical Society, Series A*, 142, 146–180.

[24] March, J. G. and Shapira, Z. (1987). Managerial perspectives on risk and risk taking. *Management Science*, 33, 1404–1418.

[25] Martin, D. W. and Gettys, C. F. (1969). Feedback and response mode in performing a Bayesian decision task. *Journal of Applied Psychology*, 53, 413–418.

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

[27] Murphy, A. H. and Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, 79, 489–500.

[28] O'Connor, M. and Lawrence, M. (1989). An examination of the accuracy of judgmental confidence intervals in time series forecasting. *Journal of Forecasting*, 8, 141–155.

[29] Peterson, C. R. and Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29–46.

[30] Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the Amererican Statistical Association*, 66, 783–801.

[31] Sharp, G. L., Cutler, B. L., and Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes*, 42, 271–283.

[32] Spetzler, C. S. and Staël von Holstein, C. S. (1975). Probability encoding in decision analysis. *Management Science*, 22, 340–358.

[33] Tversky, A. (1974). Assessing uncertainty. *Journal of the Royal Statistical Society, Series B*, 36, 148–159.

[34] Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.

[35] Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.

[36] Wallsten, T. S. and Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, 29, 151–173.

[37] West, M. (1988). Modeling expert opinion. In *Bayesian Statistics 3: Proceedings of the Third Valencia International Meeting, June 1-5, 1987*, Ed. J. M. Bernardo, pp. 493–508. Oxford: Oxford University Press.

[38] Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 62, 776–800.

[39] Winkler, R. L. (1986). On "Good Probability Appraisers." In *Bayesian Inference and Decision Techniques*, Eds. P. Goel and A. Zellner, pp. 265-278. Amsterdam: Elsevier Science Publishers.

[40] Winkler, R. L., Hora, S. C., and Baca, R. G. (1992). The Quality of Expert Judgment Elicitations. Nuclear Regulatory Commission Contract NRC-02-88-005. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

[41] Winkler, R. L. and Poses, R. M. (1991). Evaluating and combining physicians' probabilities of survival in an intensive care unit. Unpublished Manuscript. Durham, NC: Duke University.

[42] Winkler, R. L. and Murphy, A. H. (1968). Good probability assessors. *Journal of Applied Meteorology*, 7, 751–758.

[43] Wright, W. F. and Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organizational Behavior and Human Decision Processes*, 44, 68–82.