Management Science

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

Trimmed Opinion Pools and the Crowd's Calibration Problem

Victor Richmond R. Jose, Yael Grushka-Cockayne, Kenneth C. Lichtendahl Jr.

To cite this article:
Victor Richmond R. Jose, Yael Grushka-Cockayne, Kenneth C. Lichtendahl Jr. (2014) Trimmed Opinion Pools and the Crowd's Calibration Problem. Management Science 60(2):463-475. http://dx.doi.org/10.1287/mnsc.2013.1781

MANAGEMENT SCIENCE
Vol. 60, No. 2, February 2014, pp. 463–475
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2013.1781

Trimmed Opinion Pools and the Crowd's Calibration Problem

Victor Richmond R. Jose
McDonough School of Business, Georgetown University, Washington, DC 20057,
vrj2@georgetown.edu

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

We introduce an alternative to the popular linear opinion pool for combining individual probability forecasts. One of the well-known problems with the linear opinion pool is that it can be poorly calibrated. It tends toward underconfidence as the crowd's diversity increases, i.e., as the variance in the individuals' means increases. To address this calibration problem, we propose the exterior-trimmed opinion pool. To form this pool, forecasts with low and high means, or cumulative distribution function (cdf) values, are trimmed away from a linear opinion pool. Exterior trimming decreases the pool's variance and improves its calibration. A linear opinion pool, however, will remain overconfident when individuals are overconfident and not very diverse. For these situations, we suggest trimming away forecasts with moderate means or cdf values. This interior trimming increases variance and reduces overconfidence. Using probability forecast data from U.S. and European Surveys of Professional Forecasters, we present empirical evidence that trimmed opinion pools can outperform the linear opinion pool.

Keywords: trimming; probability forecasts; expert combination; linear opinion pool; underconfidence; overconfidence; scoring rules; wisdom of crowds; diversity
History: Received November 13, 2012; accepted May 30, 2013, by Rakesh Sarin, decision analysis. Published online in *Articles in Advance* November 18, 2013.

1. Introduction

One of the central findings from the forecast combination literature is the power of simple averaging. In point forecasting, the simple average is a good forecast (Clemen and Winkler 1986, Armstrong 2001). The average point forecast harnesses the wisdom of the crowd: it is more accurate than the average individual and is sometimes better than nearly all individuals (Larrick et al. 2011). The average point forecast also often outperforms more complicated point aggregation schemes (Smith and Wallis 2009).

Many organizations, such as the Federal Reserve Bank of Philadelphia and the European Central Bank, apply the idea of simple averaging to the combination of probability forecasts. Averaging probabilities is the most widely used probability aggregation method (Cooke 1991, Hora 2004, Clemen 2008). Stone (1961) gave the average probability forecast its popular name: the linear opinion pool. O'Hagan et al. (2006, p. 190) describe the linear opinion pool as "hard to beat in practice."

Soll et al. (2013) suggest that there are two primary factors that lead a point-forecasting crowd to be wise: expertise and diversity. Of the two, they emphasize the role of diversity (Soll et al. 2013, p. 777): "Diversity helps because any given perspective is likely to be wrong. People who share a perspective will all be wrong in the same way (e.g., numerical estimates which all over- or underestimate the truth), in which case there is little benefit gained from a crowd. For numerical estimates, the benefit comes when errors 'bracket' the truth and cancel out. Interestingly, diversity is so valuable that one can still benefit from averaging when individuals differ greatly in accuracy."

Although diversity benefits the average point forecast, it can hurt the average probability forecast. As the crowd's diversity (the variance in the individuals' means or point forecasts) increases, the average probability forecast becomes more spread out, or underconfident (Lichtendahl et al. 2013). When the average probability forecast is underconfident, the crowd's probability forecast, as represented by the linear opinion pool, has a calibration problem. Related shortcomings of the linear opinion pool have been studied by Dawid et al. (1995), Hora (2004), and Ranjan and Gneiting (2010). This calibration problem calls into question the power of simple averaging and motivates our search for simple and effective alternatives. In this paper, we introduce such an alternative: trimmed opinion pools. These pools

extend the idea of trimming point forecasts—first proposed by Armstrong (2001) and later studied by Jose and Winkler (2008)—to probability forecasts and represent a methodological contribution to the forecast combination literature.

To address underconfidence in the linear opinion pool, we propose the exterior-trimmed opinion pool. To form this pool, forecasts with low and high means are trimmed away before averaging. Alternatively, we consider trimming away, at each possible realization of the quantity of interest, forecasts with low and high cumulative distribution function (cdf) values. Exterior trimming decreases the pool's variance and improves its calibration. A linear opinion pool, however, will remain overconfident when individuals are overconfident or the crowd is not very diverse. For these situations, we suggest trimming away forecasts with moderate means, or cdf values. This interior trimming increases variance and reduces overconfidence.

In the next three sections, we present analytical results that illustrate how the exterior-trimmed (interior-trimmed) opinion pool can address the crowd's calibration problem by decreasing (increasing) variance relative to the linear opinion pool. In a subsequent section, we present empirical results using probability forecast data from the Federal Reserve Bank of Philadelphia's and European Central Bank's Surveys of Professional Forecasters. With these data on inflation, GDP, and unemployment forecasts and using a scoring rule that penalizes a pool for miscalibration, we find that the exterior-trimmed (interior-trimmed) opinion pool outperforms the linear opinion pool when it is underconfident (overconfident). Finally, we offer some prescriptions for practice and conclusions.

2. Linear Opinion Pool

Suppose a decision analyst elicits k experts' probability forecasts of the uncertain quantity x and combines these forecasts by averaging them. We define expert i's probability forecast as the cdf $F_i(x)$. We often use the same notation for an uncertain quantity x and its realization x, where we let the context indicate which of the two applies. We denote a forecast F's mean of x by $\mu_F = \int_{-\infty}^{\infty} x\, dF(x)$. A forecast F's variance of x is given by $\sigma_F^2 = \int_{-\infty}^{\infty} (x-\mu_F)^2\, dF(x)$. We use shorthand notation for expert i's (F_i's) mean and variance of x, using μ_i and σ_i^2, respectively. For convenience, we assume $\mu_1 \le \cdots \le \mu_k$ so that μ_i is the ith largest mean of x among the experts' means. The average of the experts' means of x is denoted by $\bar{\mu} = (1/k)\sum_{i=1}^{k} \mu_i$. The average (or mean) probability forecast of the k experts' forecasts is given by $\bar{F}(x) = (1/k)\sum_{i=1}^{k} F_i(x)$. This combined forecast is known as the linear opinion pool (Clemen 1989).

The linear opinion pool's mean of x is $\mu_{\bar{F}} = \bar{\mu}$, and its variance of x is $\sigma_{\bar{F}}^2 = (1/k)\sum_{i=1}^{k} \sigma_i^2 + (1/k)\cdot \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2$ (Giordani and Söderlind 2003, Boero et al. 2008). As location diversity $(1/k)\sum_{i=1}^{k} (\mu_i - \bar{\mu})^2$ increases, the linear opinion pool's variance of x increases. Too large a variance can result in an underconfident forecast. The average probability forecast $\bar{F}$ is underconfident (overconfident) if over multiple realizations, more (less) than y% of the realizations fall within $\bar{F}$'s y% central prediction intervals.

The idea that well-calibrated forecasts when linearly pooled can lead to an underconfident forecast is well known (Dawid et al. 1995, Hora 2004, Ranjan and Gneiting 2010, Lichtendahl et al. 2013). Even if all experts agree on the actual variance of x, say σ^2, the linear opinion pool's variance of x will be greater than σ^2 if any two experts disagree on its mean. If the experts are overconfident and not very diverse, however, the linear opinion pool may be overconfident. In a subsequent section, we propose two new types of opinion pools formed by using one of two approaches. Under our first approach, a trimmed opinion pool addresses the crowd's calibration problem by either decreasing or increasing location diversity directly. Our second approach is designed specifically for when location diversity is low.

3. Evaluating an Opinion Pool's Performance

Because it is difficult in practice to say how well calibrated an opinion pool is ex ante (before the realization of x is known), we will evaluate an opinion pool ex post. For this purpose, scoring rules are useful (Winkler 1996, Winkler and Jose 2011). An opinion pool can be evaluated using its average score ex post over multiple realizations.

One of the most popular scoring rules is the quadratic scoring rule. Assume the uncertain quantity x will fall in one of m bins. Each bin is an interval given by $B_j = (x_{j-1}, x_j]$ for $l = 1, \ldots, m$. The quadratic scoring rule takes a forecast F and the realized bin B_{j^*} and assigns the score $S(F, B_{j^*}) = 2(F(x_{j^*}) - F(x_{j^*-1})) - \sum_{j=1}^{m} (F(x_j) - F(x_{j-1}))^2$, which can range from -1 to 1, with higher scores being better. The continuous version of this score is given by $2f(x) - \int_{-\infty}^{\infty} f^2(y)\, dy$, where f is the forecast's probability density function and x is the realization. We focus on the binned version of the quadratic scoring rule to match the format of the Survey of Professional Forecasters analyzed in §5. In that survey, forecasters report probabilities over bins.

Now consider scoring the linear opinion pool on a series of independent and identical draws of x from a distribution G—a situation similar to scoring a time series of one-step-ahead forecasts of, say, inflation. As the number of draws gets large, $\bar{F}$'s average quadratic

score will go to its expected quadratic score under G. As is the case with any proper scoring rule, the expected score can be decomposed into a "penalty for miscalibration" plus a constant (Winkler 1996, p. 11). For the quadratic scoring rule, this decomposition is given by $E_G[(\bar{F}, B_j] = -\sum_{j=1}^{m}(\bar{F}(x_j) - \bar{F}(x_{j-1}) - (G(x_j) - G(x_{j-1})))^2 - \sum_{j=1}^{m}(G(x_j) - G(x_{j-1}))^2$. The quadratic scoring rule's penalty for miscalibration is the sum of the squared differences between the bin probabilities assigned by the opinion pool's forecast and the true distribution.

Any proper scoring rule's penalty for miscalibration is small when the forecast resembles the true distribution. Roughly speaking, this will occur when $\bar{F}$'s mean and variance are close to x's true mean and variance. And because the linear opinion pool's mean of x (and that of our trimmed opinion pools) has been shown to be good point forecast (Clemen and Winkler 1986, Armstrong 2001, Jose and Winkler 2008), our focus will be on a pool's variance of x.

4. Trimmed Opinion Pools

In this section, we introduce two new types of opinion pools: the exterior- and interior-trimmed opinion pools. These pools trim away forecasts from the ends and the middle of a linear opinion pool, respectively; hence, we call them trimmed opinion pools. Exterior trimming results in an inner mean probability forecast, and interior trimming produces an outer mean probability forecast. Properties of the inner mean (also called the trimmed mean) and the outer mean of a set of points are considered in Prescott and Hogg (1977). We extend the concepts of inner and outer means of points to probability forecasts. To trim forecasts, we need to order them. In the next two subsections, we propose two approaches for ordering forecasts: the mean and cdf approaches. For each approach, we illustrate some properties of exterior and interior trimming.

4.1. Mean Approach

Under the mean approach, the exterior-trimmed opinion pool is the simple average of the experts' cdfs after trimming away expert cdfs that have low and high means of x. It is given by $\hat{F}_\alpha(x) = (1/(k-2j)) \cdot \sum_{i=j+1}^{k-j} F_i(x)$, where $j = \lfloor \alpha k \rfloor$ is the greatest integer less than αk, $0 < \alpha < \frac{1}{2}$ is the level of exterior trimming, and $F_i(x)$ is the cdf of the expert with the ith largest mean of x. Any ties are broken uniformly at random. The exterior-trimmed average of the experts' means of x is denoted by $\hat{\mu}_\alpha = (1/(k-2j)) \sum_{i=j+1}^{k-j} \mu_i$, which is $\hat{F}_\alpha$'s mean of x. Empirical evidence suggests that $\hat{\mu}_\alpha$ may be a better point estimate than $\bar{\mu}$ (Jose and Winkler 2008).

Under the mean approach, the interior-trimmed opinion pool is the simple average of the experts' cdfs after trimming away expert cdfs that have neither low nor high means of x. It is given by $\check{F}_\beta(x) = (1/(2j))[\sum_{i=1}^{j} F_i(x) + \sum_{i=k-j+1}^{k} F_i(x)]$, where $j = \lfloor (\frac{1}{2} - \beta)k \rfloor$ and $0 < \beta < \frac{1}{2}$ is the level of interior trimming. The interior-trimmed average of the experts' means of x is denoted by $\check{\mu}_\beta = (1/(2j))[\sum_{i=1}^{j} \mu_i + \sum_{i=k-j+1}^{k} \mu_i]$.

This brings us to our first main result.

Proposition 1. *Under the mean approach, the linear opinion pool's variance of x is a weighted average of the trimmed pools' variances of x plus a weighted average of the square differences between each trimmed pool's mean of x and the linear opinion pool's mean of x: $\sigma^2_{\bar{F}} = w\sigma^2_{\hat{F}_\alpha} + (1-w)\sigma^2_{\check{F}_\beta} + w(\hat{\mu}_\alpha - \bar{\mu})^2 + (1-w)(\check{\mu}_\beta - \bar{\mu})^2$, where $j = \lfloor \alpha k \rfloor$, $w = (k-2j)/k$, and $\beta = \frac{1}{2} - \alpha$.*

This result offers an insight into the case when the trimmed pool's and the linear opinion pool's means of x are equal. This will occur, for instance, when the experts' means of x are symmetrically distributed. In this case, the linear opinion pool's variance is a weighted average of the exterior- and interior-trimmed opinion pools' variances. The weights are strictly between zero and one, and the result implies that when one trimmed pools' variance is greater than the linear opinion pool's variance, the other's is less than the linear opinion pool's variance.

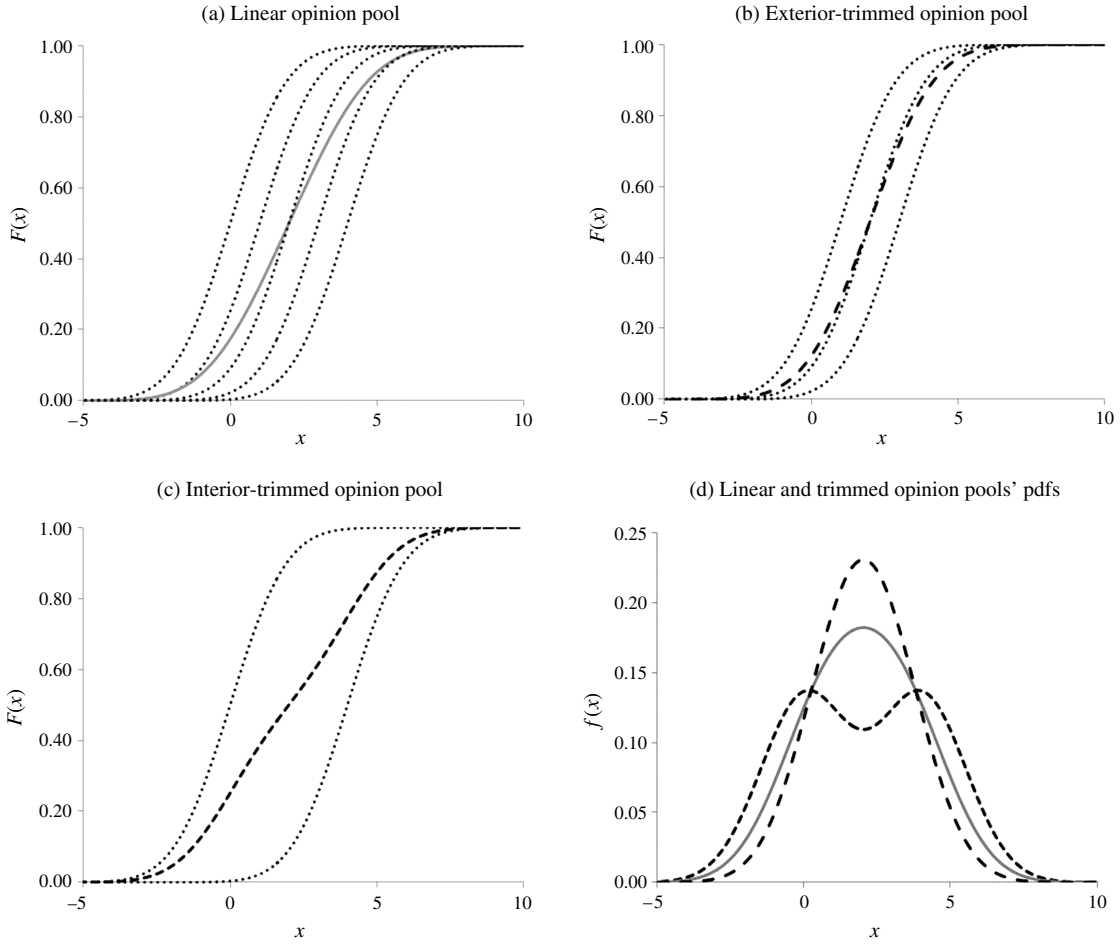
The following result concerns a situation where the exterior-trimmed opinion pool's variance will be the lower one. For this result, we define the exterior- and interior-trimmed averages of the experts' variances. They are $\hat{\sigma}^2_\alpha = (1/(k-2j)) \sum_{i=j+1}^{k-j} \sigma_i^2$ and $\check{\sigma}^2_\beta = (1/(2j)) \cdot [\sum_{i=1}^{j} \sigma_i^2 + \sum_{i=k-j+1}^{k} \sigma_i^2]$, respectively.

Proposition 2. *Under the mean approach, if experts' means are symmetrically distributed around $\bar{\mu}$ with $\hat{\sigma}^2_\alpha = \check{\sigma}^2_\beta$, then $\sigma^2_{\hat{F}_\alpha} \le \sigma^2_{\bar{F}} \le \sigma^2_{\check{F}_\beta}$, where $j = \lfloor \alpha k \rfloor$ and $\beta = \frac{1}{2} - \alpha$.*

The following example illustrates such a variance ordering.

Example 1. Suppose five experts report normal distributions with means of 0, 1, 2, 3, and 4, respectively, and all agree on the standard deviation of 1.5. Figure 1(a) depicts these five experts' forecasts and the linear opinion pool of these forecasts. Figure 1(b) shows the exterior-trimmed opinion pool with $\alpha = 0.2$, a simple average of expert 2, 3, and 4's forecasts. Figure 1(c) shows the interior-trimmed opinion pool with $\beta = 0.3$, a simple average of expert 1 and 5's forecasts. From Figure 1(d), we can see that the exterior-trimmed (interior-trimmed) opinion pool's probability density function (pdf) is narrower (wider) than the linear opinion pool's. As predicted by Proposition 2, we have the variance ordering: $\sigma^2_{\hat{F}_\alpha} = 2.92$, $\sigma^2_{\bar{F}} = 4.25$, and $\sigma^2_{\check{F}_\beta} = 6.25$.

Figure 1 **Five Experts' CDFs (Dotted) and the Linear (Gray Solid), Exterior-Trimmed (Long-Dashed), and Interior-Trimmed (Short-Dashed) Opinion Pools Using the Mean Approach from Example 1**



The variance ordering in Example 1 will typically occur in practice when location diversity's contribution to the linear opinion pool's variance is high. When location diversity's contribution is lower, however, the interior-trimmed opinion pool's variance can be the lowest if the experts with moderate means happen also to be the experts who report higher variances for x. In Example 1, if expert 3 had reported a standard deviation of 5 instead, the variance ordering among the pools would be reversed: $\sigma^2_{\hat{F}_\alpha} = 10.5$, $\sigma^2_{\bar{F}} = 8.8$, and $\sigma^2_{\check{F}_\beta} = 6.25$. Because trimming in environments with low location diversity may not affect an opinion pool's variance in a predictable way, we propose a second approach to trimming forecasts.

4.2. CDF Approach

Under the cdf approach, trimming is done pointwise. The exterior-trimmed opinion pool is the simple average of the experts' cdf values after trimming away the experts' low and high cdfs values at each possible realization of x. It is given by $\hat{F}_\alpha(x) = (1/(k-2j))\sum_{i=j+1}^{k-j} F_{(i)}(x)$, where $j = \lfloor \alpha k \rfloor$, $0 < \alpha < \frac{1}{2}$ is the level of exterior trimming, and $F_{(i)}(x)$ is the ith largest cdf value at the point x. Similarly, under this approach, the interior-trimmed opinion pool is given by $\check{F}_\beta(x) = (1/(2j))[\sum_{i=1}^{j} F_{(i)}(x) + \sum_{i=k-j+1}^{k} F_{(i)}(x)]$, where $j = \lfloor (\frac{1}{2} - \beta)k \rfloor$ and $0 < \beta < \frac{1}{2}$ is the level of interior trimming.

It is important to note that the ordering of the experts' cdfs under the cdf approach may be different at different possible realizations of x. The mean approach, on the other hand, maintains the same ordering at all possible realizations of x and produces a mixture of a subset of experts' cdfs, which is clearly a proper cdf. This raises the question whether the trimmed opinion pools under the cdf approach produce proper cdfs.

Proposition 3. *Under the cdf approach, $\hat{F}_\alpha$ and $\check{F}_\beta$ are proper cdfs.*

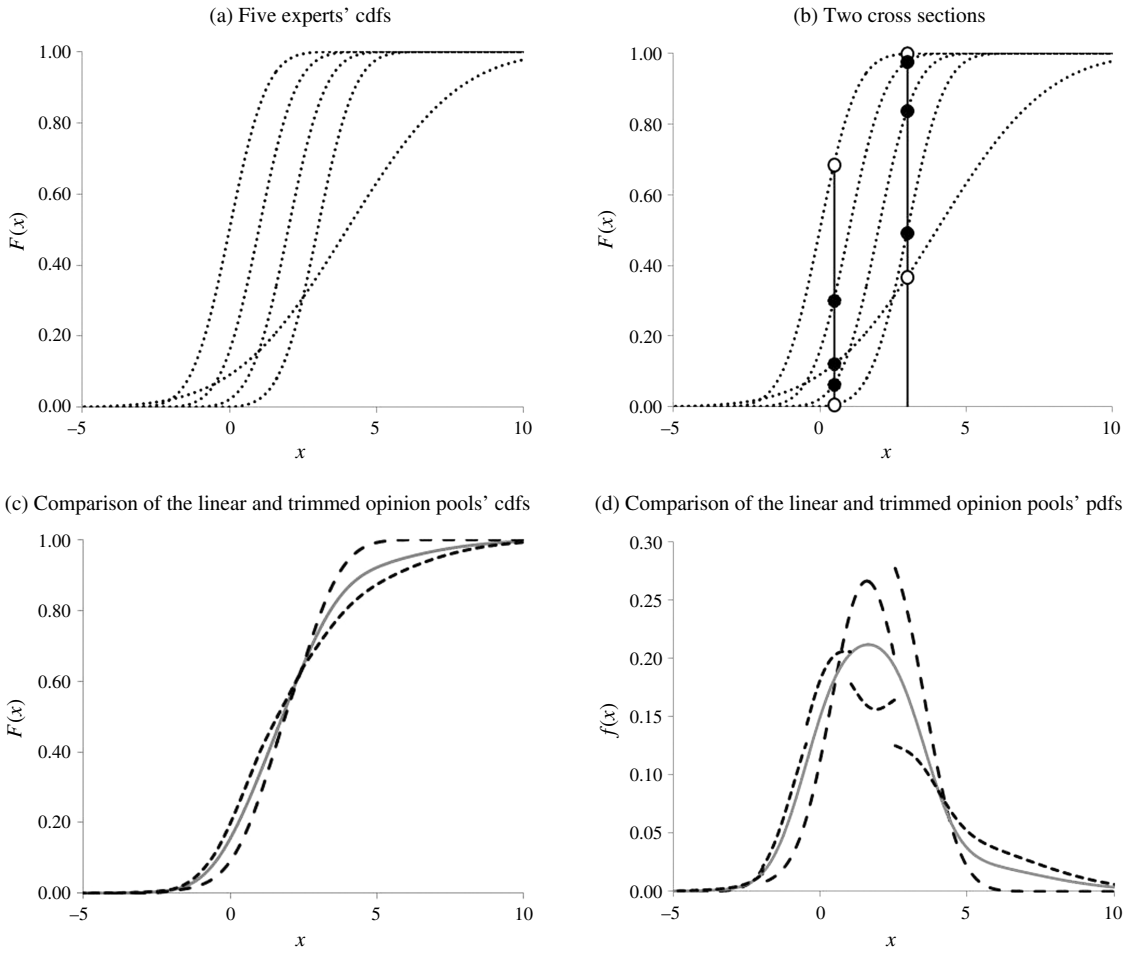
One advantage of the cdf approach is that it does not require renormalization. For instance, trimming pdf values will often result in an improper distribution.

Although low location diversity is our motivation for the cdf approach, we illustrate the construction of trimmed opinion pools using this approach with a high location-diversity example. This example serves to highlight an important property of these pools under the cdf approach: more experts can contribute to a trimmed opinion pool under the cdf approach than under the mean approach.

Example 2. Suppose five experts report normal distributions with means of 0, 1, 2, 3, and 4 and standard deviations of 1, 1, 1, 1, and 3, respectively. These five experts' cdfs are shown in Figure 2(a). In Figure 2(b), two cross sections, one at $x = 0.5$ and another at $x = 3$, are shown. At $x = 0.5$, the exterior-trimmed opinion pool includes the cdf values (marked by the filled dots) from experts 2, 5, and 3, and the interior-trimmed opinion pool includes the cdf values (marked by the open dots) from experts 1 and 4. At $x = 3$, the exterior-trimmed opinion pool includes the cdf values (marked by the filled dots) from experts 2, 3, and 4, and the interior-trimmed opinion pool includes the cdf values (marked by the open dots) from experts 5 and 1. All five experts' cdfs are included in the exterior-trimmed opinion pool at some point, whereas only 60% would be included under the mean approach. In the interior-trimmed opinion pool, 60% of the experts are included, compared to 40% under the mean approach. As in Figure 1, we can see from Figure 2(d) that the exterior-trimmed (interior-trimmed) opinion pool's pdf is narrower (wider) than the linear opinion pool's. In addition, we note that the cdf method can produce discontinuous pooled pdfs, even if the pooled cdfs are continuous. Discontinuities may occur at points where two or more experts' cdfs cross.

Whereas the mean approach is a global process that involves the removal of entire cdfs, the cdf approach is a local process. Under the cdf approach, each expert may contribute in part to the group's judgment. In this way, the cdf approach can be more egalitarian. The mean approach, on the other hand, may be easier to apply. The cdf approach will also lead to more predictable properties of a trimmed opinion pool when there is low diversity in

Figure 2 **Five Experts' CDFs (Dotted) and the Linear (Gray Solid), Exterior-Trimmed (Long-Dashed), and Interior-Trimmed (Short-Dashed) Opinion Pools Using the CDF Approach from Example 2**



the experts' means. This is the subject of our next result.

Proposition 4. *Suppose each expert reports from the same location-scale family and agrees on the location parameter, i.e.,* $\mu = \mu_1 = \cdots = \mu_k$. *Under the cdf approach, the linear opinion pool's variance of* x *is a weighted average of the trimmed pools' variances of* x: $\sigma^2_{\hat{F}} = w\sigma^2_{\hat{F}_\alpha} + (1-w)\sigma^2_{\hat{F}_\beta}$, *where* $j = \lfloor \alpha k \rfloor$, $w = (k-2j)/k$, *and* $\beta = \frac{1}{2} - \alpha$. *Moreover,* $\sigma^2_{\hat{F}_\alpha} = \hat{\sigma}^2_\alpha$ *and* $\sigma^2_{\check{F}_\beta} = \check{\sigma}^2_\beta$.

The insight from this result is similar to the one from Proposition 1. In the case of no location diversity, one trimmed pools' variance is greater than the linear opinion pool's variance and the other is less than its variance. An immediate consequence of Proposition 4 is that if $\hat{\sigma}^2_\alpha < \check{\sigma}^2_\beta$, $\sigma^2_{\hat{F}_\alpha} \le \sigma^2_{\hat{F}} \le \sigma^2_{\check{F}_\beta}$.

Roughly speaking, positive skewness in the distribution of the experts' variances will tend to produce $\hat{\sigma}^2_\alpha < \check{\sigma}^2_\beta$ and an exterior-trimmed opinion pool with a lower variance of x. Of course, positive skewness in a set of data points, such as the variances expressed by various experts, is no guarantee that the exterior-trimmed average of the points will be lower than the average of the points. Nonetheless, because variances are nonnegative numbers, one will typically encounter positive skewness in experts' variances. With the cdf approach in environments with low location diversity, the predictable effect of interior trimming will then be to increase variance. This approach to trimming will be particularly valuable when the experts are individually overconfident.

Proposition 5. *Suppose each expert reports from the same location-scale family and agrees on the scale parameter, i.e.,* $\sigma = \sigma_1 = \cdots = \sigma_k$. *Then the cdf and mean approaches result in the same exterior-trimmed (interior-trimmed) opinion pool. Moreover,* $\sigma^2_{\hat{F}_\alpha} = \sigma^2 + (1/(k-2j)) \cdot \sum_{i=j+1}^{k-j} (\mu_i - \hat{\mu}_\alpha)^2$ *and* $\sigma^2_{\check{F}_\beta} = \sigma^2 + (1/(2j))[\sum_{i=1}^{j} (\mu_i - \check{\mu}_\beta)^2 + \sum_{i=k-j+1}^{k} (\mu_i - \check{\mu}_\beta)^2]$, *where* $j = \lfloor \alpha k \rfloor$ *and* $\beta = \frac{1}{2} - \alpha$.

More generally, if the experts' cdfs can be ordered by first-order stochastic dominance, then the cdf approach is equivalent to the mean approach. Proposition 5 is a special case of this observation. When experts agree on the scale, trimming has a predictable effect on the pools' variances. An immediate consequence of Proposition 5 (together with Proposition 2) is that if $\bar{\mu} = \hat{\mu}_\alpha = \check{\mu}_\beta$, then $\sigma^2_{\hat{F}_\alpha} \le \sigma^2_{\hat{F}} \le \sigma^2_{\check{F}_\beta}$. The cdf approach, however, does not have a predictable effect on variance where experts disagree on both the location and scale of x. In the case of disagreement on location and scale, the experts' cdfs may cross many times, making it difficult to analyze the pools' variances in closed form.

5. Empirical Results

In this section, we present the results of two empirical studies that compare exterior- and interior-trimmed opinion pools to the linear opinion pool.

5.1. U.S. Survey of Professional Forecasters

We use probability forecast data from the U.S. Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters (US SPF). These data include anonymous experts' probability forecasts of inflation and output growth (as measured by GNP prior to 1992 and by GDP since 1992) over the periods 1968–2010 and 1982–2010, respectively. We do not include output growth forecasts prior to 1981:Q4 because the survey elicited forecasts of nominal GNP growth. See Diebold et al. (1999), Engelberg et al. (2009), and Clements (2010) for more details on the US SPF data.

In each quarter (Q1–Q4) during the year, expert panelists were asked for probabilities that changes in the U.S. real GDP/GNP and changes in the U.S. GDP/GNP price index between the current year and the previous year will fall in one of several predetermined bins. From the US SPF data, we have 153 quarters of inflation forecasts and realizations and 114 quarters of output growth forecasts and realizations. Realizations are based on revisions to GDP and inflation as of December 2011. On average, 35 experts responded to each inflation survey question, ranging from a low of 7 experts to a high of 106 experts. Participation on output growth survey questions was similar.

In Table 1, we present average quadratic scores for the exterior- and interior-trimmed opinion pools at what may be considered 11 different levels of trimming. At the top of the table is the linear opinion pool, which can be interpreted as 0% exterior and interior trimming. In the middle of the table, we show results for intermediate levels of trimming α, $\beta = 0.05, \ldots, 0.45$. In the last row of the table, we provide results for the median or midrange forecast. As $\alpha(\beta)$ goes to 0.5, the exterior-trimmed (interior-trimmed) opinion pool converges to the median (midrange) forecast. From these results, we see that exterior trimming outperforms the linear opinion pool of inflation forecasts. Conversely, interior trimming outperforms the linear opinion pool of output growth forecasts. At the $\alpha = 0.20$ level of exterior trimming of inflation forecasts using the cdf approach (CA), the improvement over the linear opinion pool is greatest. Interior trimming at $\beta = 0.35$ offers the highest improvement over the linear opinion pool of output growth forecasts using the mean approach (MA).

Because the SPF panelists reported probabilities over predetermined bins, a distributional assumption is required to calculate their means for the mean approach. Here, we fit continuous piecewise-linear

Table 1 US SPF's Average Quadratic Scores for the Linear, Exterior-Trimmed, and Interior-Trimmed Opinion Pools Using the Mean and CDF Approaches

Trimming level	Exterior trimming ($\hat{F}_\alpha$)				Interior trimming ($\check{F}_\beta$)			
	Inflation[a, b]		Output growth[a]		Inflation[a, b]		Output growth[a]	
α/β	MA	CA	MA	CA	MA	CA	MA	CA
Linear opinion pool	0.481		0.242		0.481		0.242	
0.05	**0.487**	**0.487**	0.228	0.231	0.476	0.477	**0.252**	**0.250**
0.10	**0.486**	**0.488**	0.210	0.216	0.474	0.474	**0.259**	**0.256**
0.15	**0.486**	**0.488**	0.198	0.208	0.468	0.468	**0.265**	**0.263**
0.20	**0.484**	**0.489**	0.185	0.199	0.459	0.459	**0.272**	**0.270**
0.25	**0.483**	**0.489**	0.178	0.192	0.448	0.449	**0.285**	**0.277**
0.30	0.481	**0.488**	0.177	0.187	0.427	0.431	**0.297**	**0.283**
0.35	0.481	**0.488**	0.172	0.182	0.393	0.399	**0.301**	**0.284**
0.40	0.479	**0.488**	0.158	0.177	0.336	0.344	**0.295**	**0.280**
0.45	0.480	**0.485**	0.151	0.176	0.214	0.226	**0.268**	**0.256**
Median/midrange	0.451	**0.483**	0.143	0.178	0.183	0.193	**0.261**	**0.242**

Note. Bold values signify scores higher than the linear opinion pool's.
[a]Excludes 1985:Q1 and 1986:Q1 due to a suspected error present in the survey.
[b]Excludes 1968: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 identified in the survey.

cdfs to the panelists' bin probabilities (Diebold et al. 1999). We note that when applying the cdf approach to bin probabilities, no distributional assumption is necessary.

In Table 2, we present the percentage of times the different trimmed opinion pools in Table 1 beat the linear opinion pool on the basis of quadratic scores. For the inflation forecasts, exterior trimming at $\alpha = 0.20$ using the cdf approach beats the linear opinion pool in 67.3% of the 153 quarters. Interior trimming at $\beta = 0.20$ beats the linear opinion pool in 58.8% of the 114 quarters of output growth.

To understand why exterior trimming performs well on inflation and interior trimming performs well on output growth, we examine the calibration of the linear and trimmed opinion pools. In Table 3, we present percentages of realizations that fall between the pools' 0.25 and 0.75 quantiles. When more (less) than 50% of the realizations fall within the pool's 50% central prediction intervals, it indicates that the pool is underconfident (overconfident). Here we see that the linear opinion pool is underconfident on inflation and overconfident on output growth. As we exteriorly trim more inflation forecasts, the percentage of realizations in the 50% central prediction interval decreases. Not surprisingly, it is close to 50% at the level of exterior trimming 0.20 that maximizes the average score. Conversely, the more we interiorly trim output growth forecasts, the more realizations fall in the 50% central prediction interval. Again, the percentage is close to 50% at the level of interior trimming 0.30 that maximizes the average score.

Next we explore how trimming address the crowd's calibration problem. In Figure 3, we present the

Table 2 US SPF's Percentages of Times That the Exterior- and Interior-Trimmed Opinion Pools Beat the Linear Opinion Pool on the Basis of Quadratic Scores

Trimming level	Exterior trimming ($\hat{F}_\alpha$) (%)				Interior trimming ($\check{F}_\beta$) (%)			
	Inflation		Output growth		Inflation		Output growth	
α/β	MA	CA	MA	CA	MA	CA	MA	CA
0.05	**63.4**	**64.1**	34.2	36.0	33.3	37.3	**57.0**	**53.5**
0.10	**68.6**	**66.7**	40.4	41.2	37.9	34.0	**61.4**	**56.1**
0.15	**66.7**	**65.4**	39.5	42.1	33.3	31.4	**58.8**	**56.1**
0.20	**64.1**	**67.3**	38.6	41.2	31.4	31.4	**58.8**	**55.3**
0.25	**64.1**	**66.7**	40.4	41.2	34.0	29.4	**57.0**	**54.4**
0.30	**64.7**	**65.4**	41.2	41.2	34.0	28.8	**58.8**	**54.4**
0.35	**64.1**	**65.4**	39.5	42.1	28.8	28.8	**57.9**	**53.5**
0.40	**61.4**	**63.4**	38.6	40.4	28.8	25.5	**57.9**	**52.6**
0.45	**64.7**	**60.8**	39.5	44.7	23.5	22.2	**53.5**	**52.6**
Median/midrange	**54.9**	**62.1**	42.1	43.0	20.3	21.6	**52.6**	**51.8**

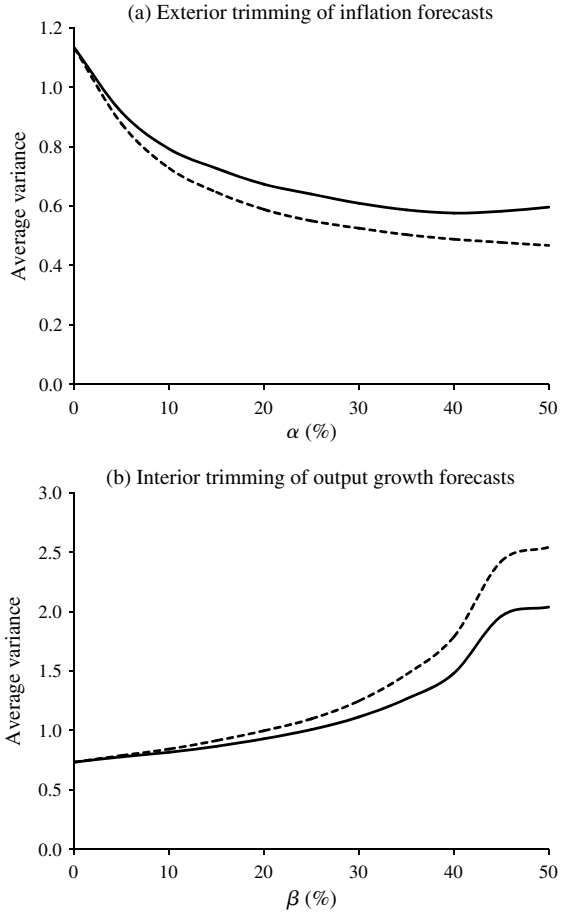
Note. Percentages above 50% are denoted in bold.

Table 3 US SPF's Percentages of Realizations That Fall Between the 0.25 and 0.75 Quantiles

Trimming level α/β	Exterior trimming ($\hat{F}_\alpha$) (%)				Interior trimming ($\check{F}_\beta$) (%)			
	Inflation		Output growth		Inflation		Output growth	
	MA	CA	MA	CA	MA	CA	MA	CA
Linear opinion pool	60.1		36.0		60.1		36.0	
0.10	56.2	57.5	35.1	36.0	63.4	62.7	36.8	36.0
0.20	51.6	54.9	30.7	34.2	68.0	65.4	39.5	37.7
0.30	48.4	52.3	30.7	34.2	78.4	73.9	48.2	46.5
0.40	48.4	51.6	31.6	33.3	86.9	83.0	63.2	61.4
Median/midrange	51.6	50.3	31.6	32.5	90.8	89.5	65.8	67.5

average variances of the trimmed opinion pools. At the 0% level, the linear opinion pool's variances are depicted. At 50% level in Figures 3(a) and 3(b), we show the average variances of the median and midrange, respectively. In Figure 3(a), the underconfident linear opinion pool of inflation forecasts has too high of an average variance. As the level of exterior trimming increases, the average variance decreases, thereby reducing the degree of underconfidence. Conversely, in Figure 3(b), the overconfident linear opinion pool of output growth forecasts has too low of an average variance. With increases in the level of interior trimming, the average variance increases, and the degree of overconfidence is reduced.

Figure 3 US SPF's Average Variances for Exterior Trimming of Inflation Forecasts and Interior Trimming of Output Growth Forecasts Under the Mean (Solid Line) and cdf (Dashed Line) Approaches



5.2. European Central Bank's Survey of Professional Forecasters

In this subsection, we consider the probability forecast data from the European Central Bank's Survey of Professional Forecasters (ECB SPF). This survey, whose design is similar to the US SPF, includes anonymous experts' probability forecasts of inflation, output growth, and unemployment over the period 1999–2010, which amounts to 48 quarters of forecasts for each of three variables. For a description of this data set and some historical perspective, see Bowles et al. (2007). Realizations are again late vintage, i.e., based on revisions as of December 2011. On average, in the ECB SPF data, 55 experts responded to each inflation survey question, ranging from a low of 46 experts to a high of 59 experts. Participation on output growth and unemployment survey questions was similar.

Table 4 shows the average quadratic scores of the linear and trimmed opinion pools of ECB SPF's forecasts. We present only the type of trimming, either exterior or interior, that leads to an improvement over the linear opinion pool. Similar to the results using the US SPF data, we find that exterior trimming benefits the pool of inflation forecasts. Interior trimming helps the pool of output growth forecasts more noticeably. Using the midrange forecast (i.e., the limit of interior trimming as β goes to 0.5), the improvement over the linear opinion pool is the highest. Interior trimming at the 0.40 level improves the performance of the pool of unemployment forecasts the most. We omit presenting the results that parallel those presented in Table 2 and Figure 3 because these results for the ECB SPF are similar to the corresponding results for the US SPF.

In Table 5, we report the ECB SPF's percentages of realizations that fall between the opinion pools' 0.25

Table 4 ECB SPF's Average Quadratic Scores for the Linear, Exterior-Trimmed, and Interior-Trimmed Opinion Pools Using the Mean and CDF Approaches

	Exterior trimming ($\hat{F}_\alpha$)		Interior trimming ($\check{F}_\beta$)			
	Inflation		Output growth		Unemployment	
Trimming level α/β	MA	CA	MA	CA	MA	CA
Linear opinion pool	0.519		0.079		0.155	
0.10	**0.525**	**0.522**	**0.088**	**0.088**	**0.175**	**0.168**
0.20	**0.522**	**0.523**	**0.108**	**0.102**	**0.188**	**0.180**
0.30	0.519	**0.520**	**0.129**	**0.121**	**0.208**	**0.196**
0.40	0.508	0.518	**0.157**	**0.154**	**0.237**	**0.222**
Median/midrange	0.516	0.510	**0.174**	**0.169**	**0.201**	**0.221**

Note. Bold values signify scores higher than the linear opinion pool's.

Table 5 ECB SPF's Percentages of Realizations That Fall Between the 0.25 and 0.75 Quantiles

	Exterior trimming ($\hat{F}_\alpha$) (%)		Interior trimming ($\check{F}_\beta$) (%)			
	Inflation		Output growth		Unemployment	
Trimming level α/β	MA	CA	MA	CA	MA	CA
Linear opinion pool	79.2		18.8		29.2	
0.10	79.2	79.2	18.8	18.8	31.3	31.3
0.20	77.1	77.1	25.0	25.0	31.3	31.3
0.30	77.1	77.1	27.1	27.1	35.4	35.4
0.40	75.0	75.0	35.4	35.4	54.2	54.2
Median/midrange	75.0	75.0	47.9	47.9	64.6	64.6

and 0.75 quantiles. The ECB SPF crowd's calibration problem on inflation and output growth is in the same direction as, but more extreme than, that of the US SPF crowd. The linear opinion pool of inflation forecasts is underconfident and therefore benefits from exterior trimming. The linear opinion pool of output growth forecasts is overconfident, which is mitigated by interior trimming. Like output growth, the linear opinion pool of ECB SPF unemployment forecasts is overconfident and becomes less so with interior trimming.

6. Prescriptions for Practice

In our empirical results above, we do not see a systematic difference in performance between the mean and cdf approaches to trimming. There is, however, one important systematic difference worth noting. Under the cdf approach, the exterior- and interior-trimmed opinion pools include segments of many more panelists' forecasts. In Table 6, we present the percentages of US SPF forecasters included in a trimmed opinion pool. From a prescriptive point of view, greater inclusion may lead to broader participation and fewer incentives to distort individual forecasts.

In the previous section, we identify ex post the types and levels of trimming that offer the most improvements over the linear opinion pool. Naturally though, we will want to know ex ante what type and level of trimming to use. It may be unrealistic, however, to make such a prescription before knowing something about how well (or poorly) calibrated the crowd is.

6.1. Split Sample Method

Suppose we train a trimmed opinion pool on the first half of the data and then test the trained pool on the second half of the data. In Table 7, we report results for this split sample method under the cdf approach. On inflation, we can see that training on the first half of the data suggests we exteriorly trim at the 0.10 level (with corresponding scores shown in bold), which in test leads to an improvement over the linear opinion pool on the second half of the data (0.540 versus 0.527). On output growth, interior trimming at the

Table 6 Percentages of US SPF Forecasters Included in the Exterior- and Interior-Trimmed Opinion Pools

	Exterior trimming ($\hat{F}_\alpha$) (%)		Interior trimming ($\check{F}_\beta$) (%)	
Trimming level α/β	MA	CA	MA	CA
Linear opinion pool	100		100	
0.10	83.0	99.8	83.0	99.8
0.20	62.7	98.4	62.7	98.4
0.30	43.0	92.5	43.0	92.5
0.40	22.9	75.2	22.9	75.2
Median/midrange	5.1	29.6	5.1	29.6

Table 7 US SPF's Average Quadratic Scores Under the CDF Approach Using a Split Sample Approach for Training the Trimmed Opinion Pool

		Inflation		Output growth	
		1969:Q1–1991:Q3	1991:Q4–2010:Q4	1982:Q1–1996:Q3	1996:Q4–2010:Q4
Linear opinion pool		0.434	0.527	0.363	0.121
	α				
Exterior trimming ($\hat{F}_\alpha$)	0.10	**0.435**	**0.540**	0.341	0.092
	0.20	0.434	0.543	0.328	0.070
	0.30	0.431	0.544	0.321	0.053
	0.40	0.431	0.544	0.313	0.041
	Median	0.430	0.536	0.313	0.042
	β				
Interior trimming ($\check{F}_\beta$)	0.10	0.430	0.517	0.372	0.140
	0.20	0.415	0.503	0.381	0.159
	0.30	0.389	0.472	**0.386**	**0.180**
	0.40	0.311	0.376	0.366	0.195
	Midrange	0.186	0.200	0.319	0.166

0.30 level does best in training and leads to a greater improvement over the linear opinion pool in testing (0.180 versus 0.121). Other methods for training a trimmed opinion pool may offer greater improvement. We find similar improvements to those in Table 7 when we train and test trimmed opinion pools using the moment approach and on the ECB SPF data.

In theory, however, if there were a regime shift (i.e., a shock to the system) somewhere in the second half of the data, there may be large penalties to using the split sample method. For instance, an underconfident linear opinion pool in the training set may become overconfident in the testing set or vice versa.

6.2. Sliding Window Method

Inspired by moving averages from time series forecasting, a more adaptive approach applies trimming to the next period based on a moving training set. For example, using a sliding lookback window of three periods, trimming for 2010:Q2 can be set to the level of trimming that maximized the average quadratic score over the previous three periods: 2010:Q1, 2009:Q4, and 2009:Q3. In this sense, the past three periods' forecasts and realizations serve as the training set for 2010:Q2. The sliding window approach is relatively simple to apply and does not require long training periods. In addition, we expect it to rapidly adapt to regime shifts in the forecasting environment, changes in the crowd's diversity and composition, and variations in individual calibration.

In Table 8, we report the average quadratic scores associated with using sliding windows of one, three, five, and seven periods looking backward. Here we find that shorter windows work better than longer windows do and work better than the linear opinion pool does. For inflation, we find that using a single period window works best, whereas for output growth we find that a sliding window of three periods works best. For comparison purposes, the average quadratic scores for the linear opinion pool in Table 8 are calculated by excluding the first 1, 3, 5, and 7 periods' scores, respectively.

Table 8 US SPF's Average Quadratic Scores Using a Sliding Lookback Window for Training the Trimmed Opinion Pool

	Inflation			Output growth		
		Trimmed opinion pool			Trimmed opinion pool	
Lookback window	Linear opinion pool	MA	CA	Linear opinion pool	MA	CA
One period	0.483	0.530	0.529	0.239	0.352	0.341
Three periods	0.484	0.529	0.519	0.229	0.284	0.273
Five periods	0.489	0.528	0.505	0.226	0.268	0.225
Seven periods	0.493	0.508	0.507	0.235	0.261	0.235

To compare the split sample and sliding window methods, we apply the sliding window method to the test set used in the split sample method above (i.e., to the second half of the US SPF data). In Table 9, we report the results of this comparison. In terms of average quadratic scores, we find that in all but four cases, the sliding window method performs better than the split sample method.

7. Conclusions

We introduce trimmed opinion pools as an alternative to the popular linear opinion pool. To form these pools, we propose two approaches for ordering experts' probability forecasts: the mean and cdf approaches. Our results indicate that trimming away forecasts from a linear opinion pool can address the crowd's calibration problem. When the linear opinion pool is underconfident, we find that exterior trimming can lead to a trimmed opinion pool with lower variance and improved calibration. Conversely, interior trimming increases variance and improves calibration when the linear opinion pool is overconfident. With improved calibration, we expect higher average scores.

Using probability forecast data from the U.S. and European Surveys of Professional Forecasters, we compare the average quadratic scores of the linear opinion pool and various trimmed opinion pools. We find that exterior trimming of inflation forecasts from both surveys performs best and better than the linear opinion pool. Interior trimming at higher levels of output growth forecasts performs best on the two respective surveys. Exterior trimming of inflation forecasts makes sense because in both surveys

Table 9 **A Comparison of the Average Quadratic Scores from Trimmed Opinion Pools Formed Using the Split Sample and the Sliding Window Methods on the Second Half of the US SPF Data**

	Inflation				Output growth			
	MA		CA		MA		CA	
Lookback window	Split sample method	Sliding window method	Split sample method	Sliding window method	Split sample method	Sliding window method	Split sample method	Sliding window method
One period	0.542	**0.603**	0.540	**0.588**	0.193	**0.243**	0.180	**0.215**
Three periods	0.542	**0.574**	0.540	**0.560**	0.193	**0.229**	**0.180**	0.173
Five periods	0.542	**0.546**	**0.540**	0.535	0.193	**0.218**	0.180	**0.185**
Seven periods	**0.542**	0.512	**0.540**	0.524	0.193	**0.201**	0.180	**0.192**

the linear opinion pool is underconfident. On output growth, interior trimming at higher levels is called for because the linear opinion pool is further away from perfect calibration, but on the overconfident side.

These empirical results suggest a straightforward prescription for practice. Choose the type and level of trimming on the basis of past performance in a training set. Then use the chosen trimmed opinion pool for aggregating future forecasts. The training set may come directly from the forecasting environment itself. As we demonstrate with the US SPF data, training a trimmed opinion pool in this way can lead to improvements in future performance above and beyond the linear opinion pool. Alternatively, a trimmed opinion pool could be trained using responses to a set of almanac questions. Like linear opinion pools, trimmed opinion pools are easy to formulate. Thus, they may be a simple and effective alternative to the linear opinion pool.

Of course, alternative aggregation schemes can be used to address the crowd's calibration problem. One natural alternative to our cdf approach is to trim cdf values horizontally as opposed to vertically. This amounts to trimming quantiles rather than probabilities. Given the benefits quantile averaging offers over probability averaging (Lichtendahl et al. 2013), there may be additional gains to be had this way. One alternative to the trimmed opinion pool is the Winsorized opinion pool. In the robust statistics literature, the effects of outliers can similarly be mitigated with a Winsorized mean instead of a trimmed mean. Winsorizing in this context is considered by some to be more attractive than trimming because it retains the same number of observations, replacing extreme values with more moderate ones. Other schemes involve the direct transformation of the linear opinion pool, such as a beta transformation (Ranjan and Gneiting 2010) or a mean-preserving scale transformation. With any of these schemes, like our trimmed opinion pools, the variance of an underconfident pool can be reduced and the variance of an overconfident pool can be increased.

Acknowledgments

The authors thank the review team for their helpful suggestions.

Appendix: Proofs of Main Results

Proof of Proposition 1. Expanding and rearranging terms of $\sigma_{\bar{F}}^2$, we have the following:

(i) $$\frac{1}{k}\sum_{i=1}^{k}\sigma_i^2 = \frac{k-2j}{k}\frac{1}{k-2j}\sum_{i=j+1}^{k-j}\sigma_i^2 + \frac{2j}{k}\frac{1}{2j}\left[\sum_{i=1}^{j}\sigma_i^2 + \sum_{i=k-j+1}^{k}\sigma_i^2\right],$$

(ii) $$\frac{1}{k}\sum_{i=1}^{k}(\mu_i-\bar{\mu})^2 = \frac{k-2j}{k}\frac{1}{k-2j}\sum_{i=j+1}^{k-j}(\mu_i-\bar{\mu})^2 + \frac{2j}{k}\frac{1}{2j}\left[\sum_{i=1}^{j}(\mu_i-\bar{\mu})^2 + \sum_{i=k-j+1}^{k}(\mu_i-\bar{\mu})^2\right],$$

(iii) $$\bar{\mu} = \frac{k-2j}{k}\frac{1}{k-2j}\sum_{i=j+1}^{k-j}\mu_i + \frac{2j}{k}\frac{1}{2j}\left[\sum_{i=1}^{j}\mu_i + \sum_{i=k-j+1}^{k}\mu_i\right] = w\hat{\mu}_\alpha + (1-w)\check{\mu}_\beta,$$

(iv) $$\begin{aligned}\frac{1}{k-2j}\sum_{i=j+1}^{k-j}(\mu_i-\bar{\mu})^2 &= \frac{1}{k-2j}\sum_{i=j+1}^{k-j}(\mu_i^2-2\mu_i\bar{\mu}+\bar{\mu}^2)\\ &= \frac{1}{k-2j}\sum_{i=j+1}^{k-j}(\mu_i-\hat{\mu}_\alpha)^2+\hat{\mu}_\alpha^2-2\hat{\mu}_\alpha\bar{\mu}+\bar{\mu}^2\\ &= \frac{1}{k-2j}\sum_{i=j+1}^{k-j}(\mu_i-\hat{\mu}_\alpha)^2+(\hat{\mu}_\alpha-\bar{\mu})^2,\end{aligned}$$ and

(v) $$\begin{aligned}&\frac{1}{2j}\left[\sum_{i=1}^{j}(\mu_i-\bar{\mu})^2+\sum_{i=k-j+1}^{k}(\mu_i-\bar{\mu})^2\right]\\ &\quad=\frac{1}{2j}\left[\sum_{i=1}^{j}(\mu_i^2-2\mu_i\bar{\mu}+\bar{\mu}^2)+\sum_{i=k-j+1}^{k}(\mu_i^2-2\mu_i\bar{\mu}+\bar{\mu}^2)\right]\\ &\quad=\frac{1}{2j}\left[\sum_{i=1}^{j}(\mu_i-\check{\mu}_\beta)^2+\sum_{i=k-j+1}^{k}(\mu_i-\check{\mu}_\beta)^2\right]+(\check{\mu}_\beta-\bar{\mu})^2.\end{aligned}$$

Putting these terms together, we have

$$\sigma^2_{\bar{F}} = \frac{k-2j}{k}\left[\frac{1}{k-2j}\sum_{i=j+1}^{k-j}\sigma_i^2 + \frac{1}{k-2j}\sum_{i=j+1}^{k-j}(\mu_i-\hat{\mu}_\alpha)^2 + (\hat{\mu}_\alpha-\bar{\mu})^2\right] + \frac{2j}{k}\left[\frac{1}{2j}\left[\sum_{i=1}^{j}\sigma_i^2 + \sum_{i=k-j+1}^{k}\sigma_i^2\right] + \frac{1}{2j}\left[\sum_{i=1}^{j}(\mu_i-\breve{\mu}_\beta)^2 + \sum_{i=k-j+1}^{k}(\mu_i-\breve{\mu}_\beta)^2\right] + (\breve{\mu}_\beta-\bar{\mu})^2\right],$$

which reduces to the result after substituting according to $\sigma^2_{\hat{F}_\alpha}$ and $\sigma^2_{\breve{F}_\beta}$. □

Proof of Proposition 2. Let

$$d = \frac{1}{k}\sum_{i=1}^{k}(\mu_i-\bar{\mu})^2, \quad \hat{d}_\alpha = \frac{1}{k-2j}\sum_{i=j+1}^{k-j}(\mu_i-\hat{\mu}_\alpha)^2 \quad \text{and}$$

$$\breve{d}_\beta = \frac{1}{2j}\left[\sum_{i=1}^{j}(\mu_i-\breve{\mu}_\beta)^2 + \sum_{i=k-j+1}^{k}(\mu_i-\breve{\mu}_\beta)^2\right].$$

From the assumption of symmetry and (ii) in the proof of Proposition 1, we have that $\bar{\mu} = \hat{\mu}_\alpha = \breve{\mu}_\beta$ and $d = w\hat{d}_\alpha + (1-w)\breve{d}_\beta$, where $w = (k-2j)/k$. Next we show that $\hat{d}_\alpha \le \breve{d}_\beta$. There are two cases:

(i) When $k-2j \ge 2$ (i.e., at least two cdfs are averaged in the exterior-trimmed opinion pool), the following inequalities hold:

$$\hat{d}_\alpha \le \frac{k-2j}{2}\frac{1}{k-2j}[(\mu_{j+1}-\bar{\mu})^2 + (\mu_{k-j}-\bar{\mu})^2] \le \frac{1}{2}\left[\frac{1}{j}\sum_{i=1}^{j}(\mu_i-\bar{\mu})^2 + \frac{1}{j}\sum_{i=k-j+1}^{k}(\mu_i-\bar{\mu})^2\right] \le \breve{d}_\beta.$$

(ii) When $k-2j=1$, $\hat{d}_\alpha = 0 \le \breve{d}_\beta$.

Because $d = w\hat{d}_\alpha + (1-w)\breve{d}_\beta$, where $w = (k-2j)/k$ and $\hat{d}_\alpha \le \breve{d}_\beta$, we have that $\hat{d}_\alpha \le d \le \breve{d}_\beta$. Finally, because $\hat{\sigma}^2_\alpha = \breve{\sigma}^2_\beta$ (by assumption), $\hat{d}_\alpha \le d \le \breve{d}_\beta$ (as just established), $\sigma^2_{\hat{F}_\alpha} = \hat{\sigma}^2_\alpha + \hat{d}_\alpha$, $\sigma^2_{\breve{F}_\beta} = \breve{\sigma}^2_\beta + \breve{d}_\beta$, and $\sigma^2_{\bar{F}} = w\sigma^2_{\hat{F}_\alpha} + (1-w)\sigma^2_{\breve{F}_\beta}$ from Proposition 1 when $\bar{\mu} = \hat{\mu}_\alpha = \breve{\mu}_\beta$, we have that $\sigma^2_{\hat{F}_\alpha} \le \sigma^2_{\bar{F}} \le \sigma^2_{\breve{F}_\beta}$. □

Proof of Proposition 3. We show that $\hat{F}_\alpha$ is a proper cdf; the arguments for $\breve{F}_\beta$ being a proper cdf follow similarly. To establish that $\hat{F}_\alpha$ is a proper cdf, we need to show that it is a nondecreasing, right-continuous function on the real line where $\lim_{x\to-\infty}\hat{F}_\alpha(x) = 0$ and $\lim_{x\to\infty}\hat{F}_\alpha(x) = 1$. First, $\lim_{x\to-\infty}\hat{F}_\alpha(x) = 0$ and $\lim_{x\to\infty}\hat{F}_\alpha(x) = 1$ because each of the experts' cdfs has these limits. Second, we show $\hat{F}_\alpha$ is nondecreasing. If we can show that for each i, $F_{(i)}(x') \le F_{(i)}(x'')$ for any $x' \le x''$ then $\hat{F}_\alpha$ is nondecreasing as it is would be a simple average of nondecreasing functions. This follows from the following: At x', $F_{(i)}(x') \le F_{(i+1)}(x') \le \cdots \le F_{(k)}(x')$. This means that there exists at least $k-i+1$ experts whose cdf values at x' are at least $F_{(i)}(x')$. At $x' \le x''$, we may not know whether the ordering remains the same; however, we know that at least $k-i+1$ experts must have cdf values greater than $F_{(i)}(x')$ since each expert's cdf is nondecreasing. Therefore, this imposes a lower bound on $F_{(i)}(x') \le F_{(i)}(x'')$. Third, we show right continuity. Let x_0 be a realization of x. Choose a realization x' that is above x_0 for which the ordering of the experts' cdfs does not change on the interval $[x_0, x']$. Because the experts' cdfs cross at most a finite number of times, we can always find such an x'. Then $\lim_{x'\to x_0^+}\hat{F}_\alpha(x') = \lim_{x'\to x_0^+}(1/(k-2j))\sum_{i=j+1}^{k-j}F_{(i)}(x') = (1/(k-2j))\sum_{i=j+1}^{k-j}\lim_{x'\to x_0^+}F_{(i)}(x') = (1/(k-2j))\sum_{i=j+1}^{k-j}F_{(i)}(x_0) = \hat{F}_\alpha(x_0)$ because each $F_{(i)}$ is right continuous. □

Proof of Proposition 4. We let $\sigma_1^2 \le \cdots \le \sigma_k^2$ without loss of generality. Also let F_0 be the cdf of the standardized member of the location-scale family. As the standardized member, it has a mean of zero and a variance of one. Expert i reports the cdf $F_0((x-\mu)/\sigma_i)$ where the scale σ_i is also expert i's standard deviation of x. At the realization $x = \mu$, each expert has the same cdf value $F_0(0)$. For realizations of x below μ, the ordering of experts' cdfs is the same as the ordering of the experts' variances: $F_0((x-\mu)/\sigma_1) \le \cdots \le F_0((x-\mu)/\sigma_k)$. And for realizations of x above μ, the ordering of the experts' cdfs is reversed: $F_0((x-\mu)/\sigma_k) \le \cdots \le F_0((x-\mu)/\sigma_1)$. Thus, the experts' cdfs only cross once at the point $x = \mu$. Under the cdf approach, the exterior-trimmed opinion pool is the simple average of the cdfs of the experts with the variances $\sigma^2_{j+1}, \ldots, \sigma^2_{k-j}$. Because $\mu_1 = \cdots = \mu_k$, we have $\sigma^2_{\hat{F}_\alpha} = \hat{\sigma}^2_\alpha = (1/(k-2j))\sum_{i=j+1}^{k-j}\sigma_i^2$. The interior-trimmed opinion pool is the simple average of the cdfs of the experts with the variances $\sigma_1^2, \ldots, \sigma_j^2, \sigma^2_{k-j+1}, \ldots, \sigma_k^2$, and $\sigma^2_{\breve{F}_\beta} = \breve{\sigma}^2_\beta = (1/(2j))[\sum_{i=1}^{j}\sigma_i^2 + \sum_{i=k-j+1}^{k}\sigma_i^2]$. The rest of the result follows directly from the decomposition of $\sigma^2_{\bar{F}}$ shown in the proof of Proposition 1 under the assumption that $\mu_1 = \cdots = \mu_k$. □

Proof of Proposition 5. Expert i reports the cdf $F_0((x-\mu_i)/\sigma)$. At each realization x, the ordering of experts' cdfs is the same as the ordering of the experts' means: $F_0((x-\mu_k)/\sigma) \le \cdots \le F_0((x-\mu_1)/\sigma)$. Under the cdf approach, the exterior-trimmed opinion pool is the simple average of the cdfs of the experts with the means $\mu_{j+1}, \ldots, \mu_{k-j}$. The interior-trimmed opinion pool is the simple average of the cdfs of the experts with the means $\mu_1, \ldots, \mu_j, \mu_{k-j+1}, \ldots, \mu_k$. Thus, the cdf and mean approaches yield the same trimmed opinion pools. In addition, because $\sigma = \sigma_1 = \cdots = \sigma_k$, we have $\sigma^2_{\hat{F}_\alpha} = \sigma^2 + (1/(k-2j))\sum_{i=j+1}^{k-j}(\mu_i-\hat{\mu}_\alpha)^2$ and $\sigma^2_{\breve{F}_\beta} = \sigma^2 + (1/(2j))[\sum_{i=1}^{j}(\mu_i-\breve{\mu}_\beta)^2 + \sum_{i=k-j+1}^{k}(\mu_i-\breve{\mu}_\beta)^2]$. □

References

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

Boero G, Smith J, Wallis KF (2008) Uncertainty and disagreement in economic prediction: The Bank of England survey of external forecasters. *Econom. J.* 118:1107–1127.

Bowles C, Friz R, Genre V, Kenny G, Meyler A, Rautanen T (2007) The ECB survey of professional forecasters (SPF): A review after eight years' experience. *Eur. Central Bank Occasional Paper Series* 59:1–68.

Clemen RT (1989) Combining forecasts: A review and annotated. *Internat. J. Forecasting* 5:559–583.

Clemen RT (2008) Comment on Cooke's classical method. *Reliability Engrg. System Safety* 93:760–765.

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.

Cooke RM (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science* (Oxford University Press, Oxford, UK).

Dawid AP, DeGroot MH, Mortera J (1995) Coherent combination of experts' opinions. *Test* 4:263–313.

Diebold FX, Tay AS, Wallis KF (1999) Evaluating density forecasts of inflation: The survey of professional 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.

Giordani P, Söderlind P (2003) Inflation forecast uncertainty. *Eur. Econom. Rev.* 47:1037–1059.

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

Jose VRR, Winkler RL (2008) Simple robust average of forecasts: Some empirical results. *Internat. J. Forecasting* 24:163–169.

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.

Lichtendahl KC Jr, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles? *Management Sci.* 59:1594–1611.

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).

Prescott P, Hogg RV (1977) Trimmed and outer means and their variances. *Amer. Statistician* 31:156–157.

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

Smith J, Wallis KF (2009) A simple explanation of the forecast combination puzzle. *Oxford Bull. Econom. Statist.* 71:331–355.

Soll JB, Mannes AE, Larrick RP (2013) The "wisdom of crowds" effect. Pashler H, ed. *Encyclopedia of the Mind*, Vol. 2 (Sage Publications, Los Angeles), 776–778.

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

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

Winkler RL, Jose VRR (2011) Scoring rules. Cohcran J, ed. *Wiley Encyclopedia of Operations Research and Management Science*, Vol. 7 (John Wiley & Sons, New York), 4733–4744.