9 Aggregating Probability Distributions

Robert T. Clemen and Robert L. Winkler*

ABSTRACT. This chapter is concerned with the aggregation of probability distributions in decision and risk analysis. Experts often provide valuable information regarding important uncertainties in decision and risk analyses because of the limited availability of hard data to use in those analyses. Multiple experts are often consulted in order to obtain as much information as possible, leading to the problem of how to combine or aggregate their information. Information may also be obtained from other sources such as forecasting techniques or scientific models. Because uncertainties are typically represented in terms of probability distributions, we consider expert and other information in terms of probability distributions. We discuss a variety of models that lead to specific combination methods. The output of these methods is a *combined probability distribution*, which can be viewed as representing a summary of the current state of information regarding the uncertainty of interest. After presenting the models and methods, we discuss empirical evidence on the performance of the methods. In the conclusion, we highlight important conceptual and practical issues to be considered when designing a combination process for use in practice.

Introduction

Expert judgments can provide useful information for forecasting, making decisions, and assessing risks. Such judgments have been used informally for many years. In recent years, the use of formal methods to combine expert judgments has become increasingly commonplace. Cooke (1991) reviews many of the developments over the years as attempts have been made to use expert judgments in various settings. Application areas have been diverse, including nuclear engineering, aerospace, various types of forecasting (economic, technological, meteorological, and snow avalanches, to name a few), military intelligence, seismic risk, and environmental risk from toxic chemicals.

In this paper, we consider the problem of using information from multiple sources. Frequently these sources are experts, but sources such as forecasting methods or scientific models can also be used. We will couch much of the discussion in this chapter in terms of experts, but the applicability of the aggregation procedures extends to other sources of information.

* Portions of this chapter are based on Clemen, R. T., and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*, 187–203.

Wu, Apostolakis, and Okrent (1990, p. 170) state that "an important issue related to knowledge representation under uncertainty is the resolution of conflicting information or opinions." Although we discuss information of various kinds, including forecasts, estimates, and probability assessments, our primary focus is on the aggregation of probability distributions. The paper does not pretend to give a comprehensive view of the topic of group judgments; the accumulated knowledge in this field springs from many disciplines, including statistics, psychology, economics, engineering, risk analysis, decision theory, and psychology. Our intent is to highlight the key issues involved in combining probability distributions and to discuss a variety of combining methods.

Because the focus in this paper is on the *combination* of experts' probability distributions, we do not discuss the process by which such probability distributions might be elicited from individual experts. For general discussions oriented toward decision and risk analysis, see Garthwaite, Kadane, and O'Hagan (2005), Hora (1992 and Chapter 8, this volume), Keeney and von Winterfeldt (1989, 1991), Merkhofer (1987), Mosleh, Bier, and Apostolakis (1987), Morgan and Henrion (1990), Otway and von Winterfeldt (1992), and Spetzler and Staël von Holstein (1975). We believe that the elicitation process should be designed and conducted by a decision analysis team composed of individuals knowledgeable about both the substantive issues of interest and individuals knowledgeable about probability elicitation. Moreover, we begin with the assumptions that the experts and the decision analysis team have ironed out differences in definitions, all agree on exactly what is to be forecast or assessed, and as much as possible has been done to eliminate individual cognitive and motivational biases. In this case, of course, it is still possible for reasonable individuals to disagree for a multitude of reasons, ranging from different analytical methods to differing information sets or different philosophical approaches. Indeed, if they never disagreed there would be no point in consulting more than one expert. Morgan and Keith (1995) note that the results of expert elicitations related to climate change "reveal a rich diversity of expert opinion." Consulting multiple experts may be viewed as a subjective version of increasing the sample size in an experiment. Because subjective information is often viewed as being "softer" than "hard scientific data," it seems particularly appropriate to consult multiple experts in an attempt to beef up the information base.

These motivations are reasonable; the fundamental principle that underlies the use of multiple experts is that a set of experts or other information sources can provide more information than a single expert. Although it is sometimes reasonable to provide a decision maker with only the individual experts' probability distributions, the range of which can be studied using sensitivity analysis, it is often necessary to combine the distributions into a single one. In many cases, for example, a single distribution is needed for input into a larger model. Even if this is not the case, it can be illuminating and valuable to generate a combined distribution as a summary of the available information.

Combination, or aggregation, procedures are often dichotomized into *mathematical* and *behavioral* approaches, although in practice aggregation might involve some aspects of each. Mathematical aggregation methods consist of processes or analytical models that operate on the individual probability distributions to produce a single *combined probability distribution*. For example, we might simply take the averages of probabilities from multiple experts. Reviews of the literature on mathematical combination of probability distributions include Winkler (1968), French (1985), Genest and Zidek (1986), Cooke (1991), and French and Ríos Insua (2000). Bunn (1988), Clemen (1989), and Armstrong (2001a) review the broader area of combining forecasts. Mathematical aggregation methods range from simple summary measures, such as arithmetic or geometric means of probabilities, to procedures based on axiomatic approaches or on various models of the information-aggregation process that require inputs regarding characteristics of the experts' probabilities, such as bias and dependence.

In contrast, behavioral aggregation approaches attempt to generate agreement among the experts by having them interact in some way. This interaction may be face-to-face or may involve exchanges of information without direct contact. Behavioral approaches consider the quality of individual expert judgments and dependence among such judgments implicitly rather than explicitly. As information is shared, it is anticipated that better arguments and information will be more important in influencing the group and that redundant information will be discounted. Because our focus is on mathematical approaches for aggregation, we do not discuss behavioral methods in which experts interact in some structured way in order to reach consensus. For more information on behavioral approaches, see Wright and Ayton (1987), Clemen and Winkler (1999), and Armstrong (2001b).

In the next section we discuss mathematical methods for combining probability distributions. Some empirical results regarding these methods are then presented, followed by a brief example on the aggregation of seismic risk distributions. To conclude, we summarize our views on the key issues in the combination of probability distributions in decision analysis and risk analysis.

Combination Models and Methods

In this section, we present a variety of mathematical methods and approaches for combining probability distributions. First, we consider axiomatic approaches, which are based on certain desirable properties or axioms. Next, we consider Bayesian approaches, which treat the probability distributions as information and use Bayesian modeling to revise probabilities on the basis of this information.

Axiomatic Approaches

Early work on mathematical aggregation of probabilities focused on axiom-based aggregation formulas. In these studies, the strategy was to postulate certain properties that the combined distribution should follow and then derive the functional form of the combined distribution. French (1985), Genest and Zidek (1986), and French and Ríos Insua (2000) provide critical reviews of this literature; our summary draws heavily on these sources.

An appealing approach to the aggregation of probability distributions is the *linear opinion pool*, so named by Stone (1961), and dating back to Laplace (Bacharach 1979):

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta), \tag{9.1}$$

where n is the number of experts, $p_i(\theta)$ represents expert i's probability distribution for the uncertain quantity θ, the weights w_i sum to one, and $p(\theta)$ represents the combined probability distribution. For simplicity, we will use p to represent a mass function in the discrete case and a density function in the continuous case, and we will ignore minor technical issues involving the difference between the two cases in order to focus on the more important underlying conceptual and practical issues. As a result, we will often use "probabilities" as shorthand for "probabilities or densities" or "probability distributions."

The linear opinion pool is clearly a weighted linear combination of the experts' probabilities, and, as such, it is easily understood and calculated. Moreover, it satisfies a number of seemingly reasonable axioms. Of particular note, the linear opinion pool is the only combination scheme that satisfies the *marginalization property* (MP). Suppose θ is a vector of uncertain quantities, and the decision maker is interested in just one element of the vector, θ_j. According to the MP, the combined probability is the same whether one combines the marginal distributions of θ_j or combines the joint distributions of the vector θ and then calculates the marginal distribution of θ_j.

The weights in Eq. 9.1 clearly can be used to represent, in some sense, the relative quality of the different experts, or sources of information. In the simplest case, the experts are viewed as equivalent, and Eq. 9.1 becomes a simple arithmetic average. If some experts are viewed as "better" than others (in the sense of being more precise because of having better information, for example), the "better" experts might be given greater weight. In some cases, it is possible for some of the weights to be negative (Genest 1984). The determination of the weights is a subjective matter, and numerous interpretations can be given to the weights (Genest and McConway 1990).

Another typical combination approach uses multiplicative averaging and is sometimes called a *logarithmic opinion pool*. In this case, the combined probability distribution is of the form

$$p(\theta) = k \prod_{i=1}^{n} p_i(\theta)^{w_i}, \tag{9.2}$$

where k is a normalizing constant, and the weights w_i satisfy some restrictions to ensure that $p(\theta)$ is a probability distribution. Typically, the weights are restricted to sum to one. If the weights are equal (i.e., each weight is $1/n$), then the combined distribution is proportional to the geometric mean of the individual distributions. This is called a logarithmic opinion pool because the logarithm of the combined distribution can be expressed as a linear combination of the logarithms of the individual distributions.

Equation 9.2 satisfies the principle of *external Bayesianity* (EB). Suppose a decision maker has consulted the experts, has calculated $p(\theta)$, but has subsequently learned some new information relevant to θ. Two choices are available. One is to use the information first to update the individual probability distributions $p_i(\theta)$ and then to combine them. The other is to use the information to update the combined $p(\theta)$ directly. A formula satisfies EB if the result is the same in each case.

Cooke's (1991) *classical method* is a linear opinion pool that has been applied widely in risk assessment, primarily in Europe, beginning in the late 1980s. In order to determine weights, Cooke's method uses probability assessments by the experts on variables for which the analyst knows the outcome in order to calculate a measure of the extent to which the experts are calibrated. (An expert is empirically calibrated if, on examining those events for which the expert has assessed an x percent chance of occurrence, it turns out that x percent actually occur.) An expert's weight in the linear opinion pool is based largely on his or her calibration measure; if the expert's assessments are sufficiently miscalibrated, however, zero weight is assigned. See Cooke (1991) for full details and examples.

Combining rules such as Eqs. 9.1 or 9.2 may be quite reasonable, but not necessarily because of connections with properties such as MP or EB. Difficulties with the axioms are discussed by French (1985) and Genest and Zidek (1986). Lindley (1985) gives an example of the failure of both axioms in a straightforward example, with the interpretation that MP ignores important information, and EB requires that the form of the pooling function not change. In addition, French (1985) points out that *impossibility theorems* exist (along the lines of Arrow's classic work on social choice theory (1951)) whereby a combining rule cannot simultaneously satisfy a number of seemingly compelling desiderata. Combining rules can also affect characteristics of the probabilities; for example, Hora (2004) shows that combining well-calibrated probability distributions can lead to a combined distribution that is not well-calibrated, and it is also possible for a combined distribution to be better calibrated than the individual distributions. Moreover, despite the work of Genest and McConway (1990), no foundationally based method for determining the weights in Eqs. 9.1 or 9.2 is available.

Bayesian Approaches

French (1985), Lindley (1985), and Genest and Zidek (1986) all conclude that for the typical decision analysis or risk analysis situation, in which a group of experts provide information for a decision maker or a risk-assessment team, a Bayesian updating scheme is the most appropriate method. Winkler (1968) provides a Bayesian framework for thinking about the combination of information and ways to assess differential weights. Building on this framework, Morris (1974, 1977) formally establishes a clear Bayesian paradigm for aggregating information from experts. The notion is straightforward. If n experts provide information $g_1, \ldots, g_n$ to a decision maker regarding some event or quantity of interest θ, then the decision maker should use Bayes' theorem to update a prior distribution $p(\theta)$:

$$p^* = p(\theta|g_1, \ldots, g_n) \propto p(\theta)L(g_1, \ldots, g_n|\theta), \tag{9.3}$$

where L represents the likelihood function associated with the experts' information. This general principle can be applied to the aggregation of any kind of information, ranging from the combination of point forecasts or estimates to the combination of individual probabilities and probability distributions. Resting on the solid ground of probability theory, including requirements of coherence as explicated by de Finetti (1937) and Savage (1954), Morris' Bayesian paradigm provides a compelling framework for constructing aggregation models. In the past two decades, attention has shifted from the axiomatic approach to the development of Bayesian combination models.

At the same time that it is compelling, the Bayesian approach is also frustratingly difficult to apply. The problem is the assessment of the likelihood function $L(g_1, \ldots, g_n \mid \theta)$. This function amounts to a probabilistic model for the information $g_1, \ldots, g_n$, and, as such, it must capture the interrelationships among θ and $g_1, \ldots, g_n$. In particular, it must account for the precision and bias of the individual g_is, and it must also be able to model dependence among the g_is. For example, in the case of a point forecast, the precision of g_i is the accuracy with which expert i forecasts θ, and bias is the extent to which the forecast tends to fall consistently above or below θ. Dependence involves the extent to which the forecast errors for different experts are interrelated. For example, if expert i overestimates θ, will expert j tend to do the same?

The notions of *bias*, *precision*, and *dependence* are also crucial, but more subtle, in the case of combining probability distributions. Bias, for example, relates to the extent to which the probabilities are calibrated, as discussed above. Precision relates to the "certainty" of probabilities; a calibrated expert who more often assesses probabilities close to zero or one is more precise. In the case of assessing a probability distribution for a continuous θ, a more precise distribution is one that is both calibrated and, at the same time, has less spread (possibly measured as variance). Dependence among such distributions refers to the tendency of the experts to report similar probabilities.

Because of the difficulty of assessing an appropriate likelihood function "from scratch," considerable effort has gone into the creation of "off-the-shelf" models for aggregating single probabilities (e.g., Lindley 1985; Clemen and Winkler 1987) and probability distributions (Winkler 1981; Lindley 1983). We will review a number of these models.

Bayesian Combinations of Probabilities

Suppose that θ is an indicator variable for a specific event, and the experts provide probabilities that $\theta = 1$ (i.e., that the event will occur). How should these probabilities be combined? Clemen and Winkler (1990) review and compare a number of different Bayesian models that might be applied in this situation. Here we discuss four of these models.

Let p_i $(i = 1, \ldots, n)$ denote expert i's stated probability that θ occurs. Expressed in terms of the posterior odds of the occurrence of θ, $q^* = p^*/(1 - p^*)$, the models are Independence, Genest and Schervish, Bernoulli, and Normal.

INDEPENDENCE

$$q^* = \frac{p_0}{1-p_0} \prod_{i=1}^{n} \frac{f_{1i}(p_i|q=1)}{f_{0i}(p_i|q=0)}, \tag{9.4}$$

where $f_{1i}(f_{0i})$ represents the likelihood of expert i giving probability p_i conditional on the occurrence (nonoccurrence) of θ, and p_0 denotes the prior probability $p(\theta = 1)$. This model reflects the notion that each expert brings independent information to the problem. Depending on how the likelihoods are modeled, adding more experts might or might not mean more certainty (i.e., q^* closer to one or zero); e.g., see Winkler (1986). For example, if all experts say that the probability is 0.6, then under some modeling assumptions for the likelihoods p^* will tend to be much higher than 0.6, whereas under other assumptions p^* will equal the common value of 0.6.

GENEST AND SCHERVISH

$$q^* = \frac{p_0^{1-n} \prod_{i=1}^{n} p_0 + \lambda_i(p_i - \mu_i)}{(1 - p_0)^{1-n} \prod_{i=1}^{n} 1 - [p_0 + \lambda_i(p_i - \mu_i)]}, \tag{9.5}$$

where μ_i is the decision maker's marginal expected value of p_i and λ_i is interpreted as the coefficient of linear regression of θ on p_i. This model is due to Genest and Schervish (1985), and it is derived on the assumption that the decision maker (or assessment team) can assess only certain aspects of the marginal distribution of expert i's probability p_i. It is similar to the independence model, but allows for miscalibration of the p_is in a specific manner.

BERNOULLI

$$p^* = \sum_{i=1}^{n} \beta_i p_i. \tag{9.6}$$

Generally, we assume that $\sum \beta_i = 1$ to ensure that p^* is a probability. This model, which is due to Winkler (1968) and Morris (1983), arises from the notion that each expert's information can be viewed as being equivalent to a sample from a Bernoulli process with parameter θ. The resulting p^* is a convex combination of the p_is, with the coefficients interpreted as being directly proportional to the amount of information each expert has.

NORMAL

$$q^* = \frac{p_0}{1 - p_0} \exp[\mathbf{q}'\boldsymbol{\Sigma}^{-1}(\mathbf{M}_1 + \mathbf{M}_0)\boldsymbol{\Sigma}^{-1}(\mathbf{M}_1 - \mathbf{M}_0)/2] \tag{9.7}$$

where $\mathbf{q}' = (\log[p_1/(1 - p_1)], \ldots, \log[p_n/(1 - p_n)])$ is the vector of log-odds corresponding to the individual probabilities, a prime denotes transposition, and the likelihood functions for $\mathbf{q}$, conditional on $\theta = 1$ and $\theta = 0$, are modeled as normal with means $\mathbf{M}_1$ and $\mathbf{M}_0$, respectively, and common covariance matrix $\boldsymbol{\Sigma}$. This model captures dependence among the probabilities through the multivariate-normal likelihood functions, and is developed in French (1981) and Lindley (1985). Clemen and Winkler (1987) use this model in studying meteorological forecasts.

These four models all are consistent with the Bayesian paradigm, yet they are clearly all different. The point is not that one or another is more appropriate overall, but that different models may be appropriate in different situations, depending on the nature of the situation and an appropriate description of the experts' probabilities. Technically, these differences give rise to different likelihood functions, which in turn give rise to the different models.

Bayesian Models for Combining Probability Distributions

Just as the models above have been developed specifically for combining event probabilities, other Bayesian models have been developed for combining probability distributions for continuous θ. Here we review three of these models.

Winkler (1981) presents a model for combining expert probability distributions that are normal. Assume that each expert provides a probability distribution for θ with mean μ_i and variance σ_i^2. The vector of means $\mu = (\mu_1, \ldots, \mu_n)$ represents the experts' estimates of θ. Thus, we can work in terms of a vector of errors, $\varepsilon = (\mu_1 - \theta, \ldots, \mu_n - \theta)$. These errors are modeled as multivariate normally distributed with mean vector $(0, \ldots, 0)$ (i.e., the estimates are viewed as unbiased) and covariance matrix $\boldsymbol{\Sigma}$, regardless of the value of θ. Let $\boldsymbol{e}' = (1, \ldots, 1)$, a conformable vector of ones. Assuming a noninformative prior distribution for θ, the posterior distribution for θ is normal with mean μ^* and variance σ^{*2}, where

$$\mu^* = \boldsymbol{e}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/\boldsymbol{e}'\boldsymbol{\Sigma}^{-1}\boldsymbol{e}, \tag{9.8a}$$

and

$$\sigma^{*2} = (\boldsymbol{e}'\boldsymbol{\Sigma}^{-1}\boldsymbol{e})^{-1}. \tag{9.8b}$$

In this model the experts' stated variances σ_i^2 are not used directly, although the decision maker may let the ith diagonal element of $\boldsymbol{\Sigma}$ equal σ_i^2. For an extension, see Lindley (1983).

The normal model has been important in the development of practical ways to combine probability distributions. The typical minimum-variance model for combining forecasts is consistent with the normal model (e.g., see Bates and Granger 1969; Newbold and Granger 1974; Winkler and Makridakis 1983). The multivariate-normal likelihood embodies the available information about the qualities of the probability distributions, especially dependence among them. Biases can easily be included in the model via a nonzero mean vector for ε. Clemen and Winkler (1985) use this normal model to show how much information is lost because of dependence, and they develop the idea of equivalent independent information sources. Winkler and Clemen (1992) show how sensitive the posterior distribution is when correlations are high, which is the rule rather than the exception in empirical studies. Chhibber and Apostolakis (1993) also conduct a sensitivity analysis and discuss the importance of dependence in the context of the normal model. Schmittlein, Kim, and Morrison (1990) develop procedures to decide whether to use weights based on the covariance matrix or to use equal weights. Similarly, Chandrasekharan, Moriarty, and Wright (1994) propose methods for investigating the stability of weights and deciding whether to eliminate some experts from the combination.

Although the normal model has been useful, it has some shortcomings. In particular, one must find a way to fit the distributions into the normal framework. If the distributions are unimodal and roughly symmetric, this is generally not a problem. Otherwise, some sort of transformation might be required. The covariance matrix Σ typically is estimated from data; Winkler (1981) derives a formal Bayesian model in which Σ is viewed as an uncertain parameter. Assessing Σ subjectively is possible; Gokhale and Press (1982), Clemen and Reilly (1999), and Clemen, Fischer and Winkler (2000) discuss the assessment of correlation coefficients via a number of different probability transformations. (Winkler and Clemen (2004) show the value of using multiple methods for assessing correlations but argue that averaging correlation judgments from multiple judges is a better strategy.) Finally, the posterior distribution is always a normal distribution and typically is a compromise. For example, suppose two experts give $\mu_1 = 2, \mu_2 = 10$, and $\sigma_1^2 = \sigma_2^2 = 1$. Then the posterior distribution will be a normal distribution with mean $\mu^* = 6$ and variance $(1 + \rho)/2$, where ρ is the correlation coefficient between the two experts' errors. Thus, the posterior distribution puts almost all of the probability density in a region that neither of the individual experts thought likely at all. In a situation such as this, it might seem more reasonable to have a bimodal posterior distribution reflecting the two experts' opinions. Lindley (1983) shows how such a bimodal distribution can arise from a *t*-distribution model.

Another variant on the normal model is a Bayesian hierarchical model (e.g., Lipscomb, Parmigiani, and Hasselblad 1998) that allows for differential random bias on the part of the experts by assuming that each expert's error mean in the normal model can be viewed as a random draw from a second-order distribution on the error means. Dependence among experts' probabilities arises through the common second-order distribution. Hierarchical models generally result in an effective shrinkage of individual means to the overall mean and tend to provide robust estimates. Because they involve another layer of uncertainty, they can be complex, particularly when the parameters are all viewed as unknown and requiring prior distributions.

Some effort has been directed toward the development of Bayesian aggregation methods that are suitable for the use of subjective judgment in determining the likelihood function. Clemen and Winkler (1993), for example, present a process for subjectively combining point estimates; the approach is based on the sequential assessment of conditional distributions among the experts' forecasts, where the conditioning is specified in an influence diagram. Formal attention has also been given to Bayesian models for situations in which the experts provide only partial specifications of their probability distributions (e.g., moments, fractiles) or the decision maker is similarly unable to specify the likelihood function fully (Genest and Schervish 1985; West 1992; West and Crosse 1992; Gelfand, Mallick, and Dey 1995). Bunn (1975) develops a model that considers only which expert performs best on any given occasion and uses a Bayesian approach to update weights in a combination rule based on past performance.

Jouini and Clemen (1996) develop a method for aggregating probability distributions in which the multivariate distribution (likelihood function) is expressed as a function of the marginal distributions. A *copula* function (e.g., Dall'Aglio, Kotz, and Salinetti et al. 1991) provides the connections, including all aspects of dependence, among the experts' judgments as represented by the marginal distributions. For example, suppose that expert i assesses a continuous density for θ, $f_i(\theta)$, with corresponding cumulative distribution function $F_i(\theta)$. Then Jouini and Clemen show that under reasonable conditions, the decision maker's posterior distribution is

$$P(\theta \mid f_1, \ldots, f_n) \propto c[1 - F_1(\theta), \ldots, 1 - F_n(\theta)] \prod_{i=1}^{n} f_i(\theta) \tag{9.9}$$

where c represents the copula density function.

In the copula approach, judgments about individual experts are entirely separate from judgments about dependence. Standard approaches for calibrating individual experts (either data based or subjective) can be used and involve only the marginal distributions. On the other hand, judgments about dependence are made separately and encoded into the copula function. Regarding dependence, Jouini and Clemen suggest using the Archimedian class of copulas, which treat the experts symmetrically in terms of dependence. If more flexibility is needed, the copula that underlies the multivariate normal distribution can be used (Clemen and Reilly 1999).

In other work involving the specification of the likelihood function, Shlyakhter (1994) and Shlyakhter, Kammen, Brodio, and Wilson (1994) develop a model for adjusting individual distributions to account for the well-known phenomenon of overconfidence, and they estimate the adjustment parameter for two different kinds of environmental risk variables. Hammitt and Shlyakhter (1999) show the implications of this model for combining probabilities.

A promising new direction for specifying the likelihood function uses Bayesian nonparametric methods. The typical use of such approaches is in specifying a nonparametric prior (e.g., based on a Dirichlet process) and then updating that prior. In combining experts' probability distributions, though, it is the likelihood function for the expert's probability distribution that is modeled using Bayesian nonparametric methods. West (1988) is the first to have developed such a model.

Building on West's approach, Lichtendahl (2005) combines a parameterized model of expert performance with a Bayesian nonparametric likelihood function for expert judgments that is capable of handling not only multiple experts, but multivariate judgments (for multiple uncertain quantities) as well.

In this section, we have discussed a number of mathematical methods for combining probability distributions. A number of important issues should be kept in mind when comparing these approaches and choosing an approach for a given application. These issues include the type of information that is available (e.g., whether full probability distributions are given or just some partial specifications of these distributions); the individuals performing the aggregation of probabilities (e.g., a single decision maker or analyst, a decision-analysis or risk-assessment team, or some other set of individuals); the degree of modeling to be undertaken (assessment of the likelihood function, consideration of the quality of the individual distributions); the form of the combination rule (which could follow directly from modeling or could be taken as a primitive, such as a weighted average); the specification of parameters of the combination rule (e.g., the weights); and the consideration of simple versus complex rules (e.g., simple averages versus more complex models). The empirical results discussed below will shed some light on some of these questions, and we will discuss these issues further in the Conclusion.

Empirical Evidence

The various combining techniques discussed above all have some intuitive appeal, and some have a strong theoretical basis given that certain assumptions are satisfied. The proof, of course, is in the pudding. How do the methods perform in practice? Do the combination methods lead to "improved" probability distributions? Do some methods appear to perform better than others? Some evidence available from experimentation, analysis of various data sets, and actual applications, including work on combining forecasts that is relevant to combining probabilities, is reviewed in this section.

Mathematical versus Intuitive Aggregation

Before comparing different combination methods, we should step back and ask whether one should bother with formal aggregation methods. Perhaps it suffices for the decision maker or the decision analysis team to look at the individual probability distributions and to aggregate them intuitively, directly assessing a probability distribution in light of the information.

Hogarth (1987) discusses the difficulty humans have in combining information from different data sources. Although his discussion covers the use of all kinds of information, his arguments apply to the aggregation of probability distributions. Hogarth shows how individuals tend to ignore dependence among information sources, and he relates this to Kahneman and Tversky's (1972) *representativeness* heuristic. In a broad sense, Hogarth's discussion is supported by psychological experimentation showing that expert judgments tend to be less accurate than statistical models based on criteria that the experts themselves claim to use. Dawes, Faust, and Meehl (1989) provide a review of this literature.

Clemen, Jones, and Winkler (1996) also study the aggregation of point forecasts. However, they use Winkler's (1981) normal model and Clemen and Winkler's (1993) conditional-distributions model, and they compare the probability distributions derived from these models with intuitively assessed probability distributions. Although their sample size is small, the results suggest that mathematical methods are somewhat better in terms of performance than intuitive assessment, and the authors speculate that this is due to the structured nature of the assessments required in the mathematical-aggregation models.

Comparisons among Mathematical Methods

Some evidence is available regarding the relative performance of various mathematical aggregation methods. In an early study, Staël von Holstein (1972) studied averages of probabilities relating to stock market prices. Most of the averages performed similarly, with weights based on rankings of past performance slightly better than the rest.

Seaver (1978) evaluated simple and weighted averages of individual probabilities. The performance of the different combining methods was similar, and Seaver's conclusion was that simple combination procedures, such as an equally-weighted average, produce combined probabilities that perform as well as those from more complex aggregation models. Clemen and Winkler (1987) reported similar results in aggregating precipitation probability forecasts.

In a follow-up study, Clemen and Winkler (1990) studied the combination of precipitation forecasts using a wider variety of mathematical methods. One of the more complex methods that was able to account for dependence among the forecasts performed best. Although a simple average was not explicitly considered, a weighted average that resulted in weights for the two forecasts that were not widely different performed almost as well as the more complex scheme.

Winkler and Poses (1993) report on the combination of experts' probabilities in a medical setting. For each patient in an intensive care unit, four individuals (an intern, a critical care fellow, a critical care attending, and a primary attending physician) assessed probabilities of survival. All possible combinations (simple averages) of these four probabilities were evaluated. The best combination turned out to be an average of probabilities from the two physicians who were, simultaneously, the most experienced and the least similar, with one being an expert in critical care and the other having the most knowledge about the individual patient.

All of these results are consistent with the general message that has been derived from the vast empirical literature on the combination of point forecasts. That message is that, in general, simpler aggregation methods perform better than more complex methods. Clemen (1989) discusses this literature. In some of these studies, taking into account the quality of the information, especially regarding relative precision of forecasts, turned out to be valuable.

The above studies focus on the combination of point forecasts or event probabilities, and mathematical methods studied have been either averages or something more complex in which combination weights were based on past data. Does the result that simpler methods work better than more complex methods carry over to the aggregation of probability distributions, especially when the quality of the probability distributions must be judged subjectively? Little specific evidence appears to be available on this topic. Clemen, Jones, and Winkler (1996) reported that Winkler's (1981) normal model and the more complex conditional-distributions model (Clemen and Winkler 1993) performed at about the same level. These results are consistent with a number of other studies on the value of decomposing judgments into smaller and more manageable assessment tasks. Ravinder, Kleinmuntz and Dyer (1988) provide a theoretical argument for the superiority of decomposed assessments. Wright, Saunders, and Ayton (1988), though, found little difference between holistic and decomposed probability assessments; on the other hand, Hora, Dodd, and Hora (1993) provide empirical support for decomposition in probability assessment. With regard to decomposition and the assessment of point estimates, Armstrong, Denniston, and Gordon (1975) and MacGregor, Lichtenstein, and Slovic (1988) found that decomposition was valuable in improving the accuracy of those estimates. Morgan and Henrion (1990) review the empirical support for decomposition in probability assessment, and Bunn and Wright (1991) do the same for forecasting tasks in general. A tentative conclusion is that, for situations in which the aggregation must be made on the basis of subjective judgments, appropriate decomposition of those judgments into reasonable tasks for the assessment team may lead to better performance.

Mathematical versus Behavioral Aggregation

Most of the research comparing mathematical and behavioral aggregation has focused on comparisons with a simple average of forecasts or probabilities rather than with more complicated mathematical combination methods. Results from these comparisons have been mixed. For example, for forecasting college students' grade-point averages, Rohrbaugh (1979) found that behavioral aggregation worked better than taking simple averages of individual group members' forecasts. Hastie (1986), Hill (1982), and Sniezek (1989) reached similar conclusions. However, Lawrence, Edmundson, and O'Connor (1986) reported that mathematical combination improved on the behavioral combination of forecasts. In Flores and White's (1989) experiment, mathematical and behavioral combinations performed at approximately the same level. Goodman (1972) asked college students to assess likelihood ratios in groups and individually; the behavioral combination showed slight improvement over the mechanical combination.

Seaver (1978) asked student subjects to assess discrete and continuous probability distributions for almanac questions. Several different conditions were used: individual assessment, Delphi, Nominal Group Technique, free-form discussion, and two other approaches that structured information sharing and discussion. Both simple averages and weighted averages of the individual probabilities were also calculated. The conclusion was that interaction among the assessors did not improve on the performance of the aggregated probabilities, although the subjects did feel more satisfied with the behavioral aggregation results.

Reagan-Cirincione (1994) used an intensive group process intervention involving cognitive feedback and a computerized group support system for a quantity-estimation task. The results show this to be the only study reviewed by Gigone and Hastie (1997) for which group judgments are more accurate than a mathematical average of the individual judgments. In general, Gigone and Hastie conclude that the evidence indicates that a simple average of individual judgments tends to outperform group judgments. Moreover, they discuss ways in which groups might improve over mathematical combinations and conclude that, "there is a limited collection of judgment tasks in which groups have a legitimate opportunity to outperform individual judgments."

Larrick and Soll (2006) study how people think about averaging. Their experiments reinforce Gigone and Hastie's (1997) conclusion. Larrick and Soll point out that the strong performance of averaging results from *bracketing*, when forecasters fall on opposite sides of the actual value. The problem, however, is that people do not appreciate what bracketing implies; instead, people tend to think that averaging forecasts tends to lead to average performance, which in turn contributes to a tendency to try to identify the best expert, when averaging forecasts would be a better strategy.

Example: Seismic Hazard

In this section, we present a brief example of three methods for aggregating subjective probability distributions. In 1989, the Nuclear Regulatory Commission reported on the seismic hazard relating to nuclear power plants in the eastern United States (Bernreuter, Savy, Mensing, and Chen 1989). As part of a subsequent study to update the hazard report, probability densities were elicited from a number of seismic-hazard experts. For example, Figure 9.1 shows probability densities for seven experts, each of whom assessed the peak ground acceleration (cm/sec^2) at a distance of 5 km from the epicenter of a magnitude-5 earthquake. Lognormal densities were fitted to the assessments for each expert.

Figure 9.2 shows three aggregated distributions. The "average" curve is the simple average of the seven individual densities. The other two aggregations use Winkler's (1981) normal model and the copula model described by Jouini and Clemen (1996). Although no formal judgments regarding the quality of the experts were made following the elicitation, we assume that appropriate elicitation procedures eliminated biases as much as possible, and so the individual densities need not be recalibrated. For the normal, we assume for each pair of experts a correlation coefficient of 0.9, a value not inconsistent with empirical findings (Clemen 1989). The copula model uses a copula from Frank's family with parameter $\alpha =$ 0.00001104 (see Jouini and Clemen 1996). This value of α encodes a level of dependence between any two experts that is consistent with Kendall's $\tau = 0.70$ and pairwise product–moment correlations on the order of 0.9. For both the normal and

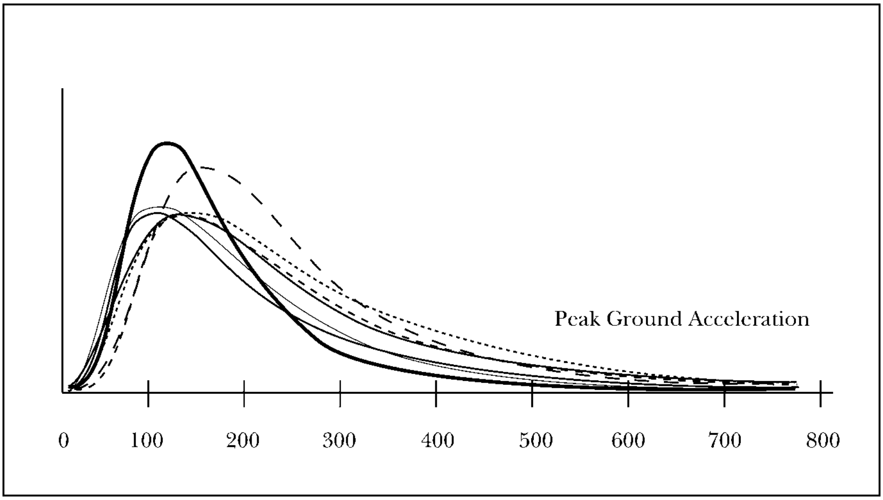

Figure 9.1. Seven subjectively assessed probability distributions for peak ground acceleration (cm/sec^2).

copula models, we transformed to log(cm/sec^2) to perform the aggregation using normal marginals, and then transformed back after aggregating.

The differences among these three models are suggestive of practical experience with aggregation methods. The average curve, which must take into account all variability in all distributions, is very spread out – in a sense understating the amount of information embodied by the experts' distributions as a whole. The model-based aggregations reflect an increase in information through the decrease in spread, placing the bulk of the aggregated density between 100 and 200 cm/sec^2,

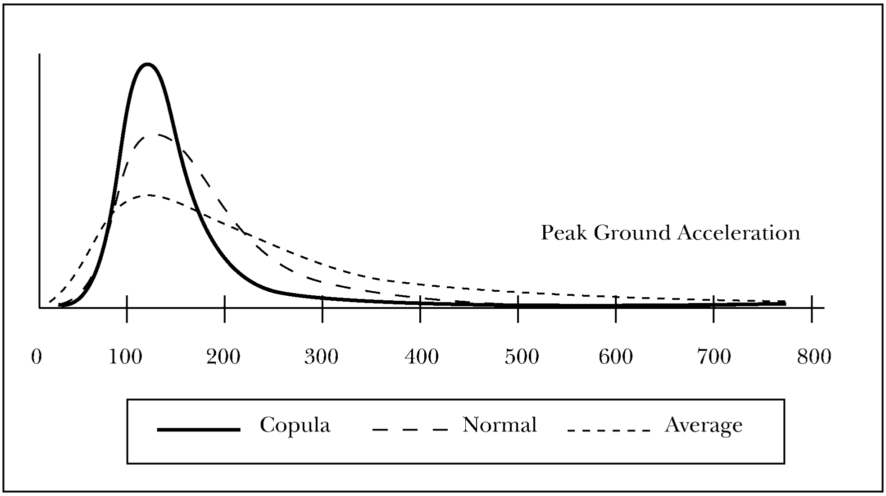

Figure 9.2. Three methods for aggregating individual probability distributions.

which is generally consistent with the individual distributions. Both are much more spread out, however, than the density (not shown) that would result from an assumption of independent experts.

Our example demonstrates how different combination methods might behave in a realistic setting. In practical applications, it is often necessary to adapt the methodology to the particular situation. Selected recent applications include Dillon, John, and von Winterfeldt (2002), who use a simulation-based combining method to combine expert cost judgments in an analysis of alternative tritium suppliers for the U.S. Department of Energy; Lacke and Clemen (2001), who apply the copula method to colorectal cancer risk estimates; Lipscomb, Parmigiani, and Hasselblad (1998), who use a hierarchical Bayes model in a study of physician staffing for U.S. Department of Veteran Affairs medical centers; and Merrick, van Dorp, and Singh's (2005) risk analysis of the Washington State ferries.

Conclusion

We have reviewed a variety of methods for combining probability distributions in decision and risk analysis. The empirical results reviewed in the previous section suggest that mathematical aggregation outperforms intuitive aggregation and that mathematical and behavioral approaches tend to be similar in performance, with mathematical rules having a slight edge. As for different mathematical combination methods, simple combination rules (e.g., a simple average) tend to perform quite well. More complex rules sometimes outperform the simple rules, but they can be somewhat sensitive, leading to poor performance in some instances. All of these conclusions should be qualified by noting that they represent tendencies over a variety of empirical studies, generally conducted in an experimental setting as opposed to occurring in the context of real-world decision analysis. These studies do not, unfortunately, directly assess the precise issue that needs to be addressed. For the purpose of the typical decision analysis in which probability distributions are to be combined, but limited past data are available, the real question is, "What is the best way to combine the probabilities?" Thus, although we should pay careful attention to available empirical results and learn from them, we should think hard about their generalizability to realistic decision-analysis applications.

Both the mathematical combination of probabilities with some modeling and the use of interaction among the experts have some intuitive appeal. It is somewhat disappointing, therefore, to see that modeling and behavioral approaches often provide results inferior to simple combination rules. We feel a bit like the investor who would like to believe that some careful analysis of the stock market and some tips from the pros should lead to high returns but finds that buying a mutual fund that just represents a stock market index such as the S & P 500 would yield better returns. On the other hand, we should remember that the simple combination rules do perform quite well, indicating that the use of multiple experts (or, more generally, multiple information sources) and the combination of the resulting probabilities can be beneficial. One message that comes from the work on the combination of probabilities is that, at a minimum, it is worthwhile to consult multiple information sources (including experts and other sources) and combine their probabilities.

Another message is that further work is needed on the development and evaluation of combination methods. The challenge is to find modeling procedures or behavioral approaches (or processes involving both modeling aspects and behavioral aspects) that perform well enough to justify the extra cost and effort that is associated with serious modeling or expert interaction. On the behavioral side, Davis (1992) states: "The engineering of increases in decision performance while maintaining [the advantages of group decision making] is a proper challenge for fundamental theory and research in applied psychology." Gigone and Hastie (1997) echo this in their concluding comments: "The quality of group decisions and judgments is an essential ingredient in democratic institutions and societies, but research on group judgment accuracy is stagnant.... Better methods and analyses will help behavioral scientists, engineers, and policymakers to design and select group decision-making procedures that will increase efficiency, justice, and social welfare."

Regarding mathematical combining procedures, we believe that simple rules will always play an important role, because of their ease of use, robust performance, and defensibility in public-policy settings where judgments about the quality of different experts are eschewed. However, we also believe that further work on Bayesian models, with careful attention to ease of modeling and assessment, as well as to sensitivity (e.g., avoiding extreme situations such as highly negative weights), can lead to improved performance. In principle, the Bayesian approach allows for careful control in adjusting for the quality of individual expert distributions (including overconfidence) and dependence among experts. One reason that the simple average works well is that it results in a distribution that is broader than any individual distribution, thereby counteracting in part typical expert overconfidence. Our preference would be to use a Bayesian model that permits the explicit modeling and appropriate adjustment for overconfidence as well as dependence. It is also worth noting that the normal and copula models allow the symmetric treatment of the probability distributions, in which case simple combining rules fall out of these models as special cases.

Generally, the process of combining probability distributions in decision and risk analysis may well involve both mathematical and behavioral aspects and should be considered in the context of the overall process for obtaining and utilizing information (including expert judgment) in a given application. For discussions of this process, see the references cited in the introduction. Important issues to consider include the following:

- *Flexibility and Process Design*. We believe that there is no single, all-purpose combining rule or combining process that should be used in all situations. Rather, the design of the combining process (as part of the overall information-gathering process) should depend on the details of each individual situation. This process design should take into account factors such as the nature and importance of the uncertainties, the availability of appropriate experts or other information sources, past evidence available about the information sources and about the quantities of interest, the degree of uncertainty about quantities of interest, the degree of disagreement among the probability distributions, the costs of different information sources (e.g., the cost of consulting an expert or bringing experts together), and a variety of other factors. We believe that a carefully structured and documented process is appropriate.

- *The Role of Modeling versus Rules*. Decision and risk analysis applications involving the combination of probability distributions or other quantities have often used simple combination rules, usually a simple average. Such simple rules are valuable benchmarks, but careful consideration should be given to modeling in order to include, in a formal fashion, factors such as the quality of the information from individual sources and the dependence among different sources. One possible scenario is that the experts or other information sources are judged to be exchangeable and their probabilities should be treated symmetrically, but this should be a conscious choice on the part of the decision analysts or decision makers. The degree of modeling will vary from case to case, ranging from fairly simple modeling (e.g., unequal weights based on judgments of relative precision) to more detailed modeling (e.g., building a full copula-based model).
- *The Role of Interaction*. This aspect of the combination process will also vary from case to case, depending on such factors as the perceived desirability of exchanging information and the ease with which such information can be exchanged. Evidence to date does not provide strong support for benefits of interaction, yet it has considerable intuitive appeal and has been used in risk analysis applications (e.g., EPRI 1986; Hora and Iman 1989; Winkler et al. 1995). We feel that the jury is still out on the impact of interaction on the quality of the resulting combined probabilities and that any benefits are most likely to come from exchanges of information (possibly including individual experts' probability distributions) as opposed to forced consensus through group probability assessments. This implies that mathematical combination will still be needed after interaction and individual probability assessment or reassessment. Also, it is important that the interaction process be carefully structured with extensive facilitation.
- *The Role of Sensitivity Analysis*. It is helpful to conduct a sensitivity analysis to investigate the variation in the combined probabilities as parameters of combining models are varied. This can help in decisions regarding the scope of the modeling effort. A related note is that reporting the individual probabilities, as well as any combined probabilities, provides useful information about the range of opinions in a given case.
- *The Role of the Analyst or the Analysis Team*. As should be clear from the above discussion, the decision analyst, or the team of decision analysts or risk analysts, plays a very important role in the combination of probability distributions as well as in all other aspects of the information-gathering process. With respect to mathematical combination, this includes performing any modeling, making any assessments of information quality that are needed in the modeling

process, and choosing the combination rule(s) to be used. On the behavioral side, it includes structuring any interaction and serving as facilitators. In general, the analysts are responsible for the design, elicitation, and analysis aspects of the combination process.

In summary, the combination of probability distributions in decision analysis and risk analysis is valuable for encapsulating the accumulated information for analysts and decision makers and providing the current state of information regarding important uncertainties. Normatively and empirically, combining can lead to improvements in the quality of probabilities. More research is needed on the potential benefits of different modeling approaches and the development of mathematical combination rules.

Acknowledgments

This research was supported in part by the National Science Foundation under Grants SBR-93–20754, SBR-94–22527, SBR-95–96176, and SES-03–17867.

REFERENCES

Armstrong, J. S. (2001a). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer, pp. 417–439.

Armstrong, J. S. (Ed.). (2001b). *Principles of forecasting: A handbook for researchers and practitioners*. Norwell, MA: Kluwer.

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

Arrow, K. J. (1951). *Social choice and individual values*. New York: Wiley.

Bacharach, M. (1979). Normal Bayesian dialogues. *Journal of the American Statistical Association, 74*, 837–846.

Bates, J. M., and Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly, 20*, 451–468.

Bernreuter, D. L., Savy, J. B., Mensing, R. W., and Chen, J. C. (1989). *Seismic hazard characterization of 69 nuclear sites east of the Rocky Mountains*. Vol. 1–8, NUREG/CR 5250, UCID-21517.

Bunn, D. W. (1975). A Bayesian approach to the linear combination of forecasts. *Operational Research Quarterly, 26*, 325–329.

Bunn, D. W. (1988). Combining forecasts. *European Journal of Operational Research, 33*, 223–229.

Bunn, D., and Wright, G. (1991). Interaction of judgmental and statistical forecasting methods: Issues and analysis. *Management Science, 37*, 501–518.

Chandrasekharan, R., Moriarty, M. M., and Wright, G. P. (1994). Testing for unreliable estimators and insignificant forecasts in combined forecasts. *Journal of Forecasting, 13*, 611–624.

Chhibber, S., and Apostolakis, G. (1993). Some approximations useful to the use of dependent information sources. *Reliability Engineering and System Safety, 42*, 67–86.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559–583.

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

Clemen, R. T., Jones, S. K., and Winkler, R. L. (1996). Aggregating forecasts: An empirical evaluation of some Bayesian methods. In D. Berry, K. Chaloner, and J. Geweke (Eds.), *Bayesian statistics and econometrics: Essays in honor of Arnold Zellner*. New York: Wiley, pp. 3–13.

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

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

Clemen, R. T., and Winkler, R. L. (1987). Calibrating and combining precipitation probability forecasts. In R. Viertl (Ed.), *Probability and Bayesian statistics*. New York: Plenum, pp. 97–110.

Clemen, R. T., and Winkler, R. L. (1990). Unanimity and compromise among probability forecasters. *Management Science, 36*, 767–779.

Clemen, R. T., and Winkler, R. L. (1993). Aggregating point estimates: A flexible modeling approach. *Management Science, 39*, 501–515.

Clemen, R. T., and Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*, 187–203.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. New York: Oxford University Press.

Dall'Aglio, G., Kotz, S., and Salinetti, G. (1991). *Advances in probability distributions with given marginals: Beyond the copulas*. Dordrecht, The Netherlands: Kluwer.

Davis, J. (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples, 1950–1990. *Organizational Behavior and Human Decision Processes, 52*, 3–38.

Dawes, R. M., Faust, D., and Meehl, P. A. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1673.

de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, *7*, 1–68. Translated in 1980 by H. E. Kyburg, Jr., Foresight. Its logical laws, its subjective sources. In H. E. Kyburg, Jr. and H. E. Smokler (Eds.), *Studies in subjective probability* (2nd ed.). Huntington, New York: Robert E. Krieger, pp. 53–118.

Dillon, R., John, R., and von Winterfeldt, D. (2002). Assessment of cost uncertainties for large technology projects: A methodology and an application. *Interfaces, 32* (Jul/Aug), 52–66.

EPRI (1986). *Seismic hazard methodology for the central and eastern United States. Vol. 1: Methodology*. NP-4/26. Palo Alto, CA: Electric Power Research Institute.

Flores, B. E., and White, E. M. (1989). Subjective vs. objective combining of forecasts: An experiment. *Journal of Forecasting, 8*, 331–341.

French, S. (1981). Consensus of opinion. *European Journal of Operational Research, 7*, 332–340.

French, S. (1985). Group consensus probability distributions: A critical survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian statistics 2*. Amsterdam: North-Holland, pp. 183–197.

French, S., and Ríos Insua, D. (2000). *Statistical decision theory*. London: Arnold.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting prior distributions. *Journal of the American Statistical Association, 100*, 680–700.

Gelfand, A. E., Mallick, B. K., and Dey, D. K. (1995). Modeling expert opinion rising as a partial probabilistic specification. *Journal of the American Statistical Association, 90*, 598–604.

Genest, C. (1984). Pooling operators with the marginalization property. *Canadian Journal of Statistics, 12*, 153–163.

Genest, C., and McConway, K. J. (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting, 9*, 53–73.

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

Genest, C., and Zidek, J. V. (1986). Combining probability distributions: A critique and annotated bibliography. *Statistical Science, 1*, 114–148.

Gigone, D., and Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*, 149–167.

Gokhale, D. V., and Press, S. J. (1982). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *Journal of the Royal Statistical Society, Series A, 145*, 237–249.

Goodman, B. (1972). Action selection and likelihood estimation by individuals and groups. *Organizational Behavior and Human Performance, 7*, 121–141.

Hammitt, J. K., and Shlyakhter, A. I. (1999). The expected value of information and the probability of surprise. *Risk Analysis, 19*, 135–152.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman and G. Owen (Eds.), *Information pooling and group decision making*. Greenwich, CT: JAI Press, pp. 129–157.

Hill, G. W. (1982). Group vs. individual performance: Are $N + 1$ heads better than one? *Psychological Bulletin, 91*, 517–539.

Hogarth, R. M. (1987). *Judgment and choice*. (2nd ed.). Chichester, England: Wiley.

Hora, S. C. (1992). Acquisition of expert judgment: Examples from risk assessment. *Journal of Energy Engineering, 118*, 136–148.

Hora, S. C. (2004). Probability judgments for continuous quantitites: Linear combinations and calibration. *Management Science, 50*, 597–604.

Hora, S. C., Dodd, N. G., and Hora, J. A. (1993). The use of decomposition in probability assessments on continuous variables. *Journal of Behavioral Decision Making, 6*, 133–147.

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

Jouini, M. N., and Clemen, R. T. (1996). Copula models for aggregating expert opinions. *Operations Research, 44*, 444–457.

Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

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

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

Lacke, C. J., and Clemen, R. T. (2001). Analysis of colorectal cancer screening regimens. *Health Care Management Science, 4*, 257–267.

Larrick, R. P., and Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*, 111–127.

Lawrence, M. J., Edmundson, R. H., and O'Connor, M. J. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science, 32*, 1521–1532.

Lichtendahl, K. C. (2005). *Bayesian nonparametric combination of multivariate expert forecasts*. Ph.D. dissertation, Duke University.

Lindley, D. V. (1983). Reconciliation of probability distributions. *Operations Research, 31*, 866–880.

Lindley, D. V. (1985). Reconciliation of discrete probability distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian statistics 2*. Amsterdam: North-Holland, pp. 375–390.

Lipscomb, J., Parmigiani, G., and Hasselblad, V. (1998). Combining expert judgment by hierarchical modeling: An application to physician staffing. *Management Science, 44*, 149–161.

MacGregor, D., Lichtenstein, S., and Slovic, P. (1988). Structuring knowledge retrieval: An analysis of decomposing quantitative judgments. *Organizational Behavior and Human Decision Processes, 42*, 303–323.

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

Merrick, J. R. W., van Dorp, J. R., and Singh, A. (2005). Analysis of correlated expert judgments from extended pairwise comparisons. *Decision Analysis, 2*, 17–29.

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

Morgan, M. G., and Keith, D. W. (1995). Subjective judgments by climate experts. *Envirionmental Science and Technology, 29*, 468–476.

Morris, P. A. (1974). Decision analysis expert use. *Management Science, 20*, 1233–1241.

Morris, P. A. (1977). Combining expert judgments: A Bayesian approach. *Management Science, 23*, 679–693.

Morris, P. A. (1983). An axiomatic approach to expert resolution. *Management Science, 29*, 24–32.

Mosleh, A., Bier, V. M., and Apostolakis, G. (1987). A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering and System Safety, 20*, 63–85.

Newbold, P., and Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society, Series A, 137*, 131–149.

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

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

Reagan-Cirincione, P. (1994). Improving the accuracy of group judgment; A process intervention combining group facilitation, social judgment analysis, and information technology. *Organizational Behavior and Human Decision Processes, 58*, 246–270.

Rohrbaugh, J. (1979). Improving the quality of group judgment: Social judgment analysis and the Delphi technique. *Organizational Behavior and Human Performance, 24*, 73–92.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Schmittlein, D. C., Kim, J., and Morrison, D. G. (1990). Combining forecasts: Operational adjustments to theoretically optimal rules. *Management Science, 36*, 1044–1056.

Seaver, D. A. (1978). *Assessing probability with multiple individuals: Group interaction versus mathematical aggregation*. Report No. 78–3, Social Science Research Institute, University of Southern California.

Shlyakhter, A. I. (1994). Improved framework for uncertainty analysis: Accounting for unsuspected errors. *Risk Analysis, 14*, 441–447.

Shlyakhter, A. I., Kammen, D. M., Brodio, C. L., and Wilson, R. (1994). Quantifying the credibility of energy projections from trends in past data: The U. S. energy sector. *Energy Policy, 22*, 119–130.

Sniezek, J. A. (1989). An examination of group process in judgmental forecasting. *International Journal of Forecasting, 5*, 171–178.

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

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

Stone, M. (1961). The opinion pool. *Annals of Mathematical Statistics, 32*, 1339–1342.

West, M. (1988). Modelling expert opinion. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian statistics, 3*, Amsterdam: New Holland, pp. 493–508.

West, M. (1992). Modelling agent forecast distributions. *Journal of the Royal Statistical Society, Series B, 54*, 553–567.

West, M., and Crosse, J. (1992). Modelling probabilistic agent opinion. *Journal of the Royal Statistical Society, Series B, 54*, 285–299.

Winkler, R. L. (1968). The consensus of subjective probability distributions. *Management Science, 15*, 361–375.

Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science, 27*, 479–488.

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

Winkler, R. L., and Clemen, R. T. (1992). Sensitivity of weights in combining forecasts. *Operations Research, 40*, 609–614.

Winkler, R. L., and Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis, 1*, 167–176.

Winkler, R. L., and Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society, Series A, 146*, 150–157.

Winkler, R. L., and Poses, R. M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science, 39*, 1526–1543.

Winkler, R. L., Wallsten, T. S., Whitfield, R. G., Richmond, H. M., Hayes, S. R., and Rosenbaum, A. S. (1995). An assessment of the risk of chronic lung injury attributable to long-term ozone exposure. *Operations Research, 43*, 19–28.

Wright, G., and Ayton, P. (Eds). (1987). *Judgmental forecasting*. Chichester, UK: Wiley.

Wright, G., Saunders, C., and Ayton, P. (1988). The consistency, coherence, and calibration of holistic, decomposed, and recomposed judgmental probability forecasts. *Journal of Forecasting, 7*, 185–199.

Wu, J. S., Apostolakis, G., and Okrent, D. (1990). Uncertainties in system analysis: Probabilistic versus nonprobabilistic theories. *Reliability Engineering and System Safety, 30*, 163–181.