Management Science

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

Is It Better to Average Probabilities or Quantiles?

Kenneth C. Lichtendahl, Jr., Yael Grushka-Cockayne, Robert L. Winkler,

To cite this article:
Kenneth C. Lichtendahl, Jr., Yael Grushka-Cockayne, Robert L. Winkler, (2013) Is It Better to Average Probabilities or Quantiles?. Management Science 59(7):1594-1611. http://dx.doi.org/10.1287/mnsc.1120.1667

MANAGEMENT SCIENCE
Vol. 59, No. 7, July 2013, pp. 1594–1611
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

informs
http://dx.doi.org/10.1287/mnsc.1120.1667
© 2013 INFORMS

Is It Better to Average Probabilities or Quantiles?

Kenneth C. Lichtendahl Jr., Yael Grushka-Cockayne
Darden School of Business, University of Virginia, Charlottesville, Virginia 22906
{lichtendahlc@darden.virginia.edu, grushkay@darden.virginia.edu}

Robert L. Winkler
Fuqua School of Business, Duke University, Durham, North Carolina 27708, rwinkler@duke.edu

We consider two ways to aggregate expert opinions using simple averages: averaging probabilities and averaging quantiles. We examine analytical properties of these forecasts and compare their ability to harness the wisdom of the crowd. In terms of location, the two average forecasts have the same mean. The average quantile forecast is always sharper: it has lower variance than the average probability forecast. Even when the average probability forecast is overconfident, the shape of the average quantile forecast still offers the possibility of a better forecast. Using probability forecasts for gross domestic product growth and inflation from the Survey of Professional Forecasters, we present evidence that both when the average probability forecast is overconfident and when it is underconfident, it is outperformed by the average quantile forecast. Our results show that averaging quantiles is a viable alternative and indicate some conditions under which it is likely to be more useful than averaging probabilities.

Key words: probability forecasts; quantile forecasts; expert combination; linear opinion pooling
History: Received March 27, 2012; accepted October 4, 2012, by Peter Wakker, decision analysis. Published online in *Articles in Advance* March 18, 2013.

1. Introduction

The most common way to combine experts' probability forecasts is to average them in what is known as a linear opinion pool (Stone 1961). This approach is appealing because it is a natural extension of the simple average that performs well for aggregating point forecasts. Many researchers have found that the average point forecast is robust and typically outperforms other methods for aggregating point forecasts (Clemen and Winkler 1986, Fildes and Makridakis 1995, Armstrong 2001, Larrick and Soll 2006). Because an expert's best point forecast is often the mean of her distribution (e.g., under a quadratic loss function), the average probability forecast is consistent with this finding: its mean of the underlying uncertain quantity is the average of the experts' means.

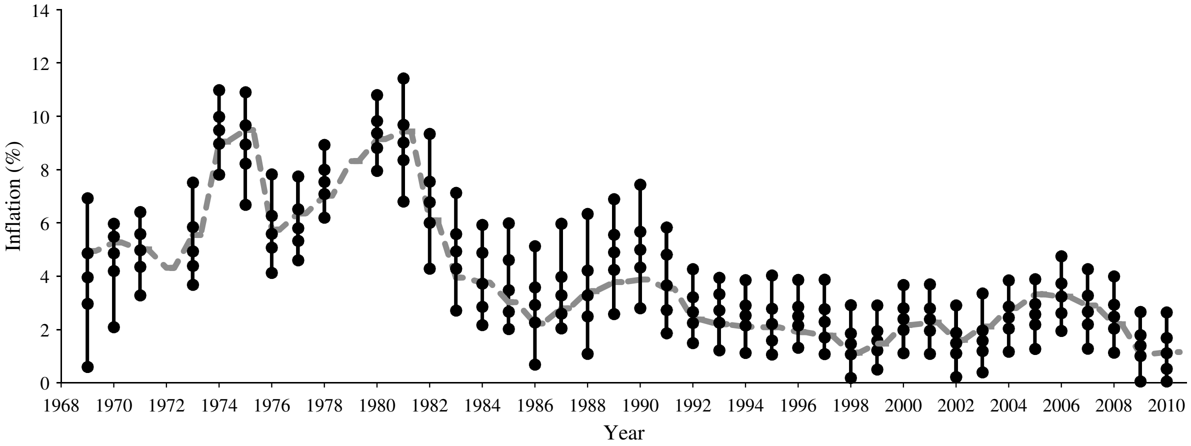
In combining probability forecasts, O'Hagan et al. (2006) note that the average probability forecast is simple and robust. Although the empirical evidence is somewhat limited, O'Hagan et al. (2006, p. 190) conclude that "In general, it seems that a simple, equally weighted opinion pool is hard to beat in practice." There are, however, well-known problems with averaging probabilities. Hora (2004) showed that when experts report perfectly calibrated probability forecasts, the average probability forecast cannot be perfectly calibrated (see also Ranjan and Gneiting 2010).

In many cases, the average probability forecast is underconfident because it spreads probability over too wide of a range (Ranjan 2009). For example, Figure 1 summarizes average probability forecasts of annual inflation reported by experts in the third quarter from 1969 to 2010 as part of the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters (SPF). The 50% central prediction intervals contain 73% of the realizations instead of the expected 50%. Thus, the 50% intervals are too wide, indicating underconfidence.

This paper considers an alternative aggregation method: averaging quantiles. This method, also known as Vincentization (Ratcliff 1979, Thomas and Ross 1980), involves horizontal averaging rather than vertical averaging of the experts' cumulative probability distributions. Although averaging quantiles, like averaging probabilities, is simple, it has not received much attention in the literature on the combination of probabilities. Our intent is to compare these two easy-to-use combining methods to see how the average quantile forecast performs and to see if it is worth adding to the mix of combining methods typically considered.

We study properties of the average probability and average quantile forecasts and compare their ability to harness the wisdom of the crowd. Our analytical results pertain to their overall performance, calibration, sharpness, and shape. In addition, empirical evidence from the SPF illustrates these analytical results and suggests that averaging quantiles is superior to averaging probabilities in that setting.

Figure 1 **Average Probability Forecasts (Solid Bars) and Realizations (Dashed Line) for Annual Inflation Reported in the Third Quarter from 1969 to 2010 as Part of the SPF**



Note. The dots in each year represent the average probability forecasts' 0.05, 0.25, 0.5, 0.75, and 0.95 quantiles.

On four scoring rules, we show analytically that the score of the average probability forecast can be no worse than the average score of the individual experts' forecasts. We also identify conditions under which the same result is true for the average quantile forecast on the quadratic scoring rule. Three factors that influence the calibration of the average probability forecast are identified: the number of experts, the degree to which the experts are overconfident, and the degree to which they disagree on the location of the uncertain quantity of interest. In terms of sharpness, the average quantile forecast is always sharper: it has lower variance than the average probability forecast. In comparing shapes, we highlight the effect that location or scale disagreement has on the shape of each average forecast.

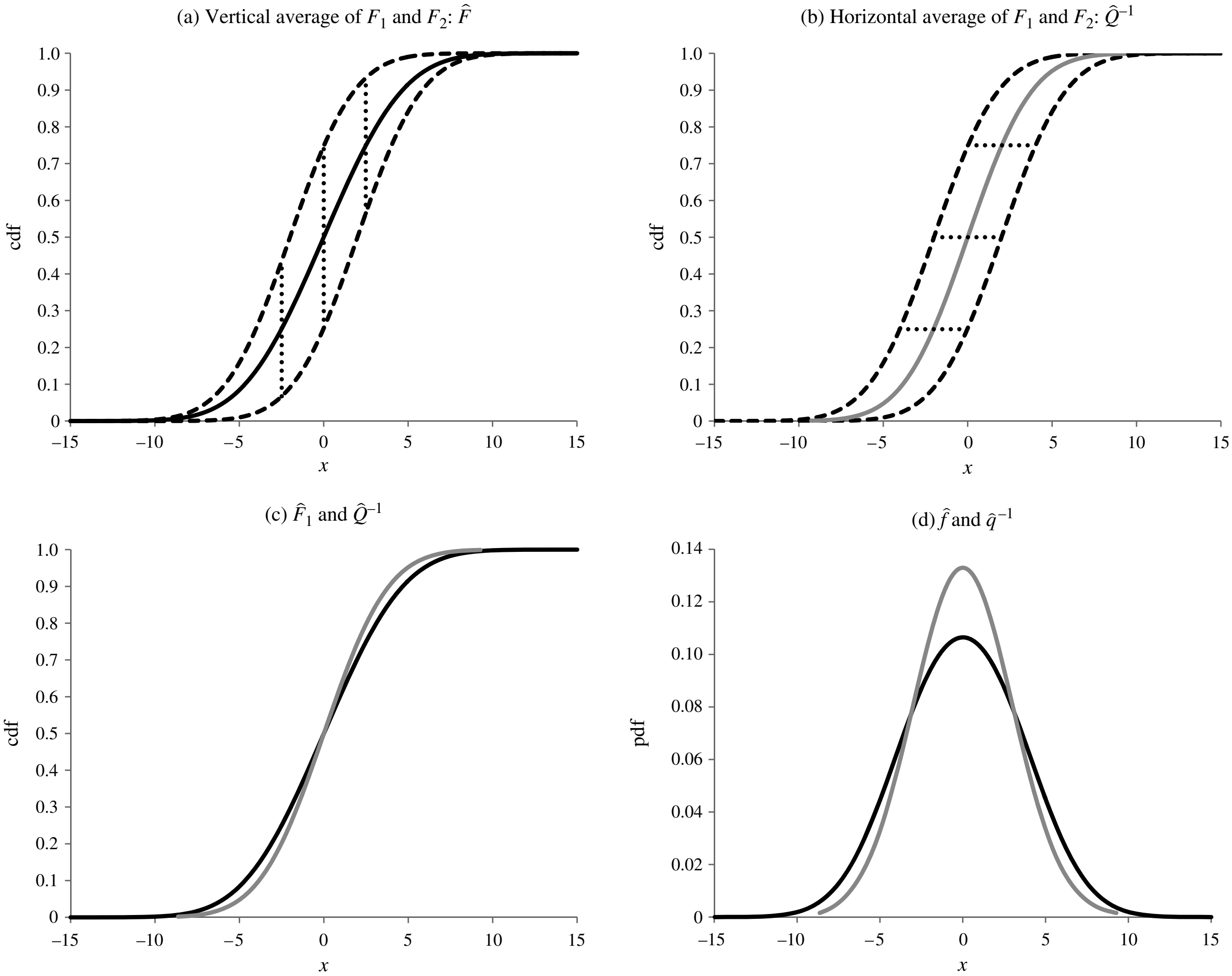
Finally, an analysis of SPF forecasts of annual gross domestic product (GDP) growth and inflation reveals that the average quantile forecast outperforms the average probability forecast in terms of average scores. We study the calibration, sharpness, and shape of the combined forecasts; compare the individual and combined forecasts; investigate the impact of factors such as location/scale disagreement; and relate the empirical results to the analytical results. Taken together, our results suggest that the average quantile forecast deserves more study and consideration and could have important implications for the way probability forecasts are combined.

2. Vertical vs. Horizontal Averaging

Suppose a decision analyst elicits k experts' probability forecasts of the continuous uncertain quantity x and combines these forecasts in one of two ways: by averaging either probabilities or quantiles. We define expert i's probability forecast as the cumulative distribution function (cdf) $F_i(x)$ with corresponding probability density function (pdf) given by $f_i(x) = dF_i(x)/dx$. Similarly, expert i's quantile forecast of x is the inverse of $F_i(x)$, denoted $Q_i(u)$. The average probability forecast of the k experts' forecasts is then given by $\hat{F}(x) = [F_1(x) + \cdots + F_k(x)]/k$, and the average quantile forecast by $\hat{Q}^{-1}(x)$, the inverse of $\hat{Q}(u) = [Q_1(u) + \cdots + Q_k(u)]/k$. We denote the pdf of the average probability forecast by $\hat{f}(x) = d\hat{F}(x)/dx$ and the pdf of the average quantile forecast by $\hat{q}^{-1}(x) = d\hat{Q}^{-1}(x)/dx$. In addition, we denote a forecast F's mean of x by $\mu = \int_{-\infty}^{\infty} x\, dF(x)$. A forecast F's even central moments of x are given by $E_F[(x-\mu)^j] = \int_{-\infty}^{\infty} (x-\mu)^j\, dF(x)$ for $j = 2, 4, 6, \ldots$. Expert i's (F_i's) mean, standard deviation, and variance of x are μ_i, σ_i, and σ_i^2, respectively. The average of the experts' means of x is denoted by $\hat{\mu} = (1/k)\sum_{i=1}^{k} \mu_i$, and the average of the experts' standard deviations of x is denoted by $\hat{\sigma} = (1/k)\sum_{i=1}^{k} \sigma_i$.

Some studies have compared the elicitation of a probability distribution using questions about probabilities [eliciting $F_i(x)$] and elicitation using questions about quantiles [eliciting $Q_i(u)$]. These two approaches may yield somewhat different distributions because they cause an expert to think about the situation in different ways, and whether one might be preferable is not clear. For example, Budescu and Du (2007) favor quantile elicitation, whereas Abbas et al. (2008) favor probability elicitation. This can be a relevant consideration in the choice of an elicitation procedure in a given situation. However, whether an expert's forecast distribution is elicited in terms of probabilities, quantiles, some of each (forcing the expert to think about the situation in different ways), or some other method does not matter in terms of our comparison of averaging probabilities and averaging quantiles. The distribution, regardless of how it is elicited, can be expressed either in terms of $F_i(x)$ or

Figure 2 Comparison of $\hat{F}$ and $\hat{Q}^{-1}$ and Their Corresponding Densities from Example 1



(a) Vertical average of F_1 and F_2: $\hat{F}$

(b) Horizontal average of F_1 and F_2: $\hat{Q}^{-1}$

(c) $\hat{F}_1$ and $\hat{Q}^{-1}$

(d) $\hat{f}$ and $\hat{q}^{-1}$

Note. F_1 and F_2 are shown as dashed black lines, $\hat{F}$ and $\hat{f}$ are shown as solid black lines, and $\hat{Q}^{-1}$ and $\hat{q}^{-1}$ are shown as solid gray lines.

$Q_i(u)$, which are equivalent because each function is the inverse of the other. On the other hand, when experts' forecasts are combined, the average probability forecast and the average quantile forecast may not be equivalent, as illustrated by Examples 1 and 2.

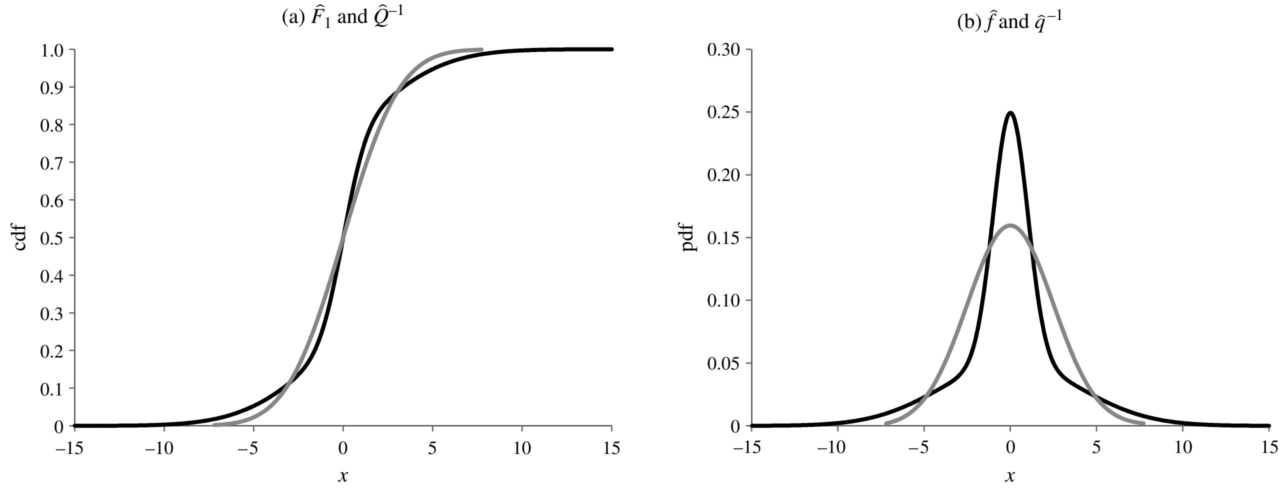
Example 1. Suppose two experts report normal cdfs for x. Both report a standard deviation of 3, but the means are −2 for the first expert and 2 for the second expert. Figures 2(a) and 2(b) depict the cdfs of the resulting average probability forecast and average quantile forecast, respectively. The dotted lines in these figures indicate that the average probability forecast is the result of "vertical averaging" and that the average quantile forecast is the result of "horizontal averaging." Figures 2(c) and 2(d) suggest that $\hat{F}$ and $\hat{Q}^{-1}$ will lead to different aggregate forecasts unless the experts' individual forecasts are identical.

Example 2. Now suppose that the two experts in Example 1 both report a mean of 3 for their normal distributions, but disagree on the standard deviation (one expert believes the standard deviation is 1, while the other believes it is 4). Figures 3(a) and 3(b) depict the cdfs and pdfs of the average probability forecast and the average quantile forecast, respectively. These figures confirm that, indeed, $\hat{F}$ and $\hat{Q}^{-1}$ will lead to different aggregate forecasts. When compared to Figures 2(c) and 2(d), these figures also demonstrate that the two aggregate forecasts might differ in a variety of ways. As we will see in §§3 and 4, the manner in which they differ in their shape has implications for their performance.

3. Analytical Results

According to Gneiting and Raftery (2007, p. 359), "the goal of probabilistic forecasting is to maximize the sharpness of the predictive distribution subject to calibration." In this section, we present analytical results concerning the calibration and sharpness of the average probability and quantile forecasts. Informally, a

Figure 3 **Comparison of $\hat{F}$ (Black Lines) and $\hat{Q}^{-1}$ (Gray Lines), and Their Corresponding Densities, from Example 2**



forecaster is "calibrated if his probability statements have their asserted coverage in repeated experience" (Rubin 1984, p. 1160). Intuitively, it makes sense that a forecaster should want to be calibrated. A forecaster should also want to be sharp; a calibrated forecast that is sharper in the sense of being more concentrated (e.g., having a smaller variance) is more informative. A scoring rule can be used to take into account both calibration and sharpness (Winkler 1996). Because both characteristics are important, we argue that the goal of probabilistic forecasting should be to maximize the score (the expected score ex ante, or the average score ex post).

3.1. Scoring Rules

A scoring rule is a function that takes a probability forecast F (often through its derivative, the pdf f) and the realization of x (denoted by x_0) and assigns the score $S(F, x_0)$. We consider four scoring rules to evaluate our forecasts, with a higher score indicating a better forecast:

(i) the linear score, $\text{Lin}\, S(F, x_0) = f(x_0)$;

(ii) the logarithmic score, $\text{Log}\, S(F, x_0) = \log(f(x_0))$;

(iii) the quadratic score, $\text{Quad}\, S(F, x_0) = 2f(x_0) - \int_{-\infty}^{\infty} [f(t)]^2\, dt$; and

(iv) the continuous ranked probability score, $\text{CRP}\, S(F, x_0) = -\int_{-\infty}^{x_0} [F(t)]^2\, dt - \int_{x_0}^{\infty} [1 - F(t)]^2\, dt$.

A scoring rule is strictly proper if an expert's expected score ex ante (or average score ex post over multiple forecasts and realizations) can be maximized only by reporting probabilities honestly. Such a rule rewards both calibration and sharpness, and most rules used in practice are strictly proper, as are three of our four rules. The linear score is not strictly proper, but it can be viewed as a predictive likelihood, which is simply the height of a forecast's density function at x_0.

The linear and logarithmic scores are local rules in that they depend only on $f(x_0)$, the density at the realization, and not on the rest of the density function (Winkler 1969). This characteristic has some intuitive appeal and makes it easy to calculate and understand scores. The logarithmic rule is the only strictly proper rule that is local (Shuford et al. 1966). The quadratic scoring rule depends not just on $f(x_0)$, but on the entire density function. Finally, the continuous ranked probability score can be attractive because it is sensitive to distance (Matheson and Winkler 1976). For more details on scoring rules, see Winkler (1996) and Gneiting and Raftery (2007).

Larrick et al. (2011, p. 7) describe the way in which the average point forecast harnesses the wisdom of the crowd: "Averaging the [point forecasts] of a crowd therefore ensures a level of accuracy no worse than the average member of the crowd and, in some cases, a level better than nearly all members." They measured the level of accuracy using a quadratic loss function (or squared error). We state a similar result for the average probability forecast using scoring rules to measure accuracy. Proofs are given in the appendix.

Proposition 1. *$S(\hat{F}, x_0) \geq (1/k) \sum_{i=1}^{k} S(F_i, x_0)$ for the logarithmic, quadratic, and continuous ranked probability scores, and the inequality is strict if $f_i(x_0) \neq f_j(x_0)$ for some $i \neq j$. Equality holds for the linear score.*

For the linear and logarithmic scores, Proposition 1 can be extended to $\hat{Q}^{-1}$. If $\hat{q}^{-1}(x_0) - \hat{f}(x_0) \geq 0$, it follows immediately that $S(\hat{Q}^{-1}, x_0) \geq S(\hat{F}, x_0)$ for the linear and logarithmic scores, because these scores depend only on the density at x_0. From Proposition 1, this yields $S(\hat{Q}^{-1}, x_0) \geq (1/k) \sum_{i=1}^{k} S(F_i, x_0)$.

The next result provides conditions under which $\hat{Q}^{-1}$'s quadratic score is no worse than the average quadratic score of the experts' forecasts. For this result, we assume that each expert reports a cdf from the same location-scale family. Location-scale families, which include the normal, t, uniform, Cauchy, Laplace, Gumbel, and elliptical, are the only families of distributions closed under quantile averaging (Thomas and Ross 1980). F_i is a member of a location-scale family with location parameter μ_i and scale parameter σ_i if and only if its density $f_i(x) = (1/\sigma_i)g((x-\mu_i)/\sigma_i)$ for some function g. Let $f(x)$ and $F(x)$ be the pdf and cdf of the standardized member of the family with $\mu_i = 0$ and $\sigma_i = 1$. The quantile function of the family is $Q_i(u) = \mu_i + \sigma_i Q(u)$, where Q is the inverse of F. If each Q_i comes from the same location-scale family, then $\hat{Q}$ is in that same family.

Proposition 2. *Assume that the experts report from the same location-scale family with* $c = \int_{-\infty}^{\infty} g(z)^2\,dz$. *Then* (i) $\operatorname{Quad}S(\hat{Q}^{-1}, x_0) \geq (1/k)\sum_{i=1}^{k} \operatorname{Quad}S(F_i, x_0)$ *if and only if the density-at-the-realization condition* $\hat{q}^{-1}(x_0) - \hat{f}(x_0) \geq cd/2$ *holds, where* $d = 1/\hat{\sigma} - (1/k)\sum_{i=1}^{k}(1/\sigma_i)$; (ii) $cd/2 \leq 0$, *and the more disagreement there is among the experts' scale parameters* σ_i *in terms of a mean-preserving spread (as defined in the appendix), the more negative* d *is and the more negative* $\hat{q}^{-1}(x_0) - \hat{f}(x_0)$ *can therefore be, with* $\hat{Q}^{-1}$ *still having a higher quadratic score than the average quadratic score of the experts' forecasts.*

Proposition 2 shows that $\hat{q}^{-1}(x_0) - \hat{f}(x_0)$ need not be positive for $\hat{Q}^{-1}$ to have a higher quadratic score than the average quadratic score for the experts. In §3.4, we show analytically how scale disagreement can lead to shape differences in $\hat{q}^{-1}$ and $\hat{f}$ that have important implications for scores. In §4, we demonstrate empirically that Proposition 2's density-at-the-realization condition holds on average. In fact, it holds to such an extent in the SPF data that $\hat{q}^{-1}(x_0) - \hat{f}(x_0) \geq 0$ on average for all quarters for both GDP growth and inflation.

3.2. Calibration

We begin with the definition of a forecast's probability integral transform, a key building block in previous work on calibration (Morris 1977, Dawid 1984, Smith 1985, Diebold et al. 1998, Hora 2004, Gneiting et al. 2007, Ranjan 2009, Mitchell and Wallis 2011, Wallis 2011).

Definition 1 (PIT). A forecast F's probability integral transform (PIT) is given by the random variable $F(x)$, where F is a random cdf and x is a random variable.

The realized PIT "is the value that the predictive cdf attains at the observation" (Gneiting et al. 2007, p. 251). An important observation from previous work on calibration is that a perfectly calibrated forecast has a uniformly distributed PIT. Below, we formally define overconfidence and underconfidence in terms of a forecast's PIT. This definition has a direct link to the traditional ex post notion of overconfidence (e.g., in a series of realized PITs, less than 50% of the PITs fall between 0.25 and 0.75).

We do not use separate notation for a random function and its realization. Instead, we rely on the context to indicate to which instance we are referring. For example, when we discuss the mean and variance of a forecast F's PIT, written $E[F(x)]$ and $V[F(x)]$, respectively, it should be understood that F is a random cdf. On the other hand, with a forecast F's mean of x, written $\int x\,dF(x)$, F is a realized cdf.

Definition 2 (Calibration). We say a forecast F is

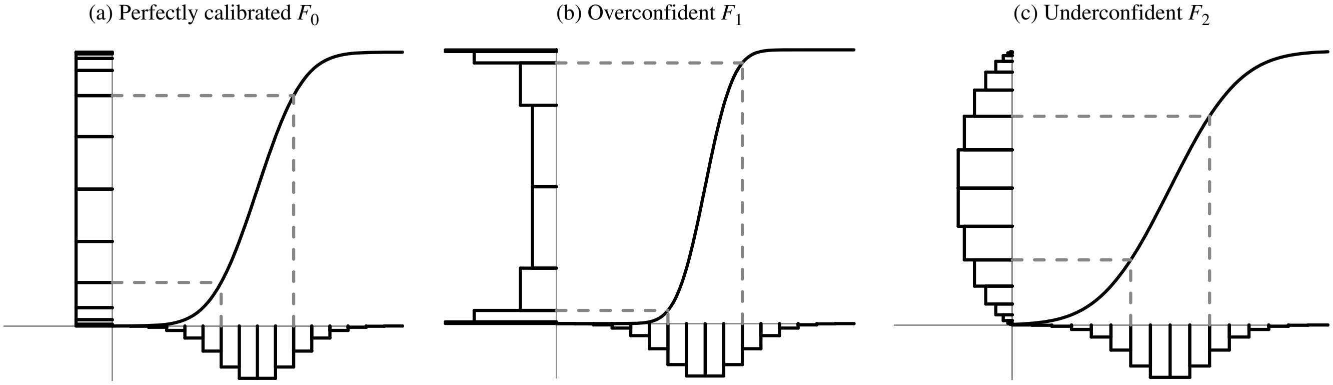
(a) perfectly calibrated if $\Pr(F(x) \leq u) = u$ for $0 < u < 1$;

(b) overconfident if $\Pr(F(x) \leq u) \geq u$ for $0 < u < 1/2$, $\Pr(F(x) \leq u) = u$ for $u = 1/2$, and $\Pr(F(x) > u) \leq u$ for $1/2 < u < 1$, with the inequalities strict for some u; and

(c) underconfident if $\Pr(F(x) \leq u) \leq u$ for $0 < u < 1/2$, $\Pr(F(x) \leq u) = u$ for $u = 1/2$, and $\Pr(F(x) > u) \geq u$ for $1/2 < u < 1$, with the inequalities strict for some u.

Example 3. Suppose x is normally distributed with mean $\mu_0 = 4$ and standard deviation $\sigma_0 = 1$. Its cdf F_0 is shown in Figure 4(a). The bars below the axes in Figure 4 correspond to F_0's probabilities that x falls

Figure 4 Comparison of Perfectly Calibrated, Overconfident, and Underconfident Forecasts



(a) Perfectly calibrated F_0

(b) Overconfident F_1

(c) Underconfident F_2

within the corresponding intervals. By Theorem 2.1.10 of Casella and Berger (2002, p. 54), $\Pr(F_0(x) \le u) = u$ for $0 < u < 1$, which means that F_0 is perfectly calibrated. The bars to the left of the axes in Figure 4 correspond to the probabilities that a forecast's PIT falls within the PIT intervals corresponding to the intervals for x. Because F_0 is perfectly calibrated, F_0's PIT probabilities are uniform.

We focus on F_0's lower tail, specifically on the event $x \le \mu_0 - \sigma_0$, to explain the PIT for the forecasts F_1 and F_2, which are normal distributions with mean $\mu_0 = 4$ and standard deviations $\sigma_1 = 0.6$ and $\sigma_2 = 1.4$, respectively. Under F_0, the probability that $x \le \mu_0 - \sigma_0$ is $F_0(\mu_0 - \sigma_0) = 0.1587$. In each plot in Figure 4, this probability is represented by the height of the left vertical dashed line. The corresponding event that F_i's PIT falls below $F_i(\mu_0 - \sigma_0)$ is given by the area under the PIT histogram below the bottom horizontal dashed line.

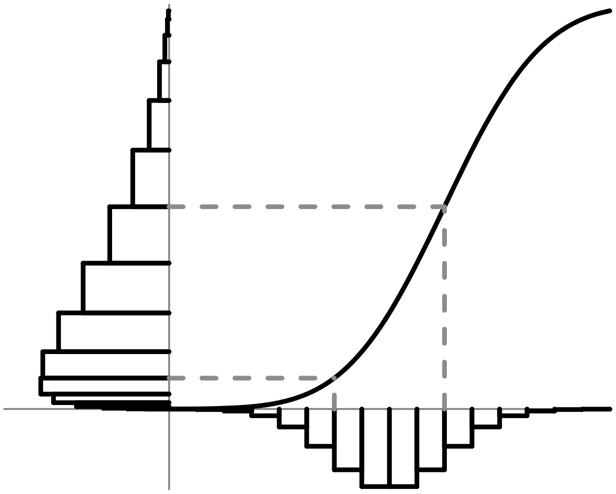
Under F_1, $x \le \mu_0 - \sigma_0$ corresponds to F_1's PIT falling below $F_1(\mu_0 - \sigma_0) = 0.0478$. And because x falls in F_0's lower tail with probability 0.1587, F_1's PIT falls below 0.0478 with probability 0.1587. Therefore, the probability that F_1's PIT falls below 0.1587 is greater than 0.1587. In other words, when x falls in F_0's lower tail, x falls further out in F_1's lower tail. According to Definition 2, F_1 is overconfident, and the density of F_1's PIT is bathtub-shaped (as shown in Figure 4(b)). For F_2, in contrast, $F_2(\mu_0 - \sigma_0) = 0.2375$, and the probability that F_2's PIT falls below 0.1587 is less than 0.1587. When x falls in F_0's lower tail, x does not fall as far out in F_2's lower tail. Thus, F_2 is underconfident and the density of F_2's PIT is hump-shaped (as shown in Figure 4(c)).

In Example 3, the distributions of the forecasts' PITs are symmetric about 1/2. Thus, their medians and means equal 1/2. If F_2's mean of x was 5, then the distribution of F_2's PIT would be right skewed, as shown in Figure 5, and its median and mean would be below 1/2. In this case, F_2's location bias prevents firm conclusions about overconfidence or underconfidence. Definition 2 requires that the median of the distribution of a forecast's PIT be equal to 1/2 for such conclusions. In the following result, we also require that its mean be 1/2, which holds when the distribution of a forecast's PIT is symmetric about 1/2.

Proposition 3. *If F is perfectly calibrated, then the variance of F's PIT is equal to* 1/12. *If $E[F(x)] = 1/2$ and F is overconfident* (*underconfident*), *then the variance of F's PIT is greater* (*less*) *than* 1/12.

Proposition 3 states, in part, that high PIT variance (assuming the PIT's mean is 1/2) is a necessary condition for overconfidence. It is, however, not a sufficient condition. All we can conclude from $V[F(x)] > 1/12$ is that F is not underconfident, or that F is possibly overconfident. Nonetheless, for ease of exposition, we say F is overconfident when $V[F(x)] > 1/12$ and take the qualifier "possibly" to be implicit.

Figure 5 **Forecast with a Right-Skewed Probability Integral Transform Distribution**



We measure calibration in terms of the variance of a forecast's PIT. Another measure is the L^1-norm, the sum of the areas between the 45-degree line and the cdf of F's PIT (Hora 2004, 2010). Such a measure captures deviations from perfect calibration, but fails to indicate the nature of the miscalibration (e.g., underconfident, overconfident, or something else). For details on other statistical tests of calibration, such as the Anderson–Darling test for PIT uniformity, see Mitchell and Wallis (2011).

This brings us to a calibration result involving the variance of $\hat{F}$'s PIT. This result is meant to illustrate some of the factors that impact the degree to which $\hat{F}$ is either underconfident or overconfident.

Proposition 4. *Assume that there are $k > 1$ experts who are exchangeable in the sense that* (i) *the variance of each expert's PIT is v, and* (ii) *the pairwise correlations of the experts' PITs are all $\rho > -1/(k-1)$. Then the variance of the average probability forecast's PIT is $V[\hat{F}(x)] = (\rho + (1-\rho)/k)v$.*

The condition $\rho > -1/(k-1)$ ensures that the variance of $\hat{F}$'s PIT is positive. From Proposition 4, we get a version of Hora's (2004) result: When each of k exchangeable experts is perfectly calibrated and their PITs are not perfectly positively correlated, the average probability forecast cannot be perfectly calibrated. (This follows directly from Proposition 3 because $\rho + (1-\rho)/k < 1$ when $\rho < 1$.) In addition, we can see that the average probability forecast becomes, loosely speaking, less overconfident (or more underconfident) as k increases, v decreases, or ρ decreases.

An important factor in Proposition 4 is ρ, which relates to the degree to which the experts disagree on the location (e.g., the mean) of x. The correlation coefficient ρ will tend to be lower when the experts disagree on the location. To see this, consider a possible scenario from Example 1: One expert reports a low mean, another expert reports a high mean, and the underlying quantity falls in between these means. In this case, the PIT will be high for the low-mean expert and low for the high-mean expert.

When a large number of exchangeable experts are overconfident (i.e., $k \gg 1$ and $v > 1/12$), the degree to which the experts disagree on the location influences whether the average probability forecast is overconfident or not. The more location disagreement; the more underconfident or less overconfident the average probability forecast will tend to be. The less location disagreement, the more overconfident or less underconfident the average probability forecast will tend to be.

The following results provide conditions under which the average probability forecast is more underconfident (or less overconfident) than the average quantile forecast. In Proposition 5, we assume that the experts' means fall symmetrically about $\hat{\mu}$. That is, if $\mu_1 < \mu_2 < \cdots < \mu_k$, $|\mu_j - \hat{\mu}| = |\mu_{k-j+1} - \hat{\mu}|$ for $1 \le j \le k/2$, and $\mu_{(k+1)/2} = \hat{\mu}$ if k is odd. We also assume that each expert reports a cdf from the same location-scale family and that the cdf is symmetric about its median.

Proposition 5. *Assume that the experts report from the same location-scale family and F, the cdf of the standardized member of the location-scale family, is symmetric about* 0 *and unimodal. In addition, assume that the experts' locations fall symmetrically about $\hat{\mu}$ and the experts agree on the scale parameter $\hat{\sigma}$. Then the average probability forecast's PIT is closer to* $1/2$ *than the average quantile forecast's PIT. That is, $\hat{F}(x) > \hat{Q}^{-1}(x)$ if $x < \hat{\mu}$, $\hat{F}(x) < \hat{Q}^{-1}(x)$ if $x > \hat{\mu}$, and $\hat{F}(x) = \hat{Q}^{-1}(x)$ if $x = \hat{\mu}$. Moreover, $\hat{F}$ and $\hat{Q}^{-1}$ are both symmetric about their median $\hat{\mu}$.*

Consider again the average forecasts in Figure 2(c) from Example 1, where two experts disagree on the location of the underlying quantity. When x_0 falls in the left tail of $\hat{F}$, obtaining a PIT value of $\hat{F}(x_0)$, it will tend to fall in the left tail of $\hat{Q}^{-1}$ at a lower value $\hat{Q}^{-1}(x_0)$. Similarly, when x_0 falls in the right tail of $\hat{F}$, $\hat{Q}^{-1}$ will tend to have a higher PIT value. The result below relates the single-crossing property from Proposition 5 to the variance of the average forecasts' PITs.

Proposition 6. *Given the assumptions in Proposition* 5 *and $E[\hat{F}(x)] = E[\hat{Q}^{-1}(x)] = 1/2$, $V[\hat{F}(x)] < V[\hat{Q}^{-1}(x)]$. Moreover, if $x - \hat{\mu}$ is distributed according to the cdf $F(x/\hat{\sigma})$ with $\sigma < \hat{\sigma}$, $V[\hat{Q}^{-1}(x)] < 1/12$.*

Thus, when experts disagree on the location, the average quantile forecast will tend to be more overconfident (or less underconfident) than the average probability forecast. When the experts disagree, however, on the scale of the underlying quantity (and agree on the location) as in Example 2, it is difficult to determine how $V[\hat{F}(x)]$ and $V[\hat{Q}^{-1}(x)]$ will compare. Figure 3(a) provides an illustration of the source of the difficulty, as $\hat{F}$ and $\hat{Q}^{-1}$ cross each other three times.

Next, we show that the average quantile forecast is sharper than the average probability forecast. In cases where there is more location disagreement and the average probability forecast is underconfident, the average quantile forecast will tend to score better because its increased sharpness leads to better calibration. Nonetheless, even in cases where there is less location disagreement and the average probability forecast is overconfident, we will show in §3.4 that the more overconfident average quantile forecast may still score better because of its shape.

3.3. Sharpness

Here, we provide two results that compare $\hat{F}$'s and $\hat{Q}^{-1}$'s moments of x.

Proposition 7. *The average probability forecast and the average quantile forecast have identical means of x equal to $\hat{\mu}$, the average of the experts' means of x.*

Definition 3 (Relative Sharpness). We say F is sharper than G if F's variance of x is less than or equal to G's.

Proposition 8. (i) *The average quantile forecast is sharper than the average probability forecast.* (ii) *Each of $\hat{Q}^{-1}$'s even central moments of x is less than or equal to $\hat{F}$'s.*

Propositions 7 and 8 state that the average quantile forecast is more concentrated than the average probability forecast around $\hat{\mu}$, their identical mean of x. In addition to its higher second central moment (the variance), part of what makes $\hat{F}$ less concentrated is its higher fourth central moment (a component of kurtosis), which indicates that it has fatter tails than $\hat{Q}^{-1}$. These results suggest that when the average probability forecast is underconfident, the average quantile forecast may score better because its increased sharpness leads to better calibration. This is particularly important in light of the fact that the miscalibration associated with the averaging of perfectly calibrated probability forecasts noted in §1 leads to average probability forecasts that are overdispersed, or underconfident (Ranjan 2009).

We conclude this section by illustrating the role that location disagreement plays in comparing the sharpness of $\hat{F}$ and $\hat{Q}^{-1}$. $\hat{F}$'s variance of x is given by $\int_{-\infty}^{\infty} (x-\hat{\mu})^2 \, d\hat{F}(x) = (1/k)\sum_{i=1}^{k} \sigma_i^2 + (1/k)\sum_{i=1}^{k} (\mu_i - \hat{\mu})^2$. Its sharpness decreases as the sample variance of the experts' means of x increases. When each expert reports from a normal distribution, $\hat{Q}^{-1}$'s variance of x is $\hat{\sigma}^2$ (Thomas and Ross 1980). This variance is less than $(1/k)\sum_{i=1}^{k}\sigma_i^2$ by Jensen's inequality and therefore less than $\hat{F}$'s variance of x. In this case, $\hat{Q}^{-1}$'s sharpness does not change as the sample variance of the experts' means of x changes. Taken together with the results in §3.2, we can see that location disagreement causes the average probability forecast to be less overconfident (or more underconfident) and less sharp.

Of course, in practice, the sharpness of $\hat{Q}^{-1}$ may be too great and lead to an overconfident $\hat{Q}^{-1}$ that one might think would score worse than $\hat{F}$. This, however, is not the case in the SPF data, where an overconfident $\hat{Q}^{-1}$ scores better, in part because of the shape properties discussed in §3.4.

3.4. Shape

In §3.4, we show that even when there is location agreement and the average probability forecast is overconfident, the more overconfident average quantile forecast may still score better. Our first two results involving shape concern the case of location agreement.

Proposition 9. *Assume that the experts report from the same location-scale family and agree on the location parameter $\hat{\mu}$ (but not the scale parameter). Then $\hat{f}(\hat{\mu}) \geq \hat{q}^{-1}(\hat{\mu})$.*

Proposition 9 is surprising because the average quantile forecast always has lower variance. How can the average probability forecast have a higher density at the mean, but be less sharp? The answer lies in the shoulders (away from the mean, but not too far away in the tails) and the tails. The next result indicates that the average quantile forecast is sharper because it has a higher density in the shoulders and a lower density in the tails. For sharpness, the variance of x is generally influenced most by the tails. We define a forecast's head, shoulders, and tails based on a partition $a_1 < a_2 < a_3 < a_4$ where (i) the head is over the region $(a_2 a_3)$, (ii) the shoulders are over the region $(a_1 a_2) \cup (a_3 a_4)$, and (iii) the tails are over the region $(-\infty, a_1) \cup (a_4, \infty)$.

We also consider a related measure, a forecast's kurtosis, the ratio of its fourth central moment of x to the square of its variance of x. Kurtosis is a shape characteristic that can be viewed as "the location- and scale-free movement of probability mass from the shoulders of a distribution into its center and tails" (Balanda and MacGillivray 1988, p. 111). A normal distribution has kurtosis 3. Distributions with kurtosis greater (less) than 3 are called leptokurtic (platykurtic) and are typically characterized by densities with sharp, narrow peaks and long, fat tails (flatter, wider peaks and thinner tails).

Proposition 10. *Assume that the experts report normal distributions and agree on the location parameter $\hat{\mu}$ (but not the scale parameter). Then* (i) *for some partition $a_1 < a_2 < \hat{\mu} < a_3 < a_4$, $\hat{f}(x) \geq \hat{q}^{-1}(x)$ in the head and tails and $\hat{q}^{-1}(x) \geq \hat{f}(x)$ in the shoulders; and* (ii) *$\hat{F}$ has higher kurtosis than $\hat{Q}^{-1}$'s kurtosis of* 3.

Example 2 (Revisited). Here two experts report normal distributions with mean 0 and standard deviations 1 and 4. As can be seen from Figure 3(b), the average quantile forecast is higher in the shoulders, with $a_4 = -a_1 = 4.885$, $a_3 = -a_2 = 1.228$, $\hat{F}(a_1) = 0.0555$, and $\hat{F}(a_2) = 0.2446$. Thus, the head is over $\hat{F}$'s central 51.1% prediction interval, and the shoulders are over $\hat{F}$'s central 88.9% prediction interval outside the head.

The next two results concern the situation when the forecasters disagree on the location parameter. For these results, we assume that the standardized member of the location-scale family from which the experts report is unimodal with its mode at 0. Any distribution from a location-scale family that is symmetric and unimodal satisfies this condition.

Proposition 11. *Assume that the experts report from the same location-scale family and agree on the scale parameter $\hat{\sigma}$ (but not the location parameter). In addition, assume that the standardized member of the location-scale family is unimodal with its mode at 0. Then $\hat{q}^{-1}(\hat{\mu}) \geq \hat{f}(\hat{\mu})$.*

Proposition 12. *Assume that the experts report normal distributions and agree on the scale parameter $\hat{\sigma}$. Then* (i) *for some partition $a_1 < a_2 < \hat{\mu} < a_3 < a_4$, $\hat{q}^{-1}(x) \geq \hat{f}(x)$ in the head and $\hat{f}(x) \geq \hat{q}^{-1}(x)$ in the shoulders and tails; and* (ii) *$\hat{F}$'s kurtosis is greater* (*less*) *than or equal to $\hat{Q}^{-1}$'s kurtosis of* 3 *if the sample kurtosis in the experts' means of x is greater* (*less*) *than* 3.

Figure 2(d) illustrates Proposition 12, depicting the densities of the average probability forecast and average quantile forecast from Example 1, where two experts report normal distributions, agree on the standard deviation, and disagree on the mean.

4. Empirical Results

In this section, we analyze the probability forecasts of annual GDP growth (1982–2009) and annual inflation (1968–2010) reported by expert panelists surveyed in the first through fourth quarters (Q1, Q2, Q3, Q4) as part of the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters (Croushore 1993). In each quarter during calendar year t (i.e., at four different lead times), panelists reported probabilities that changes in U.S. real GDP (i.e., GDP growth) and changes in U.S. GDP price index (i.e., inflation) between the years $t-1$ and t would fall in one of several predetermined bins. For example, in the third quarter of 2007 (2007:Q3), the survey asked for probabilities that changes in annual average U.S. real GDP between the years 2006 and 2007 would fall in 1 of 10 predetermined bins: below −2%, between −2% and −1%, …, between 5% and 6%, and above 6%. Once blank entries and forecasts not summing to 100% are removed, the sample consists of 3,462 individual forecasts for GDP growth and a total of 6,116 forecasts for inflation (downloaded from Philadelphia Fed's website: http://www.phil.frb.org/econ/spf/). For more details on the SPF data, see Diebold et al. (1999), Engelberg et al. (2009), and Clements (2010).

Following Diebold et al. (1999), we approximate the SPF panelists' continuous forecasts by fitting piecewise-linear cdfs to their reported bin probabilities. For the lower and upper open bins, we assume that they have the same width as the interior bins (which all have equal width). With these piecewise-linear cdfs, we average vertically and horizontally to find $\hat{F}$ and $\hat{Q}^{-1}$, respectively. The piecewise-linear approximation offers several advantages. The piecewise-linear family of cdfs is closed under both probability ("vertical") and quantile ("horizontal") averaging. In addition, statistics related to these average forecasts, such as their scores and moments, are easy to calculate.

4.1. Scores

In Table 1, we present average scores (averaged over the years separately by quarter and quantity) for the individual experts, $\hat{F}$, and $\hat{Q}^{-1}$. Average scores for the individual experts are given only for the quadratic and continuous ranked probability scores. For the linear score, the average score for the experts will always equal the average score for $\hat{F}$, and for the logarithmic score, the score is $-\infty$ for at least one expert in most years. Furthermore, in 1992:Q4 and 1996:Q4, the realizations of GDP growth fell above the upper ends of the supports of $\hat{Q}^{-1}$. Consequently, the average quantile forecast received a logarithmic score of negative infinity in these years in the fourth quarter. These infinitely bad scores, however, may be the result of rounding very small probabilities down to 0 and not the result of poorly supported beliefs. As Engelberg et al. (2009, p. 40) report, "Most probabilistic forecasts reported in the SPF are multiples of 0.05. This suggests that forecasters tend to round their responses." When we exclude these two years from the fourth quarter's data, the average logarithmic scores are −1.505 and −1.318 for $\hat{F}$ and $\hat{Q}^{-1}$, respectively.

For each scoring rule where we have average individual scores, both $\hat{F}$ and $\hat{Q}^{-1}$ have higher scores, often surprisingly higher. Moreover, from Table 2, the percentage of experts beaten by $\hat{F}(\hat{Q}^{-1})$ is over 50% in 29 (32) of 32 cases, ranging from 47.4% to 61.8% for $\hat{F}$ and from 52.3% to 64.2% for $\hat{Q}^{-1}$. These results are consistent with Propositions 1 and 2.

Comparing $\hat{F}$ and $\hat{Q}^{-1}$, we see that $\hat{Q}^{-1}$ scores better than $\hat{F}$ in most cases. On the linear and quadratic scores, $\hat{Q}^{-1}$ scores better than $\hat{F}$ in all eight quarters (four quarters for each of the two quantities). This result for the linear score means that, on average, $\hat{q}^{-1}(x_0) - \hat{f}(x_0) \geq 0$, which is stronger than the density-at-the-realization condition in Proposition 2. On the logarithmic score, $\hat{Q}^{-1}$ scores better than $\hat{F}$ in seven of the eight quarters. On the continuous ranked probability score, $\hat{Q}^{-1}$ scores better than $\hat{F}$ in six of the eight quarters. Also, $\hat{Q}^{-1}$ beats a higher percentage of experts than $\hat{F}$ in 12 of 16 cases for GDP growth and all 16 cases for inflation.

Table 1 **Average Scores for the Experts' Forecasts, $\hat{F}$, and $\hat{Q}^{-1}$**

	SPF GDP growth (1982–2009)[a]				SPF inflation (1968–2010)[a, b]			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Lin $S(\hat{F}, x_0)$	0.201	0.253	0.297	0.279	0.299	0.347	0.412	0.463
Lin $S(\hat{Q}^{-1}, x_0)$	0.219	0.271	0.305	0.297	0.335	0.372	0.427	0.494
Log $S(\hat{F}, x_0)$	−1.859[c]	−1.592	−1.507	−1.662	−1.358	−1.205	−1.028	−0.956
Log $S(\hat{Q}^{-1}, x_0)$	−1.729[c]	−1.426	−1.346	$-\infty$[d]	−1.176	−1.038	−0.897	−0.771
Quad S(experts)	0.048	0.126	0.156	−0.004	0.202	0.275	0.369	0.375
Quad $S(\hat{F}, x_0)$	0.135	0.197	0.223	0.074	0.301	0.368	0.463	0.473
Quad $S(\hat{Q}^{-1}, x_0)$	0.153	0.242	0.265	0.149	0.351	0.412	0.493	0.551
CRP S(experts)	−0.897	−0.656	−0.601	−0.630	−0.583	−0.497	−0.426	−0.380
CRP $S(\hat{F}, x_0)$	−0.775	−0.579	−0.531	−0.559	−0.459	−0.386	−0.322	−0.299
CRP $S(\hat{Q}^{-1}, x_0)$	−0.801	−0.580	−0.530	−0.551	−0.450	−0.379	−0.310	−0.285

[a] Excludes 1985:Q1 and 1986:Q1 due to an error the Philadelphia Fed suspects was present in the survey.

[b] Excludes1968:Q4, 1969:Q4, 1970:Q4, 1971:Q4, 1972:Q3–Q4, 1973:Q4, 1975:Q4, 1976:Q4, 1977:Q4, 1978:Q4, and 1979:Q2–Q4 due to an error the Philadelphia Fed has identified in the survey.

[c] Excludes 2009. In 2009, GDP growth was −3.49% and below −3%, the lower end of each F_i's support.

[d] In 1992 and 1996, the realizations of GDP growth were 3.39% and 3.74% and the upper ends of the supports of $\hat{Q}^{-1}$ were 3.22% and 3.34%, respectively.

Table 2 Percentage of Experts Beaten by $\hat{F}$ and $\hat{Q}^{-1}$ on Each of the Four Scoring Rules

		SPF GDP growth (1982–2009) (%)				SPF inflation (1968–2010) (%)			
	Score	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
$\hat{F}$	Lin S	55.3	53.7	52.5	53.7	55.3	52.2	51.1	47.4
	Log S	55.3	53.7	52.5	53.7	55.3	52.2	51.1	47.4
	Quad S	61.8	57.7	53.3	54.7	57.7	56.7	54.4	48.6
	CRP S	55.9	54.7	52.5	59.3	55.3	53.6	55.9	55.0
$\hat{Q}^{-1}$	Lin S	58.0	61.4	54.9	52.3	63.0	59.3	55.6	53.5
	Log S	58.0	61.4	54.9	52.3	63.0	59.3	55.6	53.5
	Quad S	61.2	61.9	58.0	61.3	64.2	59.7	54.8	58.1
	CRP S	55.5	57.4	55.4	62.5	59.8	57.8	63.1	63.4

CRP S is the only one of the four rules that is sensitive to distance. It rewards not just a high density at the realization, but it also rewards a forecast for having more probability close to the realization. In cases of location agreement, $\hat{Q}^{-1}$'s higher density in the shoulder opposite the realization is farther away from the realization than $\hat{F}$'s higher density in the head. This explains why $\hat{F}$ gets extra credit under CRP S for the higher probability in between $\hat{Q}^{-1}$'s two shoulders. Whether such a characteristic is important depends on whether having more probability "close to x_0" matters in the context of a decision.

For the SPF forecasts, then, averages of probabilities or quantiles from multiple experts perform better than the probabilities from the individual experts. In other words, each average forecast harnesses the wisdom of the crowd. The results also indicate that it is better to average quantiles than to average probabilities in this macroeconomic forecasting environment.

4.2. Calibration

In Tables 3–5, we report empirical results that pertain to calibration. From Table 3, the average number of experts in the pool is as low as 30 in Q3 on GDP growth and as high as 37 in Q2 on inflation. From Table 4, the average PIT means are all between 0.41 and 0.59, indicating that the PIT distributions are not perfectly symmetric about 1/2 but that the degree of asymmetry does not seem to be too large.

The statistics of greatest interest in Table 4 are the variances. The average variance of the experts' individual PITs for the 20 experts who submitted at least 10 forecasts in all quarters suggests, in comparison with the benchmark variance of 1/12 given in Proposition 3, that the experts are overconfident for GDP growth and somewhat underconfident for inflation. For both variables, the average PIT variance is higher for Q1 and Q4 than for Q2 and Q3. Table 5 gives average pairwise correlations of individual experts' PITs for the 5 experts who jointly participated in the most overlapping years. The SPF experts have PIT values that are reasonably highly correlated.

The overconfidence we find in Table 4 is based on sample PIT variances. In theory, a series of PITs should be independently and identically distributed uniform random variables (Dawid 1984, Diebold et al. 1998). In their study of the SPF inflation data, Diebold et al. (1999) found positive serial correlation in the PITs of the average probability forecast. This positive serial correlation confounds estimates of PIT variances and the conclusions we may draw about calibration from these estimates. This is because a variance estimate from a sample of positively serially correlated PITs is biased downward. In our context, however, because the impact of autoregressive PITs is likely to be similar for the average probability and quantile forecasts, this issue is less likely to confound our comparison of these average forecasts.

Both the average probability forecast and the average quantile forecast have smaller PIT variances than the experts' average variances, indicating less overconfidence or greater underconfidence. $\hat{Q}^{-1}$'s PIT variance is greater than $\hat{F}$'s PIT variance in seven of the eight quarters, with $\hat{Q}^{-1}$ exhibiting greater overconfidence (less underconfidence) than $\hat{F}$ for GDP growth (inflation). It is not surprising that $\hat{Q}^{-1}$ performs better than $\hat{F}$ in terms of the scoring rule results in Tables 1 and 2 for inflation, where $\hat{Q}^{-1}$ has better calibration in the form of less underconfidence. However, note that $\hat{Q}^{-1}$ also generally

Table 3 Minimum, Average, and Maximum Number of Experts (k) Considered in the Sample

	SPF GDP growth (1982–2009)				SPF inflation (1968–2010)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Min k	14	7	13	12	14	7	13	12
Average k	31	33	30	31	36	37	34	32
Max k	49	49	52	47	63	67	103	47

Table 4 **Average Means and Variances of the Probability Integral Transforms (PITs) of the 20 Experts Who Submitted at Least 10 Forecasts in All Four Quarters; Means and Variances of the PITs of the Average Forecasts; Estimated Variance of $\hat{F}$'s PIT**

	SPF GDP growth (1982–2009)				SPF inflation (1968–2010)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
PIT means								
$\hat{F}$ (and experts)	0.573	0.573	0.584	0.572	0.492	0.449	0.435	0.422
$\hat{Q}^{-1}$	0.581	0.577	0.589	0.574	0.485	0.446	0.435	0.413
PIT variances								
Experts	0.122	0.107	0.112	0.128	0.081	0.068	0.058	0.072
$\hat{F}$	0.106	0.090	0.090	0.114	0.064	0.055	0.040	0.047
$\hat{Q}^{-1}$	0.117	0.095	0.092	0.122	0.071	0.059	0.042	0.044
$\hat{F}$ (estimated)	0.098	0.073	0.103	0.113	0.053	0.030	0.038	0.056

Table 5 **Average and Range of Pairwise Correlations of Five Experts' Probability Integral Transforms**

	SPF GDP growth (1982–2009)	SPF inflation (1968–2010)
Q1	0.80 (0.53–0.94)	0.65 (0.34–0.80)
Q2	0.67 (0.11–0.98)	0.43 (0.07–0.76)
Q3	0.92 (0.85–0.97)	0.65 (0.41–0.95)
Q4	0.88 (0.11–0.99)	0.77 (0.60–0.93)

Notes. The five experts were those who jointly participated in the most years. This amounts to 8 years for Q1 GDP growth and inflation and Q2 GDP growth, 9 years for Q3 GDP growth and inflation, 11 years for Q2 and Q4 inflation, and 12 years of Q4 GDP growth.

performs better than $\hat{F}$ on the scoring rules for GDP growth despite the poorer calibration in the form of greater overconfidence.

The average values in Tables 3–5 provide estimates of k, v, and ρ (e.g., for GDP growth in Q1, the averages are $k = 31$, $v = 0.122$, and $\rho = 0.80$). We then plugged those averages into $V[\hat{F}(x)] = (\rho + (1-\rho)/k)v$ from Proposition 4 to get the estimated variances given in Table 4. There is considerable variation in the values of k, v, and ρ (e.g., the average ρ in Table 5 for Q3 on inflation is 0.65, but individual ρ values range from 0.41 to 0.95). Given this sort of variation, as compared with the assumption in Proposition 4 of common values for k, v, and ρ, the correspondence between the estimated and actual PIT variances for $\hat{F}$ seems quite good.

4.3. Sharpness

Table 6 pertains to results in §3.4. Consistent with Proposition 7, the means of $\hat{F}$ and $\hat{Q}^{-1}$ are both equal to $\hat{\mu}$. Moreover, $\hat{F}$'s variance of x is greater than $\hat{Q}^{-1}$'s in all of the eight quarters. This shows that $\hat{Q}^{-1}$ is sharper than $\hat{F}$, as Proposition 8 indicates, and helps explain why $\hat{Q}^{-1}$'s PIT variance is greater than $\hat{F}$'s PIT variance. Note that each of the variances in Table 6 gets smaller as we move from Q1 to Q4 (i.e., as the lead time decreases), as might be expected.

4.4. Shape

To understand better how shape might explain the scores reported in Table 1, we examine the relationship between location disagreement and shape. For the purpose of this examination, we define the head, shoulders, and tails regions for each quantity according to the partition $a_1 = \hat{F}^{-1}(0.05)$, $a_2 = \hat{F}^{-1}(0.25)$, $a_3 = \hat{F}^{-1}(0.75)$, and $a_4 = \hat{F}^{-1}(0.95)$. Thus, the head is over $\hat{F}$'s central 50% prediction interval and the shoulders are over $\hat{F}$'s central 90% prediction interval outside of the head's central prediction interval. This choice follows Gneiting et al. (2007), who used the widths of a forecast's central 50% and 90% prediction intervals to assess sharpness.

As shown by Propositions 9 and 10 (11 and 12), location agreement (disagreement) implies that the density of the average probability (quantile) forecast

Table 6 **Average Means and Variances of x for the Individual Forecasts and the Average Forecasts**

	SPF GDP growth (1982–2009)				SPF inflation (1968–2010)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
$\hat{\mu}$	2.266	2.656	2.668	2.712	4.072	4.033	4.009	3.453
$\hat{\sigma}^2$	1.146	1.018	0.779	0.475	0.884	0.811	0.694	0.493
$\frac{1}{k}\sum_{i=1}^{k}\sigma_i^2$	1.335	1.195	0.925	0.591	1.048	0.970	0.831	0.609
$E_{\hat{F}}[(x-\hat{\mu})^2]$	1.830	1.455	1.168	0.811	1.515	1.390	1.200	0.849
$E_{\hat{Q}^{-1}}[(x-\hat{\mu})^2]$	1.171	1.032	0.796	0.491	0.902	0.827	0.710	0.514

Table 7 **Given Location Agreement ($\hat{q}^{-1}(\hat{\mu})/\hat{f}(\hat{\mu}) < 1$), Average Ratio of Densities at the Mean, Probability Integral Transform Variances for $\hat{F}$ and $\hat{Q}^{-1}$, Total Counts, Percentage of Realizations in the Head and Shoulders, and Average Differences in Scores**

	SPF GDP growth (1982–2009)				SPF inflation (1968–2010)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
$\hat{q}^{-1}(\hat{\mu})/\hat{f}(\hat{\mu})$	0.933	0.952	0.939	0.948	0.962	0.928	0.950	0.964
$V[\hat{F}(x)]$	0.110	0.104	0.099	0.127	0.069	0.057	0.043	0.056
$V[\hat{Q}^{-1}(x)]$	0.115	0.106	0.0978	0.133	0.071	0.055	0.037	0.051
Count	6	16	19	19	13	17	22	18
% in the head	33.3	25.0	15.8	26.3	38.5	41.2	59.1	55.6
% in the shoulders	50.0	75.0	79.0	52.6	53.9	58.8	40.9	44.4
$S(\hat{Q}^{-1}, x_0) - S(\hat{F}, x_0)$								
Lin S	0.050	0.031	0.003	0.018	0.065	0.021	0.006	0.035
Log S	0.270	0.270	0.155	$-\infty$	0.273	0.265	0.184	0.291
Quad S	0.103	0.083	0.047	0.101	0.134	0.059	0.038	0.110
CRP S	−0.012	0.001	−0.003	0.008	0.010	−0.005	0.009	0.015

is higher in the head. We identify location agreement (disagreement) using the ratio $\hat{q}^{-1}(\hat{\mu})/\hat{f}(\hat{\mu})$. When this ratio is less (greater) than 1, we say there is location agreement (disagreement). This categorization presumes $\hat{\mu}$ is in the head, which we do find holds in each year, quarter, and quantity in our study of the SPF data.

When there is location agreement, $\hat{F}$ will tend to be overconfident and realizations of x will tend to fall disproportionately in the shoulders and tails. More realizations in the shoulders benefit the average quantile forecast in such cases because $\hat{q}^{-1}$ is higher in the shoulders and hence scores better. Of course, if $\hat{F}$ is too overconfident, then many realizations of x will fall in the tails and benefit the average probability forecast. When there is location disagreement, $\hat{F}$ will tend to be underconfident, and the realizations of x will tend to fall disproportionately in the head. More realizations in the head benefit the average quantile forecast in such cases because $\hat{q}^{-1}$ is higher in the head, and hence scores better. Tables 7 and 8 offer support for these relationships. When there is location agreement and the average quantile forecast tends to have a higher density in the shoulders, more than the expected 40% of realizations of GDP growth and inflation fall in the shoulders in all eight quarters. Conversely, when there is location disagreement and the average quantile forecast tends to have a higher density in the head, more than the expected 50% of realizations fall in the head in six of the eight quarters.

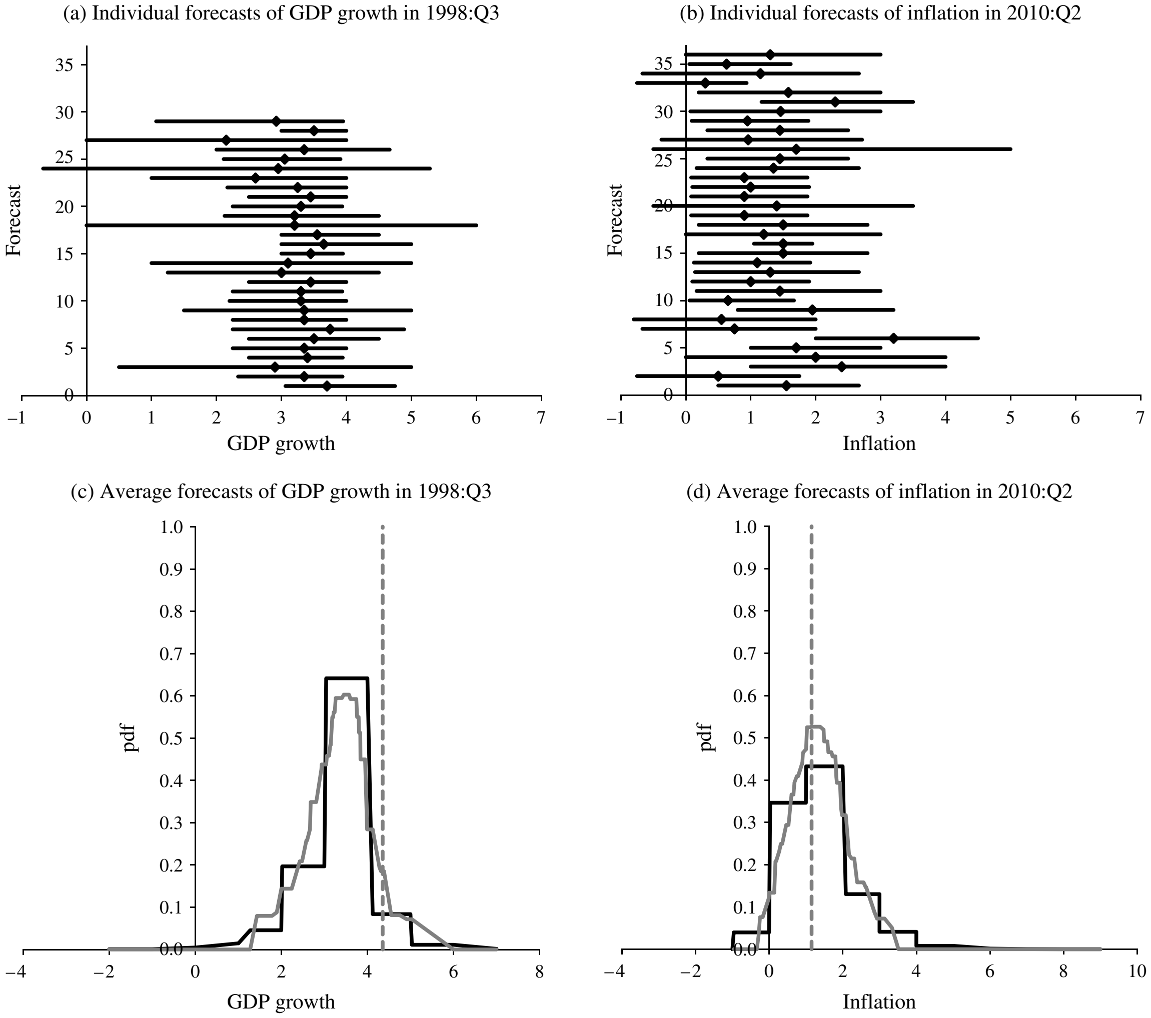
Figure 6 provides additional evidence for why the average quantile forecast scores better, on average, than the average probability forecast in our study. Figures 6(a) and 6(b) consider each individual forecast of GDP growth in 1998:Q3 and inflation in 2010:Q2, respectively. Each line goes from the forecast's 0.05 quantile to its 0.95 quantile, and the point is the forecast's mean of x. In Figure 6(a), we see more location agreement than in Figure 6(b). This could be due in part to the greater volatility in inflation than in GDP growth. The variance of

Table 8 **Given Location Disagreement ($\hat{q}^{-1}(\hat{\mu})/\hat{f}(\hat{\mu}) > 1$), Average Ratio of Densities in the Head, Probability Integral Transform Variances for $\hat{F}$ and $\hat{Q}^{-1}$, Total Counts, Percentage of Realizations in the Head and Shoulders, and Average Differences in Scores**

	SPF GDP growth (1982–2009)				SPF inflation (1968–2010)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
$\hat{q}^{-1}(\hat{\mu})/\hat{f}(\hat{\mu})$	1.162	1.089	1.089	1.161	1.133	1.160	1.196	1.220
$V[\hat{F}(x)]$	0.104	0.068	0.067	0.084	0.059	0.048	0.030	0.034
$V[\hat{Q}^{-1}(x)]$	0.117	0.076	0.0765	0.099	0.070	0.058	0.041	0.033
Count	20	11	9	9	27	24	18	14
% in the head	40.0	54.6	55.6	44.4	55.6	66.7	88.9	71.4
% in the shoulders	40.0	45.5	33.3	55.6	44.4	33.3	11.1	28.6
$S(\hat{Q}^{-1}, x_0) - S(\hat{F}, x_0)$								
Lin S	0.008	−0.001	0.019	0.020	0.022	0.028	0.027	0.026
Log S	NA[a]	0.016	0.175	0.146	0.137	0.098	0.067	0.049
Quad S	−0.008	−0.011	0.033	0.018	0.011	0.033	0.021	0.036
CRP S	−0.030	−0.004	0.011	0.010	0.009	0.016	0.017	0.013

[a] The NA stems from taking the difference between two average scores of $-\infty$.

Figure 6 Location Agreement ((a) and (c)) and Disagreement ((b) and (d)), Where $\hat{f}$ Is Shown in Black, $\hat{q}^{-1}$ Is Shown in Gray, and the Realization Is Marked with a Dashed Gray Line


(a) Individual forecasts of GDP growth in 1998:Q3

(b) Individual forecasts of inflation in 2010:Q2

(c) Average forecasts of GDP growth in 1998:Q3

(d) Average forecasts of inflation in 2010:Q2

inflation over 1968–2010 was 5.70%, and the variance of GDP growth over 1982–2009 was 3.14%. For GDP growth in 1998:Q3, $(1/k)\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2 = 0.11$ and $(1/k)\sum_{i=1}^{k}(\sigma_i-\hat{\sigma})^2 = 0.12$. For inflation in 2010:Q2, $(1/k)\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2 = 0.33$ and $(1/k)\cdot\sum_{i=1}^{k}(\sigma_i-\hat{\sigma})^2 = 0.06$. Figures 6(c) and 6(d) depict the densities that result from averaging the individual forecasts. In the case of location agreement (disagreement) in Figures 6(c) (6(d)), the density of the average quantile forecast is higher in the shoulder (head) and the realization fell in the shoulder (head).

Table 9 Average Skewness and Kurtosis of the Experts' Forecasts, $\hat{F}$, and $\hat{Q}^{-1}$

	SPF GDP growth (1982–2009)				SPF inflation (1968–2010)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Skewness								
Experts	−0.162	−0.151	−0.172	−0.083	0.128	0.109	0.070	0.004
$\hat{F}$	−0.313	−0.385	−0.487	−0.372	0.507	0.525	0.495	0.559
$\hat{Q}^{-1}$	−0.178	−0.173	−0.176	−0.085	0.157	0.127	0.096	0.028
Kurtosis								
Experts	3.132	3.187	3.218	3.119	3.035	3.073	3.118	3.191
$\hat{F}$	4.434	5.050	5.892	6.809	5.006	5.907	5.577	6.083
$\hat{Q}^{-1}$	2.883	2.843	2.818	2.670	2.762	2.731	2.691	2.607

Table 9 gives the average skewness and kurtosis for the forecasts. On average, the individual experts' distributions are slightly skewed to the left for the GDP growth forecasts and slightly skewed to the right for the inflation forecasts, with average kurtosis values slightly greater than 3. This suggests that the normal distribution used in Propositions 10 and 12 might not be a bad approximation in this situation. Compared to the average individual forecast, the skewness is in the same direction, slightly greater for $\hat{Q}^{-1}$ and considerably greater for $\hat{F}$. Finally, the average kurtosis in Table 9 is considerably greater than 3 for $\hat{F}$ and slightly less than 3 for $\hat{Q}^{-1}$, providing further evidence supporting Propositions 10 and 12.

5. Summary and Discussion

Our results indicate that averaging quantiles is as good as or better than averaging probabilities as a method for aggregating probability forecasts. We show that both the average quantile forecast and the average probability forecast have as their mean of x the simple average of the forecasters' means of x. Another key analytical result shows that the average quantile forecast is always sharper (has lower variance of x) than the average probability forecast. Moreover, in combining the individual forecasts of GDP growth and inflation from the SPF, the average quantile forecast outperforms the average probability forecast on our scoring rules. This empirical result is due, in part, to the fact that the average probability forecast is often underconfident and the average quantile forecast is sharper.

Even when the average probability forecast is overconfident, we find that the sharper average quantile forecast might still score better, in part because the average quantile forecast often has a higher density in the shoulders. In these cases, we trace the overconfidence in the average probability forecast back to overconfident experts who agree on the mean of the underlying quantity. Conversely, when experts disagree on the mean, the average probability forecast is underconfident because of the spread driven by the disagreement.

We show theoretically that the average probability forecast scores no worse (usually better) than the average score of the individual experts' forecasts on our scoring rules and provide conditions under which the same is true for the average quantile forecast on the quadratic score. We also demonstrate empirically that both average forecasts outperform the average score of the experts' forecasts in the SPF and that the average quantile forecast is the wiser of the two. These findings amount to a wisdom-of-crowds effect in the context of probability forecasting.

One natural question that arises is how averaging quantiles compares to other aggregation techniques such as weighted averaging of probabilities or logarithmic pooling. In this paper, we have not focused on other techniques for two reasons. First, we seek an aggregation method that is both simple and robust. Averaging probabilities, the most common aggregation method, is simple and quite robust (in comparison with weighted averages, for instance). Particularly in the case of subjective probability forecasts from experts with little past data, the determination of weights is problematic. The average quantile forecast is also simple to apply, and with the equal weighting, should be quite robust too. We do not claim that averaging quantiles is the "best" method in general, but the evidence we present suggests that it can hold its own against averaging probabilities. Second, although other methods may have attractive properties, they often suffer from an important shortcoming. Their aggregate distributions do not always have as their mean of x the simple average of the forecasters' means of x. For instance, in the case of the logarithmic opinion pool with equal weighting of experts' normal densities, its aggregate distribution's mean of x is a weighted average of the forecasters' means of x, with higher weights on the forecasters who are most overconfident (Wallis 2011). Thus, neither logarithmic pooling nor performance-weighted averaging of probabilities can maintain the central finding from the empirical literature on aggregating point forecasts. This finding holds that the simple average of the forecasters' point estimates (or means) is a difficult aggregate point forecast to beat. Many refer to this as the basic wisdom-of-crowds effect, and it is this point that provides additional motivation for focusing on the two methods we consider in this paper.

With the SPF data, we find that $\hat{Q}^{-1}$ does not always perform better than $\hat{F}$. Although it outperforms $\hat{F}$ overall, it can be either too sharp or not sharp enough in some cases. Just as with elicitation, where it can be useful to use two approaches because they cause an expert to think about the situation in different ways, we might want to use both $\hat{Q}^{-1}$ and $\hat{F}$, which are easy to compute once we have the individual distributions. Comparing them might provide additional insight. Moreover, we could combine them by using a vertical average: $B(x) = w_B\hat{F}(x) + (1 - w_B)\hat{Q}^{-1}(x)$. In some situations, such a blend could capitalize on the fact that $\hat{F}$ is more spread out than $\hat{Q}^{-1}$. The sharpness of B lies between the sharpness of $\hat{Q}^{-1}$ and $\hat{F}$. (See Proposition 13 in the appendix for a proof of this statement.) Another way to increase (or decrease) the sharpness of $\hat{Q}^{-1}$ is to horizontally average the average point forecast and the average quantile forecast: $C(u) = w_C\hat{\mu} + (1 - w_C)\hat{Q}(u)$. This mixture shrinks (spreads) the average quantile forecast toward (away from) the average point forecast and is sharper (less sharp) than the average quantile forecast when $0 < w_C < 1$ ($w_C < 0$). (See Proposition 14 in the appendix for a proof of the latter statement.) The inverse of $C(u)$, $C^{-1}(x)$, is a proper cumulative distribution function when $w_C < 1$. Other blends might work well, too, such as $D(x) = w_D C^{-1}(x) + (1 - w_D)\hat{F}(x)$. This type of mixture can provide more sharpness than the average probability forecast and, at the same time, protect against an outlier.

Finally, our analytical and empirical results have important implications for organizations and the way they combine expert opinions. One such organization is the Philadelphia Fed itself. Each quarter they make public the average probability forecast for annual U.S. GDP growth and inflation from the SPF. These forecasts are picked up by the press and reported widely. Businesses and governments around the world use these forecasts to make decisions about investments and policy. Our results suggest that organizations such as the Philadelphia Fed should report the average quantile forecast in place of or in addition to the average probability forecast. In other cases as well, decision makers might improve the quality of their decisions by averaging experts' quantiles instead of averaging experts' probabilities. We feel that averaging quantiles deserves serious consideration for being added to the mix of methods typically used to combine probability forecasts.

Acknowledgments

The authors are grateful to the reviewers for helpful suggestions.

Appendix. Proofs of Main Results

Proof of Proposition 1. For the linear score, $\operatorname{Lin} S(\hat{F}, x_0) = (1/k)\sum_{i=1}^k f_i(x_0) = (1/k)\sum_{i=1}^k \operatorname{Lin} S(F_i, x_0)$. For the logarithmic score, $\operatorname{Log} S(\hat{F}, x_0) = \log((1/k)\sum_{i=1}^k f_i(x_0)) \geq (1/k)\sum_{i=1}^k \log(f_i(x_0)) = (1/k)\sum_{i=1}^k \operatorname{Log} S(F_i, x_0)$ by Jensen's inequality because $\log(f)$ is concave in f. The inequality is strict if $f_i(x_0) \neq f_j(x_0)$ for some $i \neq j$ The result for the quadratic and continuous ranked probability scores follows similarly because the functions f, $-f^2$, and $-(1-f)^2$ are concave in f. □

Definition (Mean-Preserving Spread; Müller and Stoyan 2002, Definition 1.5.25). A cdf G differs from cdf F by a mean-preserving spread if they have the same finite mean and if there is an interval (a, b) such that G assigns no greater probability than F to any open subinterval of (a, b), and G assigns no smaller probability than F to any open interval either to the left or to the right of (a, b).

Proof of Proposition 2. By Thomas and Ross (1980), $\hat{q}^{-1}(x) = (1/\hat{\sigma})g((x-\hat{\mu})/\hat{\sigma})$ and $\operatorname{Quad} S(\hat{Q}^{-1}, x_0) = 2\hat{q}^{-1}(x_0) - (1/\hat{\sigma}^2)\int_{-\infty}^{\infty} g((x-\hat{\mu})/\hat{\sigma})^2\,dx$. Integration by substitution yields $\operatorname{Quad} S(\hat{Q}^{-1}, x_0) = 2\hat{q}^{-1}(x_0) - (1/\hat{\sigma}^2) \cdot \int_{-\infty}^{\infty} g(z)^2 \hat{\sigma}\,dz = 2\hat{q}^{-1}(x_0) - c/\hat{\sigma}$. Also, $(1/k)\sum_{i=1}^k \operatorname{Quad} S(F_i, x_0) = 2\hat{f}(x_0) - (1/k)\sum_{i=1}^k (1/\sigma_i^2)\int_{-\infty}^{\infty} g((x-\mu_i)/\sigma_i)^2\,dx = 2\hat{f}(x_0) - (1/k)\sum_{i=1}^k (c/\sigma_i)$, from which (i) follows immediately. For (ii), $(1/\hat{\sigma}) \leq (1/k)\sum_{i=1}^k (1/\sigma_i)$ by Jensen's inequality, and the gap between the arithmetic average and the harmonic average increases with a mean-preserving spread in the values being averaged (Mitchell 2004). □

We use the following definitions and theorems for cdfs F and G in several subsequent proofs.

Definition (Less Dangerous; Müller and Stoyan 2002, Definition 1.5.16). F is less dangerous than G if there is some t_0 such that $F(t) \leq (\geq) G(t)$ for all $t < (\geq) t_0$, and $\int_{-\infty}^{\infty} t\,dF(t) \leq \int_{-\infty}^{\infty} t\,dG(t)$.

Definition (Convex Order; Müller and Stoyan 2002, Definition 1.5.1). F is less than G in (increasing) convex order if $\int_{-\infty}^{\infty} \varphi(t)\,dF(t) \leq \int_{-\infty}^{\infty} \varphi(t)\,dG(t)$ for all (increasing) convex functions φ such that the expectations exist.

Theorem 1.5.17 (Müller and Stoyan 2002, p. 23). *F is less dangerous than G implies that F is less than G in increasing convex order.*

Theorem 1.5.3 (Müller and Stoyan 2002, p. 17). *Statements* (i) *and* (ii) *are equivalent*: (i) *F is less than G in convex order;* (ii) *F is less than G in increasing convex order and $\int_{-\infty}^{\infty} t\,dF(t) = \int_{-\infty}^{\infty} t\,dG(t)$.*

Corollary 1.5.4 (a) (Müller and Stoyan 2002, p. 17). *If F is less than G in convex order, then F's even central moments of x are less than or equal to G's* (*e.g., F's variance of x is less than or equal to G's*).

Proof of Proposition 3. The perfect calibration case follows directly from the variance of the uniform distribution. We prove the overconfident case; the underconfident case follows from a similar argument. Denote the cdf of a uniform random variable on the unit interval by $F_{U(0,1)}$. F being overconfident with $E[F(x)] = 1/2$ is equivalent to $F_{U(0,1)}$ being less dangerous than the cdf of F's PIT, which by Theorem 1.5.17 implies that $F_{U(0,1)}$ is less than the cdf of F's PIT in increasing convex order. By Theorem 1.5.3, that and the assumption that $E[F(x)] = 1/2$, imply that $F_{U(0,1)}$ is less than the cdf of F's PIT in convex order, which by Corollary 1.5.4 implies that $V[F(x)] \geq \int_0^1 (u - 1/2)^2\,dF_{U(0,1)}(u) = 1/12$. □

Proof of Proposition 4. $V[\hat{F}(x)] = (1/k^2)(kV[F_i(x)] + k(k-1)\operatorname{Cov}(F_i(x)F_j(x))) = (1/k^2)(kV[F_i(x)] + k(k-1)\rho \cdot V[F_i(x)]) = (\rho + (1-\rho)/k)v$ for $i \neq j$. □

The following lemma is used in the proof of Proposition 5.

Lemma 1. *The average quantile forecast is less than the average probability forecast in convex order.*

Proof. If φ is a convex function, then

$$\begin{aligned}\int_{-\infty}^{\infty} \varphi(x)\,d\hat{Q}^{-1}(x) &= \int_0^1 \varphi(\hat{Q}(u))\,du \\ &\leq \frac{1}{k}\sum_{i=1}^k \int_0^1 \varphi(Q_i(u))\,du \\ &= \frac{1}{k}\sum_{i=1}^k \int_{-\infty}^{\infty} \varphi(x)\,dF_i(x) = \int_{-\infty}^{\infty} \varphi(x)\,d\hat{F}(x).\end{aligned}$$

The first equality follows from integration by substitution according to $u = \hat{Q}^{-1}(x)$ and from Jensen's inequality. The second equality follows from integration by substitution according to $u = F_i(x)$ for $i = 1, \dots, k$. The third equality follows from the definition of the average probability forecast. □

Proof of Proposition 5. For $x \le \mu_1$, $\hat{F}(x) = (1/k) \cdot \sum_{i=1}^{k} F((x-\mu_i)/\hat{\sigma}) > F((x-\hat{\mu})/\hat{\sigma}) = \hat{Q}^{-1}(x)$ by the convexity of F below 0 and Jensen's inequality. F is convex below 0 because f' is positive below 0, which follows from the assumption that F is symmetric about 0 and unimodal. For $\mu_j \le x \le \hat{\mu}$ where $1 \le j \le k/2$,

$$\begin{aligned}\hat{F}(x) &= \frac{2}{k}\sum_{i=1}^{j}\left[\frac{1}{2}F\left(\frac{x-\mu_i}{\hat{\sigma}}\right) + \frac{1}{2}F\left(\frac{x-\mu_{k-i+1}}{\hat{\sigma}}\right)\right] \\ &\quad + \frac{1}{k}\sum_{i=j+1}^{k-j} F\left(\frac{x-\mu_i}{\hat{\sigma}}\right) \\ &> \frac{2}{k}\sum_{i=1}^{j}\left[(1-w_i)F(0) + w_i F\left(\frac{x-\mu_{k-i+1}}{\hat{\sigma}}\right)\right] \\ &\quad + \frac{1}{k}\sum_{i=j+1}^{k-j} F\left(\frac{x-\mu_i}{\hat{\sigma}}\right) > F\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right) = \hat{Q}^{-1}(x),\end{aligned}$$

where $w_i = (x-\hat{\mu})/(x-\mu_{k-i+1})$. The first inequality follows because, for $i \le j$,

$$\begin{aligned}&\frac{1}{2}F\left(\frac{x-\mu_i}{\hat{\sigma}}\right) + \frac{1}{2}F\left(\frac{x-\mu_{k-i+1}}{\hat{\sigma}}\right) \\ &\qquad > (1-w_i)F(0) + w_i F\left(\frac{x-\mu_{k-i+1}}{\hat{\sigma}}\right) \\ \Leftrightarrow\;& \left(\frac{1}{2}-w_i\right)F\left(\frac{x-\mu_{k-i+1}}{\hat{\sigma}}\right) \\ &\qquad > (1-w_i)F(0) - \frac{1}{2}F\left(\frac{x-\mu_i}{\hat{\sigma}}\right) \\ \Leftrightarrow\;& \left(\frac{1}{2}-w_i\right)F\left(\frac{x-\mu_{k-i+1}}{\hat{\sigma}}\right) \\ &\qquad > (1-w_i)F(0) - \frac{1}{2}\left(1 - F\left(\frac{-(x-\mu_i)}{\hat{\sigma}}\right)\right) \\ \Leftrightarrow\;& \left(\frac{1}{2}-w_i\right)F\left(\frac{x-\mu_{k-i+1}}{\hat{\sigma}}\right) + w_i F(0) > \frac{1}{2}F\left(\frac{-(x-\mu_i)}{\hat{\sigma}}\right) \\ \Leftrightarrow\;& (1-2w_i)F\left(\frac{x-\mu_{k-i+1}}{\hat{\sigma}}\right) + 2w_i F(0) > F\left(\frac{-(x-\mu_i)}{\hat{\sigma}}\right).\end{aligned}$$

The second implication follows because F is symmetric about 0, and the last implication follows by the convexity of F below 0. The second inequality above follows by the convexity of F below 0 and Jensen's inequality. For the cases where $\hat{\mu} \le x \le \mu_j$ for $k/2 \le j \le k$ and $\mu_k \le x$, the arguments follow similarly using the concavity of F above 0.

Finally, $\hat{F}$, a mixture of distributions with locations that fall symmetrically about $\hat{\mu}$, is symmetric about its median $\hat{\mu}$. For k even, we have

$$\begin{aligned}&\hat{F}(\hat{\mu}+\varepsilon) - \frac{1}{2} \\ &= \frac{1}{k}\sum_{i=1}^{k/2}\left[F\left(\frac{\hat{\mu}+\varepsilon-\mu_i}{\hat{\sigma}}\right) + F\left(\frac{\hat{\mu}+\varepsilon-\mu_{k-i+1}}{\hat{\sigma}}\right)\right] - \frac{1}{2} \\ &= \frac{1}{k}\sum_{i=1}^{k/2}\left[1 - F\left(\frac{-(\hat{\mu}+\varepsilon-\mu_i)}{\hat{\sigma}}\right) + 1 - F\left(\frac{-(\hat{\mu}+\varepsilon-\mu_{k-i+1})}{\hat{\sigma}}\right)\right] - \frac{1}{2} \\ &= \frac{1}{2} - \frac{1}{k}\sum_{i=1}^{k/2}\left[F\left(\frac{\hat{\mu}-\varepsilon-\mu_{k-i+1}}{\hat{\sigma}}\right) + F\left(\frac{\hat{\mu}-\varepsilon-\mu_i}{\hat{\sigma}}\right)\right] \\ &= \frac{1}{2} - \hat{F}(\hat{\mu}-\varepsilon),\end{aligned}$$

where the second equality follows because F is symmetric about 0. The case for k odd follows similarly. Because $\hat{Q}^{-1}$ is a member of the same location-scale family as F (Thomas and Ross 1980), it is clearly symmetric about its median $\hat{\mu}$. □

Proof of Proposition 6. $V[\hat{F}(x)] = E[(\hat{F}(x) - 1/2)^2] < E[(\hat{Q}^{-1}(x) - 1/2)^2] = V[\hat{Q}^{-1}(x)]$, where the inequality follows directly from Proposition 5. Let u be a uniform random variable on the unit interval. Then we have $V[\hat{Q}^{-1}(x)] = V[F((x-\hat{\mu})/\hat{\sigma})] = V[F((1/\hat{\sigma})\sigma F^{-1}(u))] < V[F(F^{-1}(u))] = 1/12$. □

Proof of Proposition 7. The result follows from Lemma 1 and Theorem 1.5.3. □

Proof of Proposition 8. Both (i) and (ii) follow from Lemma 1 and Corollary 1.5.4. □

Proof of Proposition 9. The average probability forecast is given by $\hat{F}(x) = (1/k)\sum_{i=1}^{k} F((x-\mu_i)/\sigma_i)$. The average quantile forecast is given by $\hat{Q}(u) = (1/k)\sum_{i=1}^{k}\mu_i + (1/k) \cdot \sum_{i=1}^{k}\sigma_i Q(u)$. Thus, $\hat{f}(x) = (1/k)\sum_{i=1}^{k}(1/\sigma_i) f((x-\mu_i)/\sigma_i)$ and $\hat{q}^{-1}(x) = (1/((1/k)\sum_{i=1}^{k}\sigma_i)) f((x-\hat{\mu})/((1/k)\sum_{i=1}^{k}\sigma_i))$. At $\hat{\mu}$,

$$\hat{f}(\hat{\mu}) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\sigma_i}f(0) > \frac{1}{(1/k)\sum_{i=1}^{k}\sigma_i}f(0) = \hat{q}^{-1}(\hat{\mu})$$

because $(1/k)\sum_{i=1}^{k}(1/\sigma_i)$ is greater than $1/((1/k)\sum_{i=1}^{k}\sigma_i)$ by Jensen's inequality. □

Proof of Proposition 10. For (i), $\hat{f}(x) = (1/k) \cdot \sum_{i=1}^{k}(1/(\sqrt{2\pi}\sigma_i))\exp(-(x-\hat{\mu})^2/(2\sigma_i^2))$ and $\hat{q}^{-1}(x) = (1/(\sqrt{2\pi}\hat{\sigma}))\exp(-(x-\hat{\mu})^2/(2\hat{\sigma}^2))$. Next we consider $G(x) = \hat{f}(x)/\hat{q}^{-1}(x) - 1$ and determine the regions of x for which it falls above and below 0. $G(x) = (1/k)\sum_{i=1}^{k}(\hat{\sigma}/\sigma_i) \cdot \exp(\Delta_i(x-\hat{\mu})^2) - 1$, where $\Delta_i = (1/(2\hat{\sigma}^2) - 1/(2\sigma_i^2))$. Because $G(x)$ is symmetric about $\hat{\mu}$ (i.e., $G(\hat{\mu}+x) = G(\hat{\mu}-x)$), we analyze the function $H(z) = (1/k) \cdot \sum_{i=1}^{k}(\hat{\sigma}/\sigma_i)\exp(\Delta_i z) - 1$, where $z = (x-\hat{\mu})^2$: (i) $H(0) > 0$ by Proposition 9; (ii) $\lim_{z\to\infty} H(z) = \infty$ because at least one σ_i is greater than $\hat{\sigma}$, which implies the corresponding $\Delta_i \ge 0$; and (iii) H is strictly convex on $[0, \infty)$ because $H''(z) = (1/k)\sum_{i=1}^{k}(\hat{\sigma}/\sigma_i)\Delta_i^2\exp(\Delta_i z) > 0$ for all $z \ge 0$. Thus, H crosses 0 twice, first from above and then from below. It must cross at least once, or $\hat{q}^{-1}(x)$ would not be a density. It cannot cross exactly once; otherwise, $\lim_{z\to\infty} H(z) < 0$. It cannot cross more than twice; otherwise, $H(z)$ would not be strictly convex. Thus, G crosses 0 four times, first from above at some $a_1 < a_2 < \hat{\mu}$, then from below at some $a_2 < \hat{\mu}$, then from above at some $a_3 > \hat{\mu}$, and finally from below at some $a_4 > a_3$. For (ii), under location agreement the kurtosis of $\hat{F}$ is $3(1/k)\sum_{i=1}^{k}(\sigma_i^2)^2/((1/k)\sum_{i=1}^{k}\sigma_i^2)^2$

by Proposition 2.2.1 of Wang (2001). By Jensen's inequality with the experts not agreeing on the scale parameter, $(1/k)\sum_{i=1}^{k}(\sigma_i^2)^2 > ((1/k)\sum_{i=1}^{k}\sigma_i^2)^2$ Thus, the kurtosis of $\hat{F}$ is greater than 3. The kurtosis of $\hat{Q}^{-1}$ is 3 when the experts report normal distributions. □

Proof of Proposition 11. Here,

$$\hat{f}(\hat{\mu}) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\sigma_i} f\left(\frac{\hat{\mu}-\mu_i}{\hat{\sigma}}\right) \le \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\sigma_i} f(0)$$

$$< \frac{1}{\hat{\sigma}} f(0) = \hat{q}^{-1}(\hat{\mu})$$

because 0 is the only mode of f and $(1/k)\sum_{i=1}^{k}(1/\sigma_i) > 1/((1/k)\sum_{i=1}^{k}\sigma_i)$ by Jensen's inequality. □

Proof of Proposition 12. For (i), we consider $G(x) = (\hat{f}(x)/\hat{q}^{-1}(x)) - 1$, as in the proof of Proposition 10, and determine the regions of x for which it falls above and below zero:

$$G(x) = \frac{1}{k}\sum_{i=1}^{k}\exp\left(-\frac{(x-\mu_i)^2-(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right) - 1$$

$$= \frac{1}{k}\sum_{i=1}^{k}\exp\left(\frac{2(\mu_i-\hat{\mu})x-\mu_i^2+\hat{\mu}^2}{2\hat{\sigma}^2}\right) - 1.$$

Four facts about G emerge: (i) $G(\hat{\mu}) < 0$ by Proposition 9, (ii) $\lim_{x\to\infty} G(x) = \infty$ because at least one μ_i is greater than $\hat{\mu}$, (iii) $\lim_{x\to-\infty} G(x) = \infty$ because at least one μ_i is less than $\hat{\mu}$, and (iv) G is strictly convex on $(-\infty, \infty)$ because $G''(x) = (1/(k\hat{\sigma}^4))\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2\exp((2(\mu_i-\hat{\mu})x-\mu_i^2+\hat{\mu}^2)/(2\hat{\sigma}^2)) > 0$ for all $x \in (-\infty, \infty)$. Thus, $G(x)$ crosses zero twice, first from above at some $a_1 < \hat{\mu}$, then from below at some $a_2 > \hat{\mu}$. For (ii), $\hat{F}$'s kurtosis is given by

$$3\frac{\hat{\sigma}^4 + 2\hat{\sigma}^2(1/k)\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2 + (1/(3k))\sum_{i=1}^{k}(\mu_i-\hat{\mu})^4}{\hat{\sigma}^4 + 2\hat{\sigma}^2(1/k)\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2 + ((1/k)\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2)^2}.$$

The first two terms in the numerator and denominator are the same. For the third terms,

$$\frac{1}{3k}\sum_{i=1}^{k}(\mu_i-\hat{\mu})^4 \ge \left(\frac{1}{k}\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2\right)^2, \quad \text{or}$$

$$\frac{\frac{1}{k}\sum_{i=1}^{k}(\mu_i-\hat{\mu})^4}{(\frac{1}{k}\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2)^2} \ge 3,$$

and

$$\frac{\frac{1}{k}\sum_{i=1}^{k}(\mu_i-\hat{\mu})^4}{(\frac{1}{k}\sum_{i=1}^{k}(\mu_i-\hat{\mu})^2)^2}$$

is the sample kurtosis in the experts' means of x. □

Proposition 13. (i) *$\hat{Q}^{-1}$ is sharper than the mixture $B(x) = w_B\hat{F}(x) + (1-w_B)\hat{Q}^{-1}(x)$ for $0 < w_B < 1$, and B is sharper than $\hat{F}$. (ii) B's mean of x is $\hat{\mu}$.*

Proof of Proposition 13. The average quantile forecast is less than the mixture B in convex order. For convex φ,

$$\int_{-\infty}^{\infty}\varphi(x)\,d\hat{Q}^{-1}(x)$$

$$= w\int_{-\infty}^{\infty}\varphi(x)\,d\hat{Q}^{-1}(x) + (1-w)\int_{-\infty}^{\infty}\varphi(x)\,d\hat{Q}^{-1}(x)$$

$$= w\int_{0}^{1}\varphi(\hat{Q}(u))\,du + (1-w)\int_{-\infty}^{\infty}\varphi(x)\,d\hat{Q}^{-1}(x)$$

$$\le w\frac{1}{k}\sum_{i=1}^{k}\int_{0}^{1}\varphi(Q_i(u))\,du + (1-w)\int_{-\infty}^{\infty}\varphi(x)\,d\hat{Q}^{-1}(x)$$

$$= w\frac{1}{k}\sum_{i=1}^{k}\int_{-\infty}^{\infty}\varphi(x)\,dF_i(x) + (1-w)\int_{-\infty}^{\infty}\varphi(x)\,d\hat{Q}^{-1}(x)$$

$$= w\int_{-\infty}^{\infty}\varphi(x)\,d\hat{F}(x) + (1-w)\int_{-\infty}^{\infty}\varphi(x)\,d\hat{Q}^{-1}(x)$$

$$= \int_{-\infty}^{\infty}\varphi(x)\,dB(x).$$

Similarly, the mixture B is less than the average probability forecast in convex order. The results follow from these convex order relations, Theorem 1.5.3, and Corollary 1.5.4. □

Proposition 14. (i) *C^{-1} is sharper (less sharp) than the average quantile forecast when $0 < w_C < 1$ ($w_C < 0$). (ii) C^{-1}'s mean of x is $\hat{\mu}$.*

Proof of Proposition 14. When $0 < w_C < 1$,

$$\int_{-\infty}^{\infty}(x-\hat{\mu})^2\,dC^{-1}(x)$$

$$= \int_{0}^{1}(w_C\hat{\mu} + (1-w_C)\hat{Q}(u) - \hat{\mu})^2\,du$$

$$\le w_C\int_{0}^{1}(\hat{\mu}-\hat{\mu})^2\,du + (1-w_C)\int_{0}^{1}(\hat{Q}(u)-\hat{\mu})^2\,du$$

$$\le \int_{0}^{1}(\hat{Q}(u)-\hat{\mu})^2\,du = \int_{-\infty}^{\infty}(x-\hat{\mu})^2\,d\hat{Q}^{-1}(x).$$

The first inequality follows from Jensen's inequality because $(x-\hat{\mu})^2$ is convex in x. When $w_C < 0$,

$$\int_{-\infty}^{\infty}(x-\hat{\mu})^2\,dC^{-1}(x)$$

$$= \int_{0}^{1}(w_C\hat{\mu} + (1-w_C)\hat{Q}(u) - \hat{\mu})^2\,du$$

$$= \int_{0}^{1}(\hat{Q}(u) - (w_C\hat{Q}(u) + (1-w_C)\hat{\mu}))^2\,du$$

$$\ge w_C\int_{0}^{1}(\hat{Q}(u)-\hat{Q}(u))^2\,du + (1-w_C)\int_{0}^{1}(\hat{Q}(u)-\hat{\mu})^2\,du$$

$$\ge \int_{0}^{1}(\hat{Q}(u)-\hat{\mu})^2\,du = \int_{-\infty}^{\infty}(x-\hat{\mu})^2\,d\hat{Q}^{-1}(x).$$

The first inequality follows from Jensen's inequality because $(\hat{Q}(u)-x)^2$ is concave in x. C^{-1}'s mean of x is $w_C\hat{\mu} + (1-w_C)\hat{\mu} = \hat{\mu}$. □

References

Abbas AE, Budescu DV, Yu H-T, Haggerty R (2008) A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Anal.* 5:190–202.

Armstrong SJ, ed. (2001) Combining forecasts. *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Kluwer Academic, Norwell, MA), 417–439.

Balanda KP, MacGillivray HL (1988) Kurtosis: A critical review. *Amer. Statistician* 42:111–119.

Budescu DV, Du N (2007) Coherence and consistency of investors' probability judgments. *Management Sci.* 53:1731–1744.

Casella G, Berger RL (2002) *Statistical Inference*, 2nd ed. (Duxbury, Pacific Grove, CA).

Clemen RT, Winkler RL (1986) Combining economic forecasts. *J. Bus. Econom. Statist.* 4:39–46.

Clements MP (2010) Explanations of the inconsistencies in survey respondents' forecasts. *Eur. Econom. Rev.* 54:536–549.

Croushore D (1993) Introducing: The survey of professional forecasters. *Federal Reserve Bank of Philadelphia Bus. Rev.* (November/December):3–13.

Dawid AP (1984) Statistical theory: The prequential approach (with discussion). *J. Roy. Statist. Soc. A* 147:278–292.

Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts, with applications to financial risk management. *Internat. Econom. Rev.* 39:863–883.

Diebold FX, Tay AS, Wallis KF (1999) Evaluating density forecasts of inflation: The survey of profession of forecasters. Engle R, White H, eds. *Festschrift in Honor of C.W.J. Granger* (Oxford University Press, Oxford, UK), 76–90.

Engelberg J, Manski CF, Williams J (2009) Comparing the point predictions and subjective probability distributions of professional forecasters. *J. Bus. Econom. Statist.* 27:30–41.

Fildes R, Makridakis S (1995) The impact of empirical accuracy studies on time series analysis and forecasting. *Internat. Statist. Rev.* 63:289–308.

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* 102:359–378.

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

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

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

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Sci.* 52:111–127.

Larrick RP, Mannes AE, Soll JB (2011) The social psychology of the wisdom of crowds. Krueger JI, ed. *Frontiers in Social Psychology: Social Judgment and Decision Making* (Psychology Press, New York), 227–242.

Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. *Management Sci.* 22:1087–1096.

Mitchell DW (2004) More on spreads and non-arithmetic means. *Math. Gazette* 88:142–144.

Mitchell J, Wallis KF (2011) Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *J. Appl. Econometrics* 26:1023–1040.

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

Müller A, Stoyan D (2002) *Comparison Methods for Stochastic Models and Risks* (John Wiley & Sons, Chichester, UK).

O'Hagan AO, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) *Uncertain Judgements: Eliciting Experts' Probabilities* (John Wiley & Sons, Chichester, UK).

Ranjan R (2009) Combining and evaluating probabilistic forecasts. Ph.D. dissertation, University of Washington, Seattle.

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

Ratcliff R (1979) Group reaction time distributions and an analysis of distribution statistics. *Psych. Bull.* 86:446–461.

Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* 12:1151–1172.

Shuford EH, Albert A, Massengill HE (1966) Admissible probability measurement procedures. *Psychometrika* 31:125–145.

Smith JQ (1985) Diagnostic checks of non-standard time series models. *J. Forecasting* 4:283–291.

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

Thomas EAC, Ross BH (1980) On appropriate procedures for combining probability distributions with the same family. *J. Math. Psych.* 21:136–152.

Wallis KF (2011) Combining forecasts—Forty years later. *Appl. Financial Econom.* 21:33–41.

Wang J (2001) Generating daily changes in market variables using a multivariate mixture of normal distributions. Peters BA, Smith JS, Medeiros DJ, Roohrer MW, eds. *Proc. 2001 Winter Simulation Conf.* (ACM, Arlington, VA), 283–289.

Winkler RL (1969) Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* 64:1073–1077.

Winkler RL (1996) Scoring rules and the evaluation of probabilities. *Test* 5:1–60.