Federal Reserve Bank of Philadelphia

Research Brief
Research Department

December 2018
https://doi.org/10.21799/frbp.rb.2018.dec

Battle of the Forecasts: Mean vs. Median as the Survey of Professional Forecasters' Consensus

Fatima Mboup
Federal Reserve Bank of Philadelphia Research Department

Ardy L. Wurtzel

Battle of the Forecasts: Mean vs. Median as the Survey of Professional Forecasters' Consensus

by **Fatima Mboup** and Ardy L. Wurtzel

In the nearly 28 years that the Federal Reserve Bank of Philadelphia has been conducting the widely followed Survey of Professional Forecasters (SPF), our analysis of the results of each new quarterly survey has emphasized the median forecast as a measure of the consensus projection.[1] The median—which we compute as the middle projection of approximately 40 individual forecasts for a broad array of major macroeconomic indicators—has proved to be a popular benchmark against which alternative forecasts can be judged. However, the median is not the only reasonable choice for a consensus projection. We could also use as the consensus the simple arithmetic average, or mean. Yet, we have avoided that approach because the mean projection is particularly sensitive to any extreme projections from our panelists. Because extreme individual responses could produce inaccurate consensus forecasts, using the mean could potentially harm the overall reliability and usefulness of the survey's projection. That said, we have never formally tested our assumption in a statistically rigorous and comprehensive manner.

In this *Research Brief*, we study whether the accuracy of the median forecast in fact exceeds that of the mean forecast. Because we want the results of our study to be as robust as possible, we examine the forecasts for six important survey variables over five forecast horizons, using four alternative measures of the realizations from which we compute the forecast errors, and four alternative sample periods. We apply the well-known Diebold–Mariano (1995) statistical test for relative forecast accuracy between the mean and median consensus projections.

Overall, our analysis shows some statistically significant differences between the accuracy of the mean and median forecasts, with the mean forecasts more often exhibiting marginally higher levels of accuracy. However, these differences in accuracy, when they exist, are often small. Moreover, because our results are sensitive to the variable, forecast horizon, sample period, and measure for the realizations, we conclude that the SPF median forecast remains a useful measure of the consensus projection when we factor in

Fatima Mboup is a research analyst and Ardy L. Wurtzel is a former real-time data specialist in the Research Department of the Federal Reserve Bank of Philadelphia.

1 For more information on the Philadelphia Fed's SPF, see www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters.

its benefits in guarding against typographical errors in the panelists' forecast submissions.[2]

Real-Time Methodological Considerations

In studying the relative accuracy of the SPF mean and median projections, we pay particular attention to the real-time nature of our statistical experiments. The survey itself is conducted in real time, not after the fact, and the panelists' projections are, accordingly, based on the macroeconomic information they had when they formulated their projections. Moreover, the Philadelphia Fed does not change the consensus projections after the survey date.[3] The real-time nature of the survey's forecasts suggests that we must exercise some caution in our choice of the historical realizations against which we will assess forecast accuracy.

It is well known that U.S. government statistical agencies frequently revise the historical realizations that we use to compute our forecast errors. These data revisions suggest that our findings could be sensitive to the measure we choose to represent the realizations. We account for revisions to the historical realizations by considering alternative measures of the realizations, depending on the degree of revision to which the realization is subject. Our four measures for the realizations are, alternatively, the first, second, and third release values published by the statistical agencies and the values as we know them today.[4]

Because the SPF began in 1968,[5] we focus our analysis on the period from 1968 to 2016, the latter date corresponding with the last observation available when we computed our results. We think our focus on the full sample period is important in our study because much previous academic work has focused on this period using the median projection, and we want to document differences in forecast accuracy between mean and median projections over the same period.[6] At the same time, good reasons exist for considering additional periods for the analysis. We also consider the period over which the National Bureau of Economic Research (NBER) and the American Statistical Association (ASA) conducted the survey (1968 to 1990) and the period over

2 It is worth noting that, although the Philadelphia Fed focuses on the median projections in our analysis of each survey's results, we provide the survey's entire history of mean projections as well as median projections at www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/data-files.

3 When we discover a typographical error in our data set, sometimes well after the survey date, we adjust the data appropriately. Such occurrences are rare and almost always have minor effects on the consensus projections.

4 Our realizations come from the Philadelphia Fed's real-time data set for macroeconomists at www.philadelphiafed.org/research-and-data/real-time-center/real-time-data.

5 The Philadelphia Fed has conducted the SPF since the second quarter of 1990. From its inception in 1968 through 1990, it was conducted by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER). The results of the survey are monitored by policymakers, research economists, and business analysts around the world for the survey's near-term outlook for the U.S. economy, as seen by a panel of professional forecasters who closely follow economic developments and generate their projections using both formal mathematical models and subjective judgments.

6 A bibliography of academic papers using the survey's data is available at www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/academic-bibliography.

which the Philadelphia Fed has been in charge of the survey (1991 to 2016). We distinguish between these periods because we want to control for rounding errors in the underlying forecast data under the NBER's tenure. We also want to allow for the possibility that the Philadelphia Fed's tenure marks one of increased attention to typographical errors and other mistakes in the survey's responses. Finally, we have a particular interest in the period from 2005 to 2016 because it covers the most recent forecasts as well as the Great Recession and the subsequent recovery, a period over which one could reasonably expect some forecasters to submit extreme forecasts that could drive an interesting wedge between the consensus mean and median projections.

The SPF itself has evolved over time. Of the 23 variables currently in the survey, only a handful appeared in the survey in 1968. Because our interest includes the entire history of the SPF, we focus on those variables with the longest history in the survey: nominal GDP, the GDP price index, industrial production, real GDP, the unemployment rate, and housing starts. The forecast horizons cover the nowcast (current) quarter and the following four quarters.[7]

The Mean-Square Error as a Measure of Forecast Accuracy

Our tests for the relative accuracy of the SPF's mean and median consensus forecasts follow the methodology of Stark (2010) in using the conventional mean-square-error (MSE) statistic, which we define as

$$mse_j\left(\tau,r\right)=\frac{1}{T}\sum_t\left(\varepsilon^{(r)}_{j,\,t+\tau|t}\right)^2,\ \ \tau=0,...,4;\ \ r=1,...,4;\ j=1,2\quad,$$

where $\varepsilon^{(r)}_{j,\,t+\tau|t}$ is the τ-quarter-ahead forecast error, defined on the r^{th} measure of the realization; j is an index to distinguish between mean and median forecasts; and T measures the number of forecast errors in the sample period. The MSE is a popular summary statistic because it gives equal consideration to two undesirable properties of forecast errors: bias and variance. Indeed, a large MSE reflects either a large bias from zero in the forecast error or a large forecast error variance, or both, and signals poor forecast accuracy.

The Diebold–Mariano (1995) statistic forms the centerpiece for testing our null hypothesis that the MSE for the SPF mean forecast (mse_1) equals the MSE for the SPF median forecast (mse_2). Formally, we follow Diebold and Mariano (1995) in designing the null and alternative hypotheses as

$$H_0: mse_1\left(\tau,r\right)=mse_2\left(\tau,r\right)$$

$$H_1: mse_1\left(\tau,r\right)\neq mse_2\left(\tau,r\right),$$

7 The forecasts for real GDP, nominal GDP, the GDP price index, and industrial production are expressed as quarter-over-quarter growth rates, compounded quarterly and expressed in annualized percentage points. (The level of the quarterly industrial production index is the quarterly average of the underlying monthly levels.) The forecasts for the quarterly average of monthly housing starts and the quarterly average of monthly unemployment rates are in millions of units and percentage points, respectively.

where, as noted above, we conduct a separate test for each variable, forecast horizon ($\tau = 0,1,\ldots,4$), realization ($r = 0,1,\ldots,4$), and sample period.

Statistical Findings

We begin with a quick inspection of the qualitative differences between the mean and median forecasts at the two-quarters-ahead horizon. The figure presents scatter plots of the median forecasts (*y* axis) against the mean forecasts (*x* axis) for each variable. The two-quarters-ahead horizon is representative of the additional horizons that we examined. A 45-degree line defines the locus along which the two forecasts are identical. Points above the locus indicate that the median forecast exceeds the mean forecast, while points below the locus indicate that the mean forecast exceeds the corresponding median forecast. We also distinguish between the consensus forecasts derived during the NBER's tenure in conducting the survey (1968 to 1990; light blue dots) and those derived during the Philadelphia Fed's tenure (1991 to 2016; dark blue dots).

The figure suggests three notable features in the SPF forecast data. First, most points cluster just above and below the 45-degree line, suggesting that the mean and median forecasts are close to each other but not identical. Second, for most variables, the differences between the mean and median forecasts, measured by the distance between the points and the 45-degree locus, seem greatest during the NBER's tenure in conducting the survey. Third, the differences between the mean and median forecasts are often greatest at the extreme values of the forecasts. It is also notable that at longer forecast horizons (not shown) we find larger differences between the mean and median projections, and the differences seem to grow with the horizon.[8]

Turning now to the formal statistical evidence, our Diebold–Mariano tests detected some statistically significant differences between the accuracy of the mean and median forecasts. However, the differences were almost always small. Tables 1 through 6 summarize our statistical findings separately for each variable, showing the number of times we rejected our null hypothesis of MSE equality between the mean and median forecasts. We computed counts separately for the cases in which the median forecast was more accurate than the mean and for the cases in which the mean forecast was more accurate.[9] We conducted a total of 420 Diebold–Mariano tests in our analysis. Our summary tables show that when a statistically significant difference exists between the mean and median forecasts, the mean forecast was superior 81 times, while the median was superior 54 times.

Taking a closer look at our results for the real GDP projections in Table 4, we note that, when a statistically significant difference exists between the mean and median forecasts, mean forecasts are superior in 19 cases versus in five

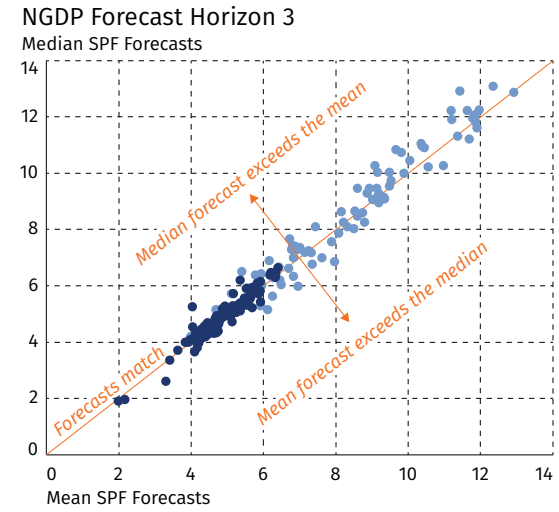
8 These results are available on request.

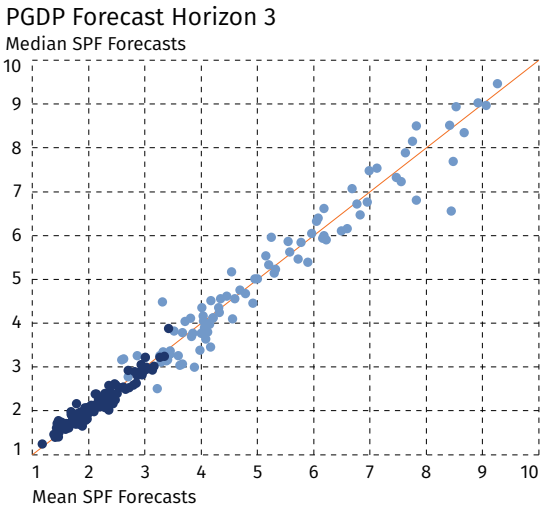
9 Our Diebold–Mariano statistics reflect the usual heteroscedasticity and autocorrelation consistent (HAC) correction using a rectangular window and truncation lag parameter of four quarters. We employ a 10 percent significance level.

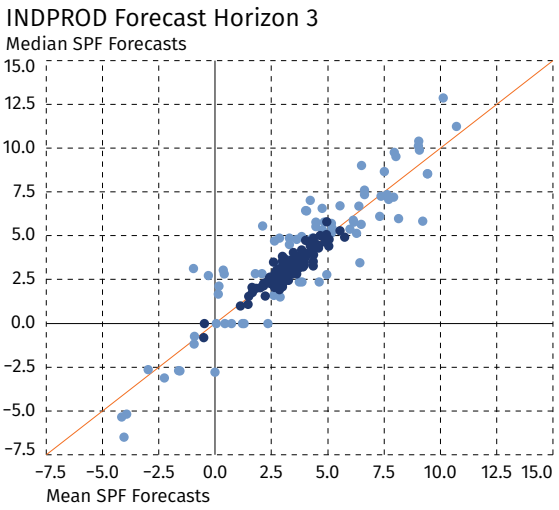
FIGURE 1.

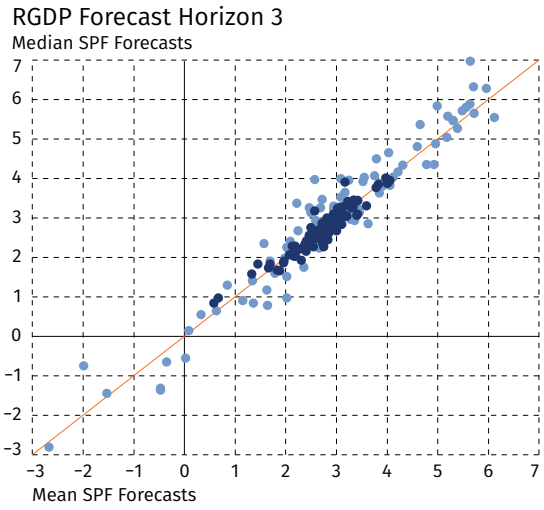
Mean and Median Forecasts, Two Quarters in the Future, 1968–2016

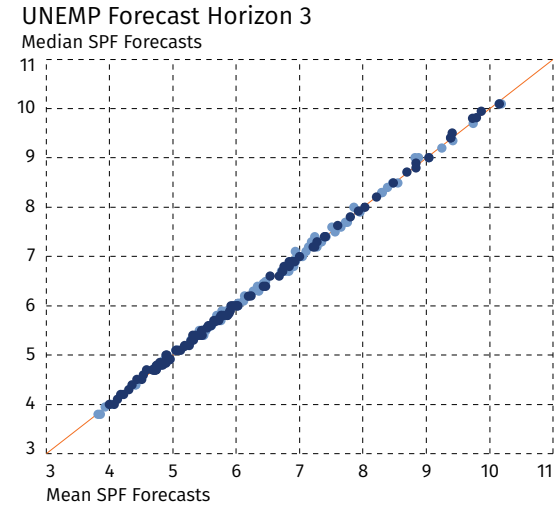
Survey by NBER (1968:4–1990:4) Survey by FRB of Philadelphia (1991:1–2016:4)

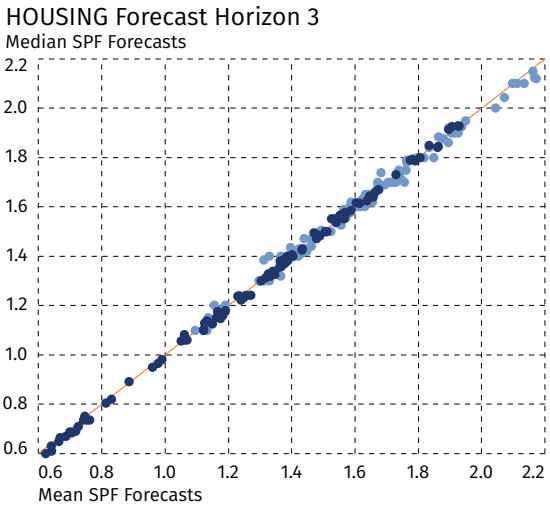












cases in the median forecasts. Over the entire sample period, 1968Q4 to 2016Q4, we found five cases of a statistically significant difference between the mean and median forecasts for real GDP, and, for all the cases, the mean forecast was superior. For the most recent period, from 2005Q1 to 2016Q4, we found nine statistically significant differences between the mean and median projections. The median projection was superior in five of the nine cases, while the mean was superior in the remaining four cases. In contrast with our findings on real GDP projections, we found almost no cases of statistically significant differences between the mean and median forecasts for the GDP inflation rate (Table 2), as we describe in more detail below.[10]

The projections for industrial production stand out because we found many statistically significant comparisons between the mean and median projections. We found 28 statistically significant cases in which the mean was more accurate than the median but only seven cases in which the median was more accurate than the mean (Table 3). However, even where we found statistical significance in the difference between the mean and median projections for industrial production, a close inspection of the detailed results suggests the differences are usually small. Of the 28 cases in which the mean was superior, only three had a percent difference in root-mean-square error (RMSE) greater than 5 percentage points. All other statistically significant differences were associated with a smaller than 5 percentage point difference in the RMSE.

The projections for housing starts also stand out, not only because we found many cases of statistically significant differences between the mean and median forecasts but because the differences were not economically negligible. Table 6 shows that among these cases of statistically significant differences, the median forecasts were superior 30 times, while the mean forecasts were superior 12 times (as measured by p-values of less than 10 percent).[11] In fact, for all 12 cases in which the mean was the superior consensus forecast, the absolute value of the percent difference in the RMSEs was greater than 5 percent. It's interesting to note that all 12 cases appear in the period 1968 to 1990, namely during NBER's tenure. We see in Table 7 that, for the representative case of realizations measured by their final-release values and the sample period from 1968 to 1990, the percent difference in the RMSE was negative and greater than 5 percent in magnitude for the last three horizons, indicating that the mean forecast was more accurate than the median. In our

10 An interesting question we do not address in this paper is whether relative forecast accuracy between the mean and median projections depends on turning points in the economy. Dovern, Fritsche, and Slacalek (2012) have documented a cyclically sensitive cross-sectional forecast variation, a phenomenon that also holds in the SPF. This finding suggests that the number of panelists reporting extreme forecasts in a survey is cyclically sensitive. Further, because extreme projections can weigh more heavily on the mean consensus forecast than on the median, relative forecast accuracy between the mean and median could well depend on the business cycle. In future work, we plan to investigate relative forecast accuracy of the mean and median projections conditional on the business cycle, using the test proposed by Giacomini and White (2006).

11 Operationally, we conduct our Diebold–Mariano tests by running the Diebold–Mariano regression given by $\varepsilon^2_{1,t+\tau|t} - \varepsilon^2_{2,t+\tau|t} = \mu + \upsilon_{t+\tau|t}$, where the dependent variable is the difference in the squared forecast errors, as previously defined; μ is the population difference in the MSE; and $\upsilon_{t+\tau|t}$ is the regression error.

Table 1

Nominal GDP

Sample Period	Median	Mean
1968:4–2016:4	1	1
1968:4–1990:4	1	5
1991:1–2016:4	2	2
2005:1–2016:4	5	0
	9	**8**

Table 2

GDP Price Index

Sample Period	Median	Mean
1968:4–2016:4	0	0
1968:4–1990:4	0	0
1991:1–2016:4	0	0
2005:1–2016:4	0	5
	0	**5**

Table 3

Industrial Production

Sample Period	Median	Mean
1968:4–2016:4	4	9
1968:4–1990:4	0	8
1991:1–2016:4	3	10
2005:1–2016:4	0	1
	7	**28**

Table 4

Real GDP

Sample Period	Median	Mean
1968:4–2016:4	0	5
1968:4–1990:4	0	2
1991:1–2016:4	0	8
2005:1–2016:4	5	4
	5	**19**

Table 5

Unemployment Rate

Sample Period	Median	Mean
1968:4–2016:4	1	3
1968:4–1990:4	1	3
1991:1–2016:4	0	2
2005:1–2016:4	1	1
	3	**9**

Table 6

Housing Starts

Sample Period	Median	Mean
1968:4–2016:4	0	0
1968:4–1990:4	0	12
1991:1–2016:4	18	0
2005:1–2016:4	12	0
	30	**12**

Table 7

Housing Starts: Final-Release Realizations

1968:04–1990:04	$(RMSE_1 / RMSE_2)-1$	P-Value
Horizon 1	0.00500	0.68998
Horizon 2	-0.02299	0.31853
Horizon 3	-0.05809	0.06717
Horizon 4	-0.08251	0.02913
Horizon 5	-0.08512	0.01546

Note: RMSE1 is the root-mean-square error for the mean SPF forecasts, while RMSE2 is the root-mean-square error for the median SPF forecasts.

most recent sample period from 2005 to 2016, we find 12 cases of statistical significance between the mean and median forecasts for housing starts in which the median forecast dominates the mean.

In contrast with the forecasts for industrial production and housing starts, inflation forecasts have the fewest occurrences of statistically significant differences between the mean and median forecasts. Table 2 shows that the inflation forecasts have only five cases in which the mean and median are statistically different. We see all five cases of statistical significance in the most recent period, 2005 to 2016. The results (available on request) show that the fifth forecast horizon always has a percent difference in RMSE greater than or equal to 5 percentage points in magnitude, suggesting that the mean is the better forecast for that horizon.

Conclusion

The Philadelphia Fed has always emphasized the SPF's median projection as the appropriate measure of the survey's consensus forecast. In this *Research Brief*, we test whether the median projection is as accurate a consensus projection as the mean projection. Looking over many variables, alternative sample periods, different measures of realizations, and five quarterly forecast horizons, we generally find statistically insignificant differences between mean and median forecast accuracy. When we do find statistical significance, the difference in forecast accuracy is often small because the percent difference in the RMSE is usually less than 5 percent. The one exception is the survey's forecast for housing starts. For this variable, we more often find statistically and economically significant differences between the mean and median forecasts, with the median forecast generally more accurate than the mean forecast. In thinking about our results in the broadest possible terms, across all variables, sample periods, forecast horizons, and measures of the historical realizations, we conclude that the survey's median projections are as good a choice for measuring the consensus projection.

Of particular interest is that our results suggest generally no differences between the mean and median forecasts for the unemployment rate, real GDP, and inflation. Specifically, for real GDP and unemployment, we found almost always that the mean and median forecasts were statistically indistinguishable or that the percent difference in the RMSEs was usually less than 5 percent. In contrast to real GDP and unemployment, inflation had even fewer occurrences of statistical significance between the mean and median forecasts. RB

References

Diebold, Francis X., and Robert S. Mariano. "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13:3 (July 1995), pp. 253–263.

Dovern, Jonas, Ulrich Fritsche, and Jiri Slacalek. "Disagreement Among Forecasters in G7 Countries," *Review of Economics and Statistics*, 94:4 (2012), pp. 1081–1096.

Giacomini, Raffaella, and Halbert White. "Tests of Conditional Predictive Ability," *Econometrica*, 74:6 (2006), pp. 1545–1578.

Stark, Tom. "Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters," Federal Reserve Bank of Philadelphia *Research Rap Special Report* (May 28, 2010).