Decision Analysis

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Median Aggregation of Distribution Functions

Stephen C. Hora, Benjamin R. Fransen, Natasha Hawkins, Irving Susel

To cite this article:
Stephen C. Hora, Benjamin R. Fransen, Natasha Hawkins, Irving Susel (2013) Median Aggregation of Distribution Functions. Decision Analysis 10(4):279-291. http://dx.doi.org/10.1287/deca.2013.0282

DECISION ANALYSIS
Vol. 10, No. 4, December 2013, pp. 279–291
ISSN 1545-8490 (print) | ISSN 1545-8504 (online)

http://dx.doi.org/10.1287/deca.2013.0282

Median Aggregation of Distribution Functions

Stephen C. Hora
Center for Risk and Economic Analysis of Terrorism Events, University of Southern California, Los Angeles, California 90089, hora@usc.edu

Benjamin R. Fransen
Office of Infrastructure and Protection, National Protection and Programs Directorate, Department of Homeland Security, Washington, DC 20528, benjamin.fransen@dhs.gov

Natasha Hawkins, Irving Susel
Office of Strategy, Planning, Analysis and Risk, Office of Policy, Department of Homeland Security, Washington, DC 20528 {natasha.l.hawkins@hq.dhs.gov, irving.susel@dhs.gov}

When multiple redundant probabilistic judgments are obtained from subject matter experts, it is common practice to aggregate their differing views into a single probability or distribution. Although many methods have been proposed for mathematical aggregation, no single procedure has gained universal acceptance. The most widely used procedure is simple arithmetic averaging, which has both desirable and undesirable properties. Here we propose an alternative for aggregating distribution functions that is based on the median cumulative probabilities at fixed values of the variable. It is shown that aggregating cumulative probabilities by medians is equivalent, under certain conditions, to aggregating quantiles. Moreover, the median aggregate has better calibration than mean aggregation of probabilities when the experts are independent and well calibrated and produces sharper aggregate distributions for well-calibrated and independent experts when they report a common location-scale distribution. We also compare median aggregation to mean aggregation of quantiles.

Key words: expert judgment; calibration; scoring rules; terrorists; expert combination
History: Received on April 16, 2013. Accepted by Editor-in-Chief Rakesh Sarin on August 24, 2013, after 1 revision.

Aggregation of Probability Functions

Often during the construction of risk and decision analyses it is beneficial to employ multiple subject matter experts to provide information about uncertain quantities. This apparent redundancy may be beneficial for two reasons:

1. Different experts may approach a problem from various viewpoints or using different tools, data sources, etc. The differences between judgments provides a better understanding of the uncertainty about the target quantities.

2. When properly aggregated into a single representation of uncertainty, the aggregate may more accurately reflect the information and uncertainty about the target quantity.

This second reason assumes that multiple judgments are somehow aggregated into a single expression of uncertainty. Such as, aggregation may be done

1. to improve accuracy as stated in 2, above;

2. so that a single uncertainty distribution is available to propagate uncertainty through models;

3. as a representation of consensus;

4. to avoid the potential for presenting conflicting results from multiple experts.

Aggregation methods are roughly categorized as behavioral and mathematical. Behavioral methods involve some type of negotiation such as a group coming to a consensus or a group leader or facilitator helping the group reach a satisfactory representation of their collective judgment (Kaplan 1992). Some behavioral methods intentionally avoid face-to-face interaction such as in the Delphi method (Dalkey 1967). Discussions of behavioral methods can be found in Wright and Ayton (1987), Clemen and Winkler (1999), and Armstrong (2001).

In contrast, mathematical methods impose an unambiguous rule. Various approaches have been taken in developing these rules. The simplest rule,

and perhaps the one most widely used, is to take an unweighted average of probabilities, probability densities, or distribution functions. Stone (1961) named this method the linear opinion pool although it dates to Laplace (Clemen and Winkler 2007). Lichtendahl et al. (2013) examined averaging quantiles of continuous distributions given by multiple experts rather than averaging probabilities. These authors demonstrate both properties and performance of the rule. We will refer to the linear opinion pool as the mean probability aggregate and the quantile average as the mean quantile aggregate.

Various other approaches have been taken in developing aggregation rules. One approach is to begin with axioms or properties that are desirable and derive rules consistent with the axioms and properties. Genest and Zidek (1986) present an excellent discussion of axioms and the types of rules that result. Too many assumptions are not a good thing however, as Dalky (1972) has shown in his impossibility theorem. Another approach is to examine the performance of rules by comparing the aggregated probabilities to realizations of the values or events that the probabilities forecast. This may be accomplished retrospectively using empirical data (Lichtendahl et al. 2013) or through methods of mathematical statistics (Hora 2010). Cooke (1991) proposed and utilized a rule he calls the classical method that incorporates both features of an axiomatic approach and empirical verification of an expert's capabilities based on test questions. A third mathematical approach is based on Bayesian methods (Morris 1977, Winkler 1981).

Median Aggregation

Here we consider an aggregation based on the median distribution function, which is defined as

$$F_M(x) = \text{median}[F_1(x), F_2(x), \ldots, F_n(x)] \text{ for each } x. \quad (1)$$

The median distribution function, or median aggregate, will be unique when n is odd. When n is even we may take the median aggregate to be the average of the $n/2$ and $n/2+1$ ranked cumulative probabilities at each value of x. The left panel of Figure 1 shows the median aggregate with three distributions and the right panel shows the median aggregate with four distributions. In what follows, we show that this method of aggregation has good properties. In particular, it admits aggregation of quantiles similar to the method of Lichtendahl et al. (2013). Median aggregation also has an agreement preservation property, does not require normalization, and has good calibration and sharpness properties. Median aggregation is a special case of trimmed aggregation where all but one or two or the values are trimmed (see Jose et al. 2013).

As a preliminary, we note that medians are preserved under monotone transformations. That is, if $h(x)$ is a monotone transformation and n is odd, then

$$\text{median}[h(x_1), \ldots, h(x_n)] = h[\text{median}(x_1, \ldots, x_n)]. \quad (2)$$

Note that when $h(x)$ is not strictly monotone, several among $h(x_1), \ldots, h(x_n)$ may be equal, but one of these must be $h[\text{median}(x_1, \ldots, x_n)]$ so that (2) holds. The result (2) is not shared by the mean as one may ascertain from Jensen's inequality (Merkle 2005). We will also require the definition of a quantile. Following Severini (2005), the qth quantile of a distribution is the value x_q of the random variable X defined by

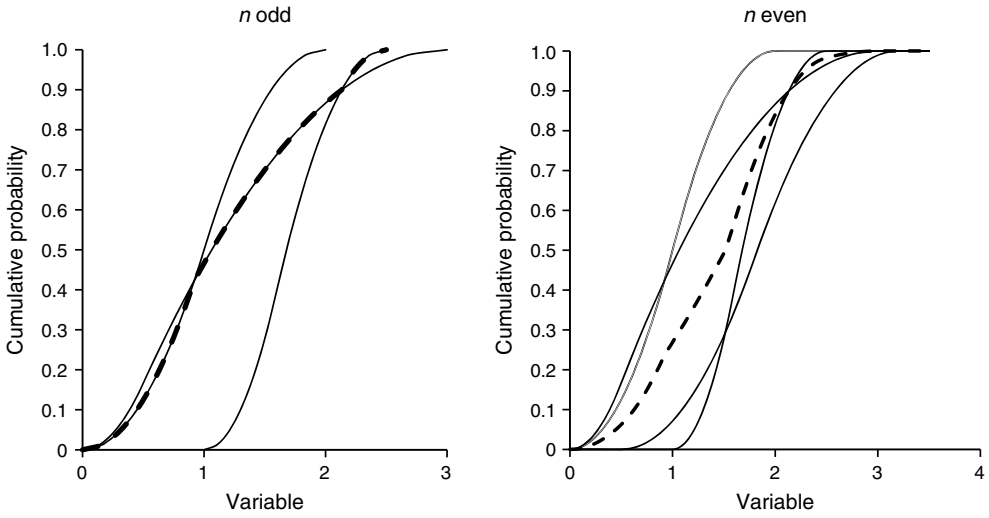
$$x_q = \inf\{x\colon P(X \le x) \ge q\}. \quad (3)$$

This definition provides unique quantiles for discrete distribution functions as well as continuous distribution functions. This function (3) is sometimes referred to as the left quantile function.

Proposition 1 (Median Quantile Aggregation). *Let $F_1(x), F_2(x), \ldots, F_n(x)$ be distribution functions with qth quantiles $x_{q1}, x_{q2}, \ldots, x_{qn}$. Then, if n is odd, $F_M(x_{qM}) = q$, where $x_{qM} = \text{median}(x_{q1}, x_{q2}, \ldots, x_{qn})$ and $F_M(x)$ is given in* (1).

This result follows immediately from the distribution function being monotone and (2). There is practical importance to this result. Assessment for continuous quantities is often conducted by assessing quantiles of distribution functions, completing the distributions by using maximum entropy, or fitting to a family of distributions and then averaging the probabilities. If one employs median aggregation, the steps of completing the distributions and aggregation can be exchanged. In some cases, such as when n is large, this may result in a less labor-intensive process.

Figure 1 Median Aggregate



Proposition 2 (Agreement). *If $F_i(x) - F_i(y) = \alpha$ for $i = 1, \ldots, n$ then $F_M(x) - F_M(y) = \alpha$ and as a special case, if $F_i(x) = \alpha$ for $i = 1, \ldots, n$ then $F_M(x) = \alpha$.*

We note that agreement is stronger than, and thus implies the zero set property of, McConway (1981), which requires that if all experts agree that an event has a zero probability, the aggregate must also have a zero probability for that event.

Agreement is demonstrated here for the continuous case knowing that the discrete case follows using the same line of reasoning. Now, agreement about the interval $[x, y]$ implies that $F_i(y) - F_i(x) = F_j(y) - F_j(x) = \alpha$ for some $0 \le \alpha \le 1$ and all $i, j = 1, \ldots, n$. Thus, $F_i(y) = F_i(x) + \alpha$, $i = 1, \ldots, n$, so that $F_i(y)$ is a monotone transformation of $F_i(x)$ for $i = 1, \ldots, n$. Then, the median preservation property shows that when n is odd, the median aggregates at x, and y are to be found on the same constituent distribution function, say the distribution function with index i, and thus $F_M(y) - F_M(x) = \alpha$. For even n, we note that the constituent distributions both agree on the probability of the interval and thus

$$\begin{aligned} &\tfrac{1}{2}[F_i(y) + F_j(y)] - \tfrac{1}{2}[F_i(x) + F_j(x)] \\ &\quad = \tfrac{1}{2}\{[F_i(y) - F_i(x)] + [F_j(y) - F_j(x)]\} \\ &\quad = \tfrac{1}{2}(2\alpha) = \alpha. \end{aligned} \tag{4}$$

Proposition 3 (Normalization). *The median aggregate is a distribution function and does not require normalization.*

This result is obvious from the construction of the median aggregate (see Figure 1). This property is shared with both the mean probability and mean quantiles aggregates but is found not to hold for multiplicative rules such as a geometric mean. Failure of this property implies that the rule meets neither the weak nor strong setwise function properties (McConway 1981). These properties are akin to "independence of irrelevant alternatives" (Genest and Zidek 1986, p. 117) properties in that they require the aggregate probability to depend only on the constituent probabilities of that event or value and not on the probabilities of other events.

Even though the median aggregate does not require normalization, it fails to have the weak setwise property (Genest and Zidek 1986) as we will demonstrate by counterexample. Table 1 contains three probability mass functions for a random variable that takes on one of the three values 1, 2, or 3. The right set of mass functions are similar to the left set with the exception that the probabilities from expert 1 are in a different order. Both sets of mass functions, however, assign the same probabilities to the value 2. When these mass functions are converted to distribution functions, aggregated using the median aggregate, and then differenced to recover the probabilities of 1, 2, and 3, one obtains 0.4, 0.4, 0.2 for the left set and 0.6, 0.3, 0.1 for the right set. Clearly, the probabilities assigned to $X = 1$ and $X = 3$ have moderated the probability of $X = 2$.

Table 1 Three Distributions

	$X=1$	$X=2$	$X=3$	$X=1$	$X=2$	$X=3$
Expert 1	0.1	0.3	0.6	0.6	0.3	0.1
Expert 2	0.6	0.2	0.2	0.6	0.2	0.2
Expert 3	0.4	0.5	0.1	0.4	0.5	0.1

Calibration refers to the faithfulness of the probabilities of values or intervals of values in the sense that they can be empirically verified, at least in a conceptual sense. For, example, a sequence of distribution functions is well calibrated for an associated sequence of realizations if intervals bounded by the pth and qth quantiles contain the realizations in a fraction $|p-q|$ of the time for all $p \neq q$. Thus, one would find 25% of the values in each of the quartiles of the distributions. We consider calibration in terms of the calibration trace that is a plot of cumulative probabilities of the realizations versus the cumulative relative frequencies of these probabilities in a set of probability forecasts. A well-calibrated set of cumulative probabilities will have a calibration trace that is a 45-degree line.

Proposition 4 (Calibration Trace). *Let $F_i(x)$ for $i=1,\ldots,n$ and n odd, be given by well-calibrated, independent experts. Then, $F_M(x)$ has a calibration trace that is a beta distribution function with parameter $[(n+1)/2, (n+1)/2]$ given by*

$$F_\beta(u \mid a,b) = \int_0^u [\beta(a,b)]^{-1} x^{a-1}(1-x)^{b-1}\,dx \quad \text{for } a>0,\ b>0,\ 0\le u\le 1 \tag{5}$$

and $\beta(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$. For even n, the calibration trace of $F_M(x)$ is given by

$$\begin{aligned} G_n(w) &= \frac{2(n-1)!}{[(n/2)!]^2}\{\beta(n/2+1,n/2)F_\beta(w\mid n/2+1,n/2) \\ &\quad + \sum_{i=0}^{n/2}\binom{n/2}{i}(2w)^i(-1)^{m/2-i}\beta(n/2-i+1,n/2) \\ &\quad \times [F_\beta(2w\mid n/2-i+1,n/2) \\ &\quad - F_\beta(w\mid n/2-i+1,n/2)]\} \qquad \text{for } 0\le w\le 1/2; \\ &= 1-G_n(1-w) \qquad \text{for } 1/2<x\le 1. \end{aligned} \tag{6}$$

The proposition is not difficult to prove. First, we recognize that the cumulative probability of the target quantity will behave as a uniform random variable when the expert provides a well-calibrated distribution function (Hora 2004). Then if the experts are both well calibrated and independent, the cumulative probabilities at the target value will behave as a sample of n uniform random variables on the unit interval and the distribution of the median of these uniform random variables will be a beta distribution function with parameter $[(n+1)/2, (n+1)/2]$ (David and Ngaraja 2003).

For the even n case, we require the joint density of the $n/2$ and $n/2+1$ order statistics from a sample of n independent uniform random variables, which is given by David and Ngaraja (2003) as

$$f_n(x,y) = \frac{n!}{[(n/2-1)!]^2}x^{(n/2)-1}(1-y)^{(n/2)-1} \quad \text{for } 0\le x\le y\le 1. \tag{7}$$

Then, for $w \le 1/2$

$$\begin{aligned} &\text{prob}[(x+y)/2\le w] \\ &= \int_0^w\int_0^y f_n(x,y)\,dx\,dy + \int_w^{2w}\int_0^{2w-y} f_n(x,y)\,dx\,dy \\ &= \frac{2(n-1)!}{[(n/2)!]^2}\{\beta(n/2+1,n/2)F_\beta(w\mid n/2+1,n/2) \\ &\quad + \sum_{i=0}^{n/2}\binom{n/2}{i}(2w)^i(-1)^{m/2-i}\beta(n/2-i+1,n/2) \\ &\quad \times [F_\beta(2w\mid n/2-i+1,n/2) \\ &\quad - F_\beta(w\mid n/2-i+1,n/2)]\}. \end{aligned} \tag{8}$$

The expansion in the last term in (8) results from a binomial expansion of the terms in $2w-y$.

For $w>1/2$,

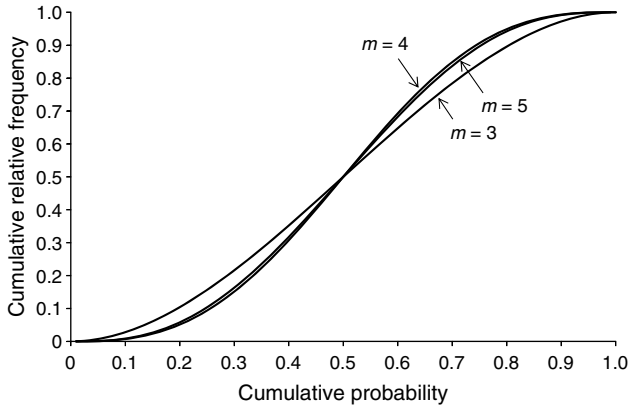
$$\text{prob}[(x+y)/2\le w] = 1-\text{prob}[(x+y)/2\le 1-w].$$

For a similar result entailing a density with support on the entire real line see Desu and Rodine (1969).

Applying the inverse of the calibration trace will generate cumulative probabilities that behave as a uniform random variable on the unit interval and thus are well calibrated. This finding has two important implications. The first is insight into how median aggregation operates on calibration. The second is that it provides a method for correcting miscalibration induced by the aggregation.

When the constituent distributions are well calibrated and the experts independent, the calibration trace is similar to those shown in Figure 2. The curve

Figure 2 **Calibration Traces, $n = 3, 4$, and Five Independent Well-Calibrated Experts**



is indicative of more spread than is justified and, therefore, in the direction of underconfidence. This result is qualitatively similar to that found by Hora (2004) for mean probability aggregate and the tendency serves as a counterbalance to experts being overconfident. The impact on calibration with median aggregation is not as great as that with mean probability aggregation, however. This can be seen by examining the derivative of the calibration trace, which is, in fact, a density function. With arithmetic aggregation the variance of the density associated with the trace will be $1/(12n)$, whereas median aggregation results in a variance of $1/[4(n+2)]$, which is the variance of the beta distribution with parameter $[(n+1)/2, (n+1)/2]$ and is larger than the former variance for $n = 2, \ldots$. Thus, one would find that the cumulative probabilities of the target values will be clustered closer to 0.5 with mean probability aggregation than with median aggregation. For a further discussion of recalibrating mean aggregates, see Ranjan and Gneiting (2010). Lichtendahl et al. (2013) compare the variance of calibration traces of the mean quantile aggregate to the trace of the mean probability aggregate and conclude that when experts are well calibrated and report distributions from a location-scale family in a manner such that the reported means are symmetric about some value, the variance of the calibration trace from mean quantile aggregation will be greater than that of the trace from the mean probability aggregate suggesting better calibration.

Figure 2 depicts three calibration traces derived from (5) and (6) that correspond to the median aggregate of three, four, and five well-calibrated independent experts. Surprisingly the calibration with four experts is worse than that with five experts. We attribute this to the use of the mean of two distribution functions in the even case ($n = 4$).

It is clear from this discussion that the median aggregation produces densities that are more spread than they should be when the constituent distributions are well calibrated and independent. The reality is, however, that experts tend to provide distributions that are too narrow and thus, as with the linear opinion pool, the aggregation tends to counterbalance overconfidence. Here, we have a case of two wrongs that may make a right.

But what about the impact of dependence? We will demonstrate that the impact of increasing the dependence among experts on calibration is to decrease the tendency toward underconfidence. We employ a copula and again assume that the experts are well calibrated so that the cumulative probabilities of the target value behave as n dependent uniform $[0, 1]$ random variables. The chosen copula is the Clayton (1978) copula given by

$$c(u_1, \ldots, u_n \mid \theta) = \left[\sum_{i=1}^{n} u_i^{-\theta} - 1\right]^{-1/\theta} \quad \text{for } 0 < \theta, \qquad (9)$$

which induces positive dependence in the sense that Kendall's tau is $\tau = \theta/(\theta+1)$.

The median of the cumulative probabilities at the target value will then have the same distribution as the median of n dependent uniform random variables (David and Ngaraja 2003) and assuming equal correlation among the experts, the calibration trace is given by

$$H(u) = \sum_{m=(n+1)/2}^{n} (-1)^{m-n/2} \binom{m-1}{n/2-1} \binom{n}{m} \cdot \Pr(U_{m:m} \le u), \qquad (10)$$

where $U_{m:m}$ is the largest value in a dependent uniform sample of size m and, assuming the Clayton copula, $\Pr(U_{m:m} \le u)$ is given by $\Pr(U_{m:m} \le u) = c(u, \ldots, u \mid \theta) = [mu^{-\theta} - (m-1)]^{-1/\theta}$ so that

$$H(u) = \sum_{m=(n+1)/2}^{n} (-1)^{m-n/2} \binom{m-1}{n/2-1} \binom{n}{m} \cdot [mu^{-\theta} - (m-1)]^{-1/\theta}. \qquad (11)$$

Figure 3 Calibration Traces with Dependent Experts

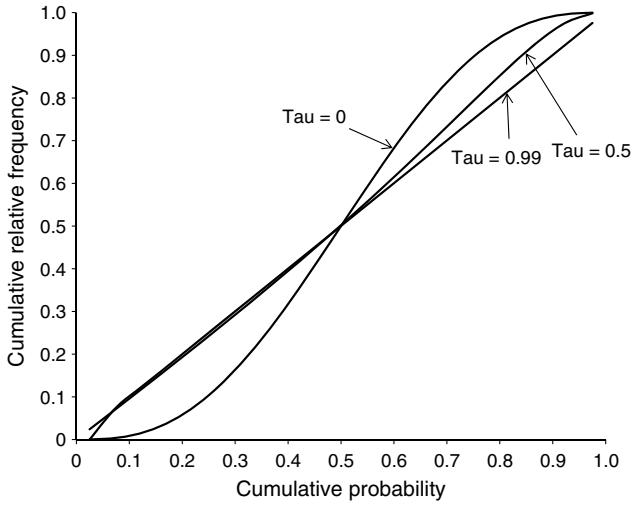


Figure 3 shows calibration traces for the median aggregate of five well-calibrated experts where the mutual dependency is varied so that $\tau = 0, 0.5, 0.99$. Recalling that perfect calibration results in a 45-degree calibration trace, the tendency toward underconfidence is seen to be largely done away with $\tau = 0.5$.

Proposition 5 (Sharpness). *Given that* $F_1(x), \ldots, F_n(x)$ *are from a common location-scale family of distributions with a common variance,* σ^2*, and* n *is odd, the variance of* $F_M(x)$ *is* σ^2.

Although calibration is an important property of an aggregation rule, it is the degree of sharpness that measures the information in the aggregate. A well-calibrated set of distributions that are very diffuse tell us little about the value of the variable. Gneiting et al. (2007) suggest that an appropriate approach for dealing with calibration and sharpness is to insist on calibration and then maximize sharpness subject to maintaining calibration. A different viewpoint is argued by Winkler (2009) who suggests that sharpness is a more important consideration as calibration can be improved by training, whereas improvements in sharpness require improved knowledge, which is more difficult to obtain.

Here, we examine how the sharpness of individual judgments are transformed by median aggregation and compare the results to similar results with mean probability and mean quantile aggregation. We employ the paradigm of Hora (2010) who shows how one might generate sequences of well-calibrated distributions. Specifically, we employ a family of location-spread distributions:

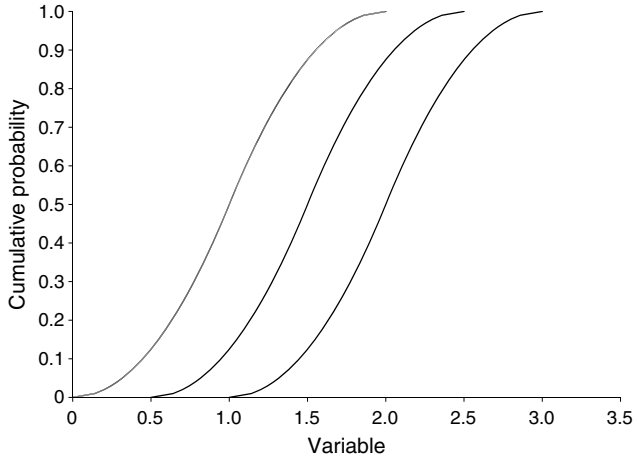
$$F(x \mid \mu, \sigma) = F\left(\frac{x-\mu}{\sigma} \,\middle|\, 0, 1\right) = P(X \le x), \qquad (12)$$

and generate n sequences of distributions so that each sequence is well calibrated. We assume that the variance exists for this family and that it is equal to σ^2. The sequence is conceptually generated by assuming a value for the target quantity, say x, and associating each location parameter with a uniform random variable so that $\mu_{ij} = x + \sigma F^{-1}(u_{ij})$ for $i = 1, \ldots, n$, $j = 1, 2, \ldots$ where $F^{-1}(u)$ is the inverse of $F(x)$. Following Hora (2010), we take $x = 0$ in without loss of generality as it is the difference between x and the mean that determines the value and not the specific value of x.

In this simple setup, the variances of the distributions in each sequence are constant and equal across sequences. The median aggregate distribution will be the distribution among $\{F_1(x), \ldots, F_m(x)\}$ that has the location parameter that is the median of $\{\mu_1, \ldots, \mu_m\}$. Thus, the median aggregate will have variance σ^2. This result becomes obvious when one examines the distribution functions in Figure 4 where it is apparent that one of the distribution functions will be the median aggregate as the distribution functions do not cross. Arithmetic aggregation of quantiles, for different reasons, has the same variance, σ^2. This can be seen by noting that the quantile function is $Q_i(p) = \mu_i + \sigma\Phi^{-1}(p)$ and thus the mean quantile function is $\bar{Q}(p) = \bar{\mu} + \sigma\Phi^{-1}(p)$ so that the inverse of the aggregated quantile function is a cumulative distribution function (CDF) with variance σ^2 as with the median aggregate.

Hora (2010) shows that for the case examined here, the variance of the mean probability aggregate is $(2n-1)\sigma^2/n$. This variance is an increasing function of n, the number of experts, which is a somewhat perverse result as the expressed uncertainty increases with the number of experts—more expertise results in more uncertainty. For $n > 1$, the variance of the median aggregate is always less than that of the arithmetic average and approaches ½ relative to the variance of the mean probability aggregate as n grows large indicating that the spread with mean probability aggregation will be $\sqrt{2}$ times that with median aggregation.

Figure 4 Densities from the Same Family with Equal Variances



As calibration appears to be better with median aggregation than mean probability aggregation and sharpness also appears to be better, at least in the simple case examined here, one would anticipate that median aggregation would also perform better in terms of proper scoring rules. This is indeed the case. Specifically, we examine the case described in the previous section, assume the normal family of distributions, and calculate the expected Brier score (Brier 1950). The Brier score is given by

$$\mathrm{BS}(x, f) = 2f(x) - \int_{-\infty}^{\infty} f^2(x)\,dx, \quad \text{where } f(x) = dF(x)/dx, \tag{13}$$

where higher scores are preferred.

Proposition 6 (Expected Brier Score). *Given that $F_1(x), \ldots, F_n(x)$ are normal distribution functions sharing a common variance, σ^2, and have means $\mu_1, \ldots, \mu_n$ independently drawn from a normal distribution with mean equal to the target quantity and variance σ^2, the expected Brier score of the median aggregate for large n is*

$$E_{F_M}\,\mathrm{BS}[0, dF_M(x)/dx] \approx \frac{\sqrt{2}}{\sigma\sqrt{\pi}\sqrt{1+\pi/(2n)}} - \frac{1}{2\sigma\sqrt{\pi}} = \frac{1}{\sigma\sqrt{\pi}}\left[\frac{\sqrt{2}}{\sqrt{1+\pi/(2n)}} - \frac{1}{2}\right]. \tag{14}$$

Let $f_N(x \mid \mu, \sigma) = (1/(\sqrt{2\pi}\sigma))e^{-(1/2)[(x-\mu)/\sigma]^2}$ denote the normal density. The following result will be used to simplify the calculations:

$$\int_{-\infty}^{\infty} f_N(x \mid 0, \sigma) f_N(x \mid 0, s)\,dx = \frac{1}{\sqrt{2\pi(\sigma^2+s^2)}}. \tag{15}$$

For the first term in the Brier score, $2f(x)$, we have

$$E[2dF_M(x)/dx|_{x=0}] = 2E_{\mu_M}[f_N(0 \mid \mu_M, \sigma)]. \tag{16}$$

The expectation in (16) is found in terms of expected normal order statistics using the exact distribution of the median (David and Ngaraja 2003). However, the result is not in a closed form as the expected normal order statistics do not have a closed form. One can, however, obtain a relatively simple expression by considering the large sample distribution of the median. Thus, with large n, one has

$$\begin{aligned} &2E_{\mu_M}[f_N(0 \mid \mu_M, \sigma)] \\ &\quad = 2\int_{-\infty}^{\infty} f_N(0 \mid \mu_M, \sigma) f_N(\mu_M \mid 0, s_M)\,d\mu_M \\ &\quad = 2\int_{-\infty}^{\infty} f_N(\mu_M \mid 0, \sigma) f_N(\mu_M \mid 0, s_M)\,d\mu_M \\ &\quad = \frac{2}{\sqrt{2\pi(\sigma^2+s_M^2)}} = \frac{\sqrt{2}}{\sigma\sqrt{\pi}\sqrt{1+\pi/(2n)}}, \end{aligned} \tag{17}$$

where $s_M^2 = 1/[4nf_N^2(0 \mid 0, \sigma)] = \pi\sigma^2/(2n)$ is found from the large sample variance of the sample median (Severini 2005). The second term in (13) is

$$\int_{-\infty}^{\infty} f_N(x \mid \mu, \sigma) f_N(x \mid \mu, \sigma)\,dx = \int_{-\infty}^{\infty} f_N(y \mid 0, \sigma) f_N(y \mid 0, \sigma)\,dy = \frac{1}{2\sigma\sqrt{\pi}}. \tag{18}$$

Subtracting the second term from the first gives

$$E_{F_M}\,\mathrm{BS}[0, dF_m(x)/dx] = \frac{\sqrt{2}}{\sigma\sqrt{\pi}\sqrt{1+\pi/(2n)}} - \frac{1}{2\sigma\sqrt{\pi}} = \frac{1}{\sigma\sqrt{\pi}}\left[\frac{\sqrt{2}}{\sqrt{1+\pi/(2n)}} - \frac{1}{2}\right]. \tag{19}$$

Hora (2010) gives a similar result for the mean probability aggregate:

$$E_{\mu_1, \ldots, \mu_n}\,\mathrm{BS}\left(0, \frac{1}{n}\sum_{i=1}^{n} f_N(0 \mid \mu_i, \sigma)\right) = \frac{1}{\sigma\sqrt{\pi}}\left(1 - \frac{1}{2n} - \frac{n-1}{2n}\sqrt{\frac{1}{2}}\right). \tag{20}$$

Table 2 Evaluation of the Expected Brier Score

	Median aggregation			Mean probability aggregation	Mean quantile aggregation
Experts	Numerical	Normal approximation	Monte Carlo		
1	0.28209	0.21553	0.28039	0.28209	0.28209
2		0.31504	0.36675	0.32341	0.36937
3	0.38176	0.36431	0.38032	0.33718	0.40889
4		0.39401	0.41763	0.34406	0.43155
5	0.42207	0.41392	0.42185	0.34819	0.44627
6		0.42821	0.44214	0.35095	0.45660
7	0.44386	0.43898	0.44294	0.35292	0.46426
8		0.44738	0.45587	0.35439	0.47016
9	0.45749	0.45412	0.45613	0.35554	0.47484
10		0.45966	0.46532	0.35646	0.47866
Inf.		0.51579		0.36472	0.51579

For the mean quantile aggregate, we note that the quantile function for the ith expert is given by $Q(p) = \mu_i + \sigma\Phi^{-1}(p)$, where μ_i and σ^2 are the mean and variance of the distribution, respectively, and $\Phi^{-1}(p)$ is the inverse of the standard normal distribution function. Averaging these quantile functions gives $\bar{Q}(p) = \bar{\mu} + \sigma\Phi^{-1}(p)$, where $\bar{\mu}$ is the average of the means. For the experts to be providing well-calibrated quantiles, the variance of the means μ_i must also be σ^2 so that $\bar{\mu}$ is normal with variance σ^2/n. This result may then be used to evaluate the expected Brier score yielding

$$E_{\bar{\mu}}\,\mathrm{BS}(0, f_N(0 \mid \bar{\mu}, \sigma)) = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\sigma^2 + \sigma^2/n}} - \frac{1}{2\sqrt{\pi}\sigma}. \tag{21}$$

Table 2 contains the expected Brier scores for both an odd and even number of experts. The columns correspond to results obtained by numerical integration of (16) for odd n the normal approximation in (16), and Monte Carlo simulation results with 10,000 samples of normal distributions of the indicated size and the expressions given in (19) and (20) for the expected scores mean probability and mean quantile aggregates.

There is close agreement between the numerical integration results and the Monte Carlo results as one would expect. The normal approximation does not produce close results for small n so its primary value is to obtain the limit of the expected Brier score. As n grows without bound, the ratio of the median and mean aggregate expected Brier scores approaches $\sqrt{2}$. Median and mean quantile aggregation share the same limiting expected Brier score.

For finite $n > 1$, mean quantile aggregation performs fractionally better than median aggregation and both perform substantially better than mean probability aggregation. In fact, median aggregation with $n = 3$ and mean quantile aggregation with $n = 2$ have higher expected scores than mean probability aggregation with an infinite number of experts. However, this case entails two strong assumptions: calibrated and independent experts; assumptions that are unlikely to hold in practice. We examine one more case in which we allow both assumptions to be relaxed. Here we again use the setting provided in Hora (2010) but introduce overconfidence and dependence. Overconfidence is injected by allowing the variance of the means to exceed the expressed variance implied by the quantile functions provided by the experts. Dependence is created by drawing the means from an equicorrelated multivariate normal density. Hora (2010) provides an analytic expression for the mean probability aggregate expected score and we use the quantile function given earlier to develop the expected Brier score for the mean quantile aggregate. It is given by the following:

$$E_{\bar{\mu}}\,\mathrm{BS}(0, f_N(0 \mid \bar{\mu}, \sigma)) = \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\sigma^2 + \sigma_{\bar{\mu}}^2}} - \frac{1}{2\sqrt{\pi}\sigma}. \tag{22}$$

Unfortunately, we do not have the tools necessary to develop an analytic expression for the expected score with median aggregation when the experts are both overconfident and dependent. Instead, Monte Carlo simulation is used to obtain estimates. The case with five experts is considered, and both the degree of overconfidence and the amount of dependence are varied. Overconfidence is inserted by using ratios of $\sqrt{\mathrm{var}(\mu)}/\sigma$ equal to 1 (no overconfidence), 2, and 3. Equicorrelations of 0, 0.2, 0.4, 0.6, and 0.8 are used. Table 3 shows the results for each of the three aggregation rules using analytic results from Hora (2010) for mean probability aggregate, (22) for the mean quantile aggregate, and 10,000 samples drawn by Monte Carlo simulation for the median aggregate.

The columns are labeled 1, 2, and 3 for the amount of overconfidence and the rows labeled with the equicorrelation values. The cells of the last sub-table indicate which aggregation method performed best for the particular combination of overconfidence and

Table 3 Expected Brier Scores with Overconfidence and Dependence

	Calibration index		
Mean prob.	1	2	3
Correlation			
0	0.348194	0.227690	0.162142
0.2	0.339562	0.225609	0.161417
0.4	0.329358	0.223393	0.160670
0.6	0.317040	0.221029	0.159900
0.8	0.301758	0.218497	0.159104
Mean quant.	1	2	3
Correlation			
0	0.446271	0.312613	0.194732
0.2	0.402086	0.228698	0.105392
0.4	0.365075	0.172542	0.052690
0.6	0.333487	0.131589	0.016925
0.8	0.306114	0.100023	−0.00938
Median	1	2	3
Correlation			
0	0.420317	0.261299	0.140664
0.2	0.413784	0.250308	0.130780
0.4	0.392220	0.214240	0.095525
0.6	0.359652	0.166807	0.045592
0.8	0.322114	0.121082	0.003980
	1	2	3
Correlation			
0	Q	Q	Q
0.2	M	M	P
0.4	M	P	P
0.6	M	P	P
0.8	M	P	P

dependency. Here P stands for the mean probability aggregate, Q for the mean quantile aggregate, and M for the median aggregate. There are some conclusions that can be drawn. Median aggregation is best with dependent but well-calibrated experts. In contrast, mean quantile aggregation is best with independent experts regardless of overconfidence. Mean probability aggregation appears to be robust when both overconfidence and dependence are present although it does not fare so well when either is absent.

It is useful to inquire where in this table one would find "typical" circumstances. Experience from working with experts (Hora et al. 1992) and examination of published results of experiments (see Cooke 1991, Table 4.1) indicate that realized values frequently appear in extreme tails about one-third of the time rather than the 2% or 10% that they should. This suggests that experts provide distributions that are too tight and have standard deviations about one-half of what they should be. This translates to an overconfidence index of 2. Recent unpublished experiments conducted by the author involving forecasts of football contests have turned up correlations among experts that are close to 0.5. If one uses these two values, 2 and 0.5 to enter the table, one concludes that mean probability aggregation would have a slight advantage over median aggregation, which, in turn, provides modest improvement over mean quartile aggregation.

Recalibration

Recall that the median aggregate has a known calibration trace when experts are well calibrated and independent. It is then possible to recalibrate the cumulative probabilities by applying the calibration trace to the distribution function produced by the aggregate. Thus, for odd n,

$$F_M^*(x) = F_\beta[F_M(x) \mid (n+1)/2, (n+1)/2] \qquad (23)$$

will be a well-calibrated distribution function when the experts are individually well calibrated and independent and n is odd. To see the impact of recalibration on the expected Brier score, we require the density associated with (23), which is obtained by application of the chain rule for differentiation of composite functions:

$$\frac{dF_M^*(x)}{dx} = f_\beta[F_M(x) \mid (n+1)/2, (n+1)/2]\frac{dF_M(x)}{dx}. \qquad (24)$$

Returning to the setup used earlier in this section, where an odd number of well-calibrated independent experts provide normal distributions with equal variances, we numerically calculate the expected Brier scores using Simpson's rule with 1,001 points. In Table 4 the calculated expected Brier scores are given and compared to expected Brier scores using a multiplicative model that simply multiplies the expert densities and normalizes the result to produce a density. The results for the multiplicative model are taken from Hora (2010). This model was chosen for comparison as it performs well under these idealized circumstances. The performance of the multiplicative rule, however, is very sensitive to the assumptions of calibrated and independent experts and is not a good choice for practical application.

Table 4 Expected Brier Scores

Experts	Recalibrated median	Multiplicative
1	0.2821	0.2821
3	0.4221	0.4886
5	0.5278	0.6308
7	0.6160	0.7464
9	0.6932	0.8463
11	0.7627	0.9356
101	2.2672	2.8350
501	5.0403	6.3141
1,001	7.1228	8.9251

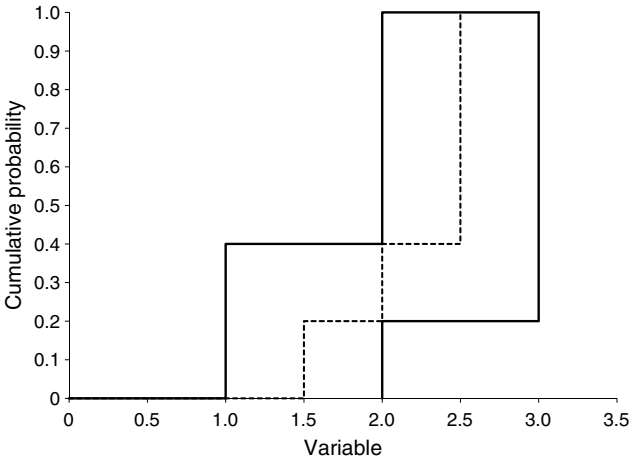
The performance of the recalibrated density is superior to the density derived from the median aggregate without recalibration. The score of the uncalibrated density becomes asymptotic to 0.5158 (see Table 2) and the calibrated counterpart appears to grow without bound as n increases. Although the recalibrated density does not fare as well as the density of the multiplicative aggregate, it performs well compared to the mean probability aggregate and the mean quantile aggregate (compare to Table 2), particularly with a large number of experts.

Discrete Distributions

Unfortunately, the median aggregation of quantiles can provide unacceptable results when applied to discrete distribution functions when n is even. This is also true of the mean quantile aggregate for any $n > 1$ regardless of whether n is odd or even. Consider two probability mass functions, one that places lumps of probability of 0.4 and 0.6 at 1 and 2, respectively, and another that places lumps of probability of 0.2 and 0.8 at 2 and 3, respectively. The CDFs are shown in Figure 5 along with the dotted line that is both the median quantile aggregate and the mean quantile aggregate. Positive probability results at both 1.5 and 2.5, even though both of these values were given zero probability and thus the zero set property is violated. A similar problem can exist when multimodal continuous distributions are aggregated. With an odd n, however, the median aggregate CDF will always coincide with one or the other of the assessed CDFs at each value so that the problem does not occur.

A word of caution is warranted regarding the aggregation of probability mass functions rather than cumulative distribution functions. Median aggregation of probability mass functions with more than two outcomes may easily result in probabilities that do not sum to one and is not advocated here.

Figure 5 Median and Mean Quantile Aggregates



Application Setting

An application is now presented that provided the motivation for this research. The Department of Homeland Security (DHS) addresses risks to the nation arising from such threats as terrorism, natural and manmade disasters, cyber attacks, and transnational crime. There are many DHS programs whose mission it is to protect the nation from these threats, with activities covering prevention, detection, interdiction, protection, response, and recovery. To measure the effectiveness of these programs, DHS employs formal elicitation of subject matter experts (SME) when data are too sparse or conflicting or when evidence from various sources must be integrated into a coherent understanding of what is known and what is uncertain. Given this reliance on SME judgments, DHS strives to continuously improve the quality and transparency of the elicitation process to enhance the usefulness of SME judgments and the confidence in them. It is important to select appropriate techniques for aggregating judgments of multiple SMEs so as to represent the overall sense of the SMEs by correctly representing the range of uncertainty provided by the experts.

As a step toward improving its expert judgment processes, a proof of concept study was undertaken by DHS to improve SME judgments of program effectiveness. Program effectiveness refers to the ability of

a particular activity to successfully complete its mission. At an airport, for example, luggage screening is conducted to detect concealed explosives. The activity is effective if it can successfully detect and interdict the contraband. However, effectiveness may depend on a sequence of events such as initial screening success, selection for secondary screening, etc. Additionally, a number of factors or conditions such as type of explosive device, the material the luggage is made from, or even the destination of the luggage, should some destinations receive more intensive screening, may moderate the effectiveness.

The assessment of overall program effectiveness was accomplished by first decomposing effectiveness into relevant steps and then conditioning assessment for each step on a number of factors that moderated effectiveness. The elicitation of judgments was conducted for all relevant combinations of these factors. Three SMEs were used for the assessment of program effectiveness with SMEs sometimes providing assessments for several activities or steps in an activity for which they were cognizant. As part of this effort, a new method of aggregating judgments based on the median of assessed quantiles was developed. An analysis of the properties of this method made subsequent to proof of concept study were detailed earlier in this article.

In partnership with the United States Coast Guard (USCG), the DHS Office of Strategy, Planning, Analysis, and Risk (SPAR) designed and executed the study to increase confidence in and utility of SME judgments. The work focused on estimating the effectiveness of the USCG in detecting terrorist teams that attempt to enter the United States by sea at port. Although other DHS components, such as immigration and customs enforcement and customs and border protection, also contribute directly toward detecting such terrorist teams, these were not included in this pilot effort because of time and resource constraints. However, focusing on the USCG alone was sufficient for the purposes of evaluating the approach.

The methodology is designed to simplify the judgments that SMEs must make to better harness their expertise and knowledge, as well as make their judgments more transparent. With this approach, SMEs do not directly estimate effectiveness—the "target estimate." Instead, the approach starts by SPAR working with SMEs to decompose the target estimate into its key steps. The SMEs then make narrower and more well-defined judgments for each step conditional on the different combinations of underlying factors, in a facilitated process. To compute the target estimate, SPAR then applies a model utilizing the SMEs more granular judgments. In the demonstration project, the SMEs provided judgments for 24 combinations of underlying factors. Even though the number of estimates needed was much greater with the decomposition modeling, all of the SMEs strongly preferred this approach because the judgments needed were much easier to make and document, and took less time as well.

For each of the 24 combinations of factors, each SME provided a low estimate (5th percentile), a high estimate (95th percentile), and a most likely value for effectiveness at each of the steps. The most likely value, the mode, was used because it is the easiest central value for the SMEs to conceptualize. For the purposes of aggregation, the median can be calculated from the mode and the two assessed quantiles. Because the demonstration study provided illuminating results, the resulting reports have been classified as sensitive security information and are not available to the public. We will demonstrate the aggregation process, however, with a simple, synthetic example in the absence of the actual study data.

We begin with three-point assessments that are the 0.05 and 0.95 quantiles and the mode. These values are used to compute the medians assuming triangular densities. Aggregation by the median method is performed at both directly assessed quantiles and at the computed medians. Next, the aggregated density is completed by assuming a triangular density fit to the aggregated quantiles. Table 5 shows the hypothetical assessments, the computed median, and the aggregated quantiles, and the mode inferred from the aggregated quantiles assuming a triangular density. Figure 6 shows CDFs formed by assuming triangular densities for each of the SMEs assessments, the median CDF (heavy dashed line) computed from the completed CDFs, and the CDF formed by aggregating the assessed quantiles and then fitting a triangular density to the aggregated quantiles (heavy solid line). The difference between the two aggregates (completing the density/CDF then aggregating

Figure 6 **Example CDFs from Triangular Densities, Median Aggregate, and Approximation**

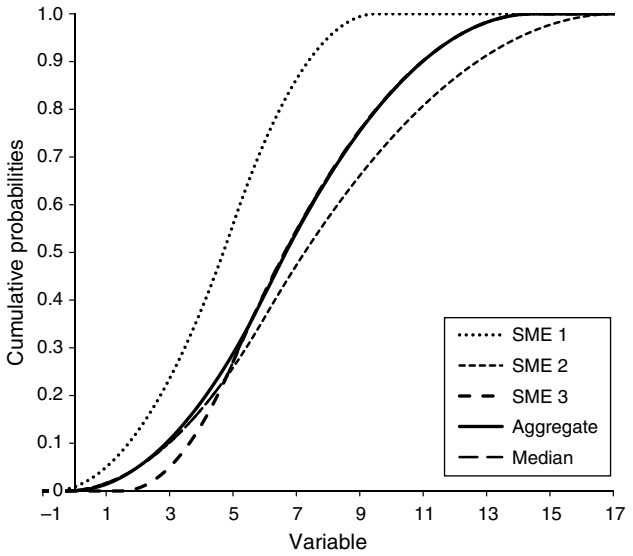


Table 5 **Assessment and Aggregation Example**

SME	0.05	Mode	0.95	Median (computed)
1	1	5	8	4.71
2	2	6	14	7.10
3	3	5	12	6.26
Aggregate	2	6.1433 (computed)	12	6.634

versus aggregating the quantiles then completing the density/CDF) is small as one would expect. The two aggregates share the same 0.05, 0.50, and 0.95 quantiles and are visually indistinguishable everywhere except between the 0.06 and 0.33 quantiles.

Conclusions

The median method of aggregation shows promise on several dimensions. SPAR was seeking a method that was expedient when the number of assessments required was large. It was also desired to have a method of aggregation that was transparent and easy to apply—one that did not require judgments about the judgments. Assessing the minimum number of quantities (3) needed to specify a triangular density and then aggregating by the median method held promise as a simple solution.

Fortunately, the properties of the method are good. Its performance holds up well when compared to mean probability aggregation and is similar to that of mean quantile aggregation. We anticipate that three-point assessments and median aggregation will become the methods of choice at SPAR when a large number of assessments must be made by each SME, when the exact shape of the uncertainty distribution can be approximated, and when extreme tail behavior does not drive the analysis. If the exact shape of the distribution is important, and there is sufficient information to derive this shape, or when interest is concentrated on the tail(s) of the distribution, then more refined analyses may be warranted. In other cases, the combination of three- point assessments and median aggregation provides sufficient resolution for policy and decision-making purposes.

Acknowledgments

This research was supported by the United States Department of Homeland Security through the National Center for Risk and Economic Analysis of Terrorism Events (CREATE) [Cooperative Agreement No. 2010-ST-061-RE0001]. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the United States Department of Homeland Security or the University of Southern California.

References

Armstrong JS, ed. (2001) *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Kluwer, Boston).

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78(1950):1–3.

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65(1):141–151.

Clemen RT, Winkler RL (1999) Combining probability distributions from experts in risk analysis. *Risk Anal.* 19(2):187–203.

Clemen RT, Winkler RL (2007) Aggregating probability distributions. Edwards W, Miles R, von Winterfeldt D, eds. *Advances in Decision Analysis* (Cambridge University Press, Cambridge, UK), 154–176.

Cooke RM (1991) *Experts in Uncertainty* (Oxford University Press, Oxford, UK).

Dalkey NC (1967) Delphi. Report, RAND Corporation, Santa Monica, CA.

Dalkey NC (1972) An impossibility theorem for group probability functions. Report, RAND Corporation, Santa Monica, CA.

David HA, Ngaraja HN (2003) *Order Statistics* (John Wiley & Sons, Hoboken, NJ).

Desu M, Rodine R (1969) Estimation of the population median. *Scandinavian Actuarial J.* 1969(1–2):67–70.

Genest C, Zidek JV (1986) Combining probability distributions: A critique and annotated bibliography. *Statist. Sci.* 1(1):114–148.

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *J. Royal Statist. Soc. B* 69(2):243–268.

Hora SC (2004) Probability judgments for continuous quantities: Linear combinations and calibration. *Management Sci.* 50(5): 597–604.

Hora SC (2010) An analytic method for evaluating the performance of aggregation rules for probability densities. *Oper. Res.* 58(5):1440–1449.

Hora SC, Hora JA, Dodd NG (1992) Assessment of probability distribution for continuous random variables: A comparison of bisection and fixed value methods. *Organ. Behav. Human Decision Processes* 51:135–155.

Jose VRR, Grushka-Cockayne Y, Lichtendahl KC Jr (2013) Trimmed opinion pools and the crowd's calibration problem. *Management Sci.* Forthcoming.

Kaplan S (1992) Expert information versus expert opinions. Another approach to the problem of eliciting/combining/using expert knowledge in PRA. *J. Reliability Engrg. System Safety* 35(1):61–72.

Lichtendahl KC Jr, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles? *Management Sci.* Forthcoming.

McConway KJ (1981) Marginalization and linear opinion pools. *J. Amer. Statist. Assoc.* 76:410–414.

Merkle M (2005) Jensen's inequality for medians. *Statist. Probab. Lett.* 71(3):277–281.

Morris PA (1977) Combining expert judgments: A Bayesian approach. *Management Sci.* 23(7):679–693.

Ranjan R, Gneiting T (2010) Combining probability forecasts. *J. Royal Statist. Soc. A* 72(1):71–91.

Severini TA (2005) *Elements of Distribution Theory* (Cambridge University Press, Cambridge, UK).

Stone M (1961) The linear opinion pool. *Ann. Math. Statist.* 32:1339–1342.

Winkler RL (1981) Combining probability distributions from dependent information sources. *Management Sci.* 27(4):479–488.

Winkler RL (2009) Calibration. Kattan MW, ed. *Encyclopedia of Medical Decision Making*, Vol. 1 (SAGE Publications, Thousand Oaks, CA), 106–109.

Wright B, Ayton P, eds. (1987) *Judgemental Forecasting* (Wiley, Chicester, UK).

Stephen C. Hora serves as director of CREATE (the National Center of Excellence for Risk and Economic Analysis of Terrorism Events) at the University of Southern California, where he is appointed in the Viterbi School of Engineering and the Price School of Public Policy. Professor Hora's research has focused on the analysis of risks from technological hazards and terrorism. Most prominent has been his research into the use of experts to quantify risk and decision models.

Benjamin R. Fransen is a computer scientist working for the Defense Information Systems Agency (DISA). He earned his bachelor's degree in physics from St. Mary's College of Maryland and his doctoral degree in computer science from Pennsylvania State University. His research interests include autonomous systems, high performance computing, and applied statistics.

Natasha Hawkins is a section chief for the Department of Homeland Security's Office of Strategy, Planning, Analysis and Risk, leading modeling and analysis of homeland security events, developing a range of products that support strategic risk considerations, and providing technical support and subject matter expertise to execute risk and decision analysis for homeland security decision making.

Irving Susel serves as senior scientist for the Department of Homeland Security's Office of Strategy, Planning, Analysis and Risk, providing technical input, advice, and leadership for the development of a broad range of strategic assessments, risk management analyses, expert elicitation, and risk management tools. He served as a consultant for more than 30 years developing advanced quantitative approaches for setting risk-based priorities, especially where expert judgment is required.