Combining forecasts: An Application to Election Forecasts

Andreas Graefe, Karlsruhe Institute of Technology
J. Scott Armstrong, The Wharton School, University of Pennsylvania
Randall J. Jones, Jr., University of Central Oklahoma
Alfred G. Cuzán, University of West Florida

Abstract. A meta-analysis published in 2001 suggested that by combining forecasts one can expect error reductions of about 12%. Research published since then suggests that the gains in accuracy are substantially higher if three or more forecasts are used, especially if based on different methods and different data. Using data on U.S. Presidential elections from 1992 through 2008, this study examined the error reduction obtained by averaging forecasts within and across four groups of methods (polls, the Iowa Electronic Markets prediction market, econometric models, and expert judgment). This combining procedure yielded error reductions ranging from 10% to 58% compared to the average errors of the uncombined individual forecasts.

This is a revised and expanded version of a paper prepared for presentation at the 2011 meeting of the American Political Science Association Seattle, WA, 1-4.

It is well known that combining forecasts improves accuracy. The error of the simple average obtained by combining two or more forecasts will always be no larger than the average of each component's forecast errors. Hereafter we will refer to the former as the *error of the combined forecast* or *the average error,* and to the latter as the *error of the typical forecast* or the *typical error*. If the true value lies within the range of individual forecasts – a situation referred to as *bracketing* (Larrick & Soll, 2006) – the average forecast will always be more accurate than the typical forecast. Over all conditions, combining forecasts protects against picking the worst method and, thus, obtaining large errors.

How combining reduces forecast error

Imagine two methods forecasting a nation's GDP growth rate (assume the actual growth rate turned out to be 1%); method A predicts an increase by 2% and method B predicts an increase by 3%. Then, method A yielded an error of 1 percentage point, while method B missed by two percentage points. Thus, the typical error of an individual method would have been 1.5 percentage points, the average of the two methods' errors. Since no bracketing occurred, the typical individual forecast was as accurate as the average of both forecasts, which was 2.5%. Although combining did not increase accuracy, it did not harm it, and it reduced the likelihood of a larger error.

Now, imagine a situation in which two forecasts yield identical forecast error as in the previous example but lie on either side of (i.e., *bracketing*) the true value. If method C predicts zero growth and method B keeps its forecast of 3%, the forecasts again yield errors of 1 and 2 percentage points, and a typical error of 1.5 percentage points. However, due to bracketing, the average of the two forecasts (i.e., 1.5%) missed the true value only by 0.5 percentage points, an error reduction of 67% compared to the typical individual forecast. (As to calculations, since the typical error is 1.5 and the error from bracketing is 0.5, bracketing reduced the typical error by 1.0. Thus in percentage terms the error reduction is 1.0 divided by 1.5, or 67%.)

Combining forecasts can help even when one knows, in advance, which method is most accurate. For example, averaging the forecast of method C with a (less accurate) forecast from method D that lies within the range 2% and 4% would result in a combined forecast with a lower absolute error than the original forecast of method C. As a general rule, given that the second forecast enables bracketing, it would need to yield an error three times larger than the original forecast in order to harm accuracy (measured in terms of absolute error). See Herzog and Hertwig (2009) or Soll and Larrick's (2009) PAR model for an illustration of when combining is better than picking a single forecast, even if one has complete knowledge about which forecast is more accurate.

Armstrong (2001) conducted a meta-analysis of 30 studies and estimated that, on average, the combined forecast yielded a 12% error reduction compared to the error of the typical forecast; the reductions of forecast error ranged from 3 to 24%. In addition, the combined forecasts were often more accurate than the most accurate component.

Combining is applicable to almost all estimation and forecasting problems. The only exceptions would be where one has strong evidence about which method is best *and* where the likelihood of bracketing is very low.

In the next sections, we discuss conditions of when combining is most useful and report findings from an application of the combining principle for forecasting U.S. presidential elections. By combining forecasts within and across four different methods, large gains in accuracy were achieved, much larger than previously estimated.

Conditions when combining is most useful

Accuracy gains from the combining method are expected to be highest if the component forecasts meet certain *ex ante* conditions: (1) the forecasts draw upon different methods and data, (2) there is uncertainty about which method is best, and (3) a number of evidence-based forecasts can be obtained (Armstrong, 2001). By evidence-based, we mean that they should be consistent with proper forecasting procedures for the given situation.

Use of a number of evidence-based forecasts

Based on his meta-analysis of 30 empirical comparisons, Armstrong (2001) recommended using at least five forecasts. Adding more forecasts helps, though at a diminishing rate of improvement. Nine of these studies were based on combining forecasts from two methods; four of these studies used forecasts from the same method. None of the studies combined forecasts from four or more different methods.

Vul and Pashler (2008) plotted the errors for combinations of a varying number of individual estimates. Error reduction increased as more estimates were included in the combination, although at a diminishing rate, of course. A study by Yaniv and Milyavsky (2007, Exp. 1) revealed similar findings: Given three estimates (i.e., participants' initial estimate and two sources of advice), the average error reduction of was 9% compared to participants' final individual estimates. Given five estimates (four sources of advice), error reduction through averaging was 15%. Further advice did not improve the gains from combining. Given nine estimates (eight sources of advice), error reduction was also 15%.

Use forecasts that draw upon different methods and data

Differences among forecasts can be achieved by selecting forecasts from different valid methods that draw upon different data. The underlying assumption is that forecasts from different methods that use different data are likely share different biases and/or random errors and thus lead to bracketing and low correlations of errors.

Herzog and Hertwig (2009) showed that single individuals could improve their decision-making by thinking in ways that encourage bracketing. In this experiment, 50 of 101 student participants were asked to revise their initial date-estimates of 40 historical events by using a consider-the-opposite strategy. [Mention time lag?] That is, when making their revised estimate, these participants were specifically instructed to draw upon knowledge that they previously ignored or which contradicts their existing beliefs. Compared to the participants' initial estimates, the average of the first and the second estimate reduced error by 4.1%. By comparison, for the control groups that were given no specific instructions for revising their initial estimate, error reduction was only 0.3%. Thus, participants' estimates were more accurate when they used different information and assumptions when making the second estimate. However, two heads were still better than one: combining an individual's first and second estimate did not achieve the gains that could be reached by combining the initial estimates from two different individuals (7.1%).

Vul and Pashler (2008) suggested another approach that enables individual decision-makers to harness the benefits of combining. In this experiment, 428 participants provided initial estimates for eight factual questions. Then, half of the participants were asked to provide a second estimate immediately after completing the questionnaire. The other half provided their second estimate three weeks later, which was assumed to increase the independence of estimates. As expected, error reduction of the average of both estimates compared to the initial estimate was higher in the delayed conditions (about 16%) than in the immediate condition (6.5%). Also, similar to the results by Herzog and Hertwig (2009), combining within individuals was inferior to combining two estimates from different participants.

Batchelor and Dua (1995) analyzed combinations of forecasts made by 22 U.S. economic forecasters. These forecasts differed in their underlying economic theories (e.g., Keynesian, Monetarism, or Supply Side) and methods (e.g., judgment, econometric modeling, or time-series analysis). The authors found that the extent and probability of error reduction through combining were higher the greater the differences in the underlying theory or method of the component forecasts. For example, when combining real GNP forecasts of two forecasters, combining the 5% of forecasts that were most similar in their underlying theory reduced the error of the typical forecast by 11%. By comparison, combining the 5% of forecasts that were most diverse in their underlying theory yielded an error reduction of 23%. Similar effects were obtained regarding the underlying forecasting methods. Error reduction from combining the forecasts derived from most similar methods was 2%, compared to 21% for combinations of forecasts derived from most diverse methods.

Winkler and Clemen (2004) analyzed the relative accuracy of combinations of estimates from different experts and/or different methods. For their analysis, the authors used data from a laboratory experiment in which 90 MBA students estimated the relationship between five pairs of variables (e.g., height & weight or math & verbal SAT scores). For each pair of variables, participants were instructed to use six different methods (e.g., rate the strength of relationship on a scale from 1 to 7 or estimate the correlation coefficient). As expected, the accuracy of combinations from forecasts within and across participants and methods increased with an increasing number of component forecasts. For example, the mean absolute error (MAE) from combining the estimates from five participants was 69% lower than the MAE of the typical participant. Combining five different estimates made by the same participant reduced the error by 34%, compared to a single forecast from this participant. Combining forecasts across participants was generally more accurate than combining across methods (i.e., within participants). On average, combining the forecasts from two different participants was more accurate than combining four forecasts from one participant who used different methods. However, the 'methods' used by participants differed only in how participants were instructed to think about the problem. The participants underlying information did not differ.

Uncertainty about the best method

In most real world situations, analysts do not know about the accuracy of individual sources at the time they make a forecast. Such uncertainty not only makes it difficult to identify the most accurate forecast but also increases the chances that a poor forecast will be chosen.

Hibon and Evgeniou (2005) showed empirically that combining reduces the risks associated with choosing an individual forecast. The authors compared the relative risk associated with two strategies for predicting the 3,003 time series used in the M3-competition based on forecasts from 14 methods: choosing an individual forecast or relying on various combinations of forecasts. Risk was measured as the incremental forecast error results from failing to identify the best individual forecast but randomly picking an individual forecast. Compared to randomly picking an individual forecast, choosing a random combination of all possible combination forecasts reduced risk by 56%.

One way to assess uncertainty is to list all possible relevant methods and to seek rankings from experts. Another is use prior research on which method works best in the situation.

Evidence from a study of election forecasting

Several methods have been used for forecasting U.S. Presidential elections. These include *polls, experts*, *prediction markets*, and *econometric models*. The field of forecasting U.S. presidential elections provides an ideal opportunity for testing the *ex ante* conditions of combining: (1) there are several valid methods that are commonly used for predicting election outcomes, (2) each method uses a different approach and draws upon different data, and (3) it is unclear which method is most accurate.

Polls

Campaign – or trial heat – polls reveal voter support for candidates in an election campaign. Typically, voters are asked which candidate they would support if the election were held today. Thus, polls do not provide predictions but rather snapshots of current opinion. Nonetheless, the media and the public routinely interpret polls as forecasts and project the results to Election Day.

Experts

Another way to forecast elections is to ask unbiased political experts to make a prediction of who will win. Experts are commonly assumed to be relatively unbiased and to have experience in reading polling results, discounting their importance throughout the election cycle, assessing the effect of campaigns and estimating the effects of recent or expected events.

Models

For three decades now, economists and political scientists have used econometric models to estimate the impact of certain variables on the outcome of U.S. presidential elections. Then, the researchers would often use these models to provide a forecast of the two-party vote received by the incumbent party candidate in the next election. As summarized by Jones and Cuzán (2008), most models have between two and five variables and typically include an indicator of economic conditions and a measure of public opinion.

Prediction markets

Prediction (or betting) markets have a long history in election forecasting. Rhode and Strumpf (2004, p.127) studied historical markets that existed for the 15 presidential elections from 1884 through 1940 and found that these markets "did a remarkable job forecasting elections in an era before scientific polling". In comparing forecasts from the *Iowa Electronic Markets* (IEM) to 964 polls for the five presidential elections from 1988 to 2004, Berg et al. (2008) found that 74% of the time the IEM forecasts were closer to the actual election results than polls conducted the same day. (However, Erikson and Wlezien (2008) found polls to be more accurate than IEM forecasts when the polls' pre-election lead times were discounted by regressing the vote on polls taken at comparable times across elections.)

Combining procedure

A two-step procedure was used to combine forecasts *within* and *across* the four component methods. Combining *within* component methods was used to group and combine forecasts that use a similar approach and/or similar information. The goal of this first step was to reduce the impact of a large number of forecasts that use a similar method or draw upon similar information. For example, while only one prediction market was available, we could draw on a large number of model forecasts, many of which were based upon a similar method and/or information. In such a situation, a simple average of all available forecasts would over-represent models and under-represent prediction markets. We then averaged the forecasts from these four component methods. We expected this procedure to meet the *ex ante* conditions for combining described earlier.

Predictions from polls, models, and the IEM were available for the five presidential elections held between 1992 and 2008. Expert forecasts were available for only the two most recent elections.

Within component combining

We combined polls by calculating rolling averages of three most recent polls. In case more than three polls were published per day, all polls of that day were averaged. That way, we obtained one *poll average* forecast per day. Over all five elections, we used the results from 773 polls: 93 in 1992, 84 in 1996, 329 in 2000, 112 in 2004 and 155 in 2008.

For the 2004 and 2008 elections, we formed a panel of experts and contacted them periodically for their estimates of the incumbent's share of the two-party vote on Election Day. Most experts were from the ranks of academia, though a few were at think tanks, in the media, or former politicos. We deliberately excluded election forecasters who developed econometric models, because that method is represented as a separate component. The number of respondents in the three surveys conducted in 2004 ranged from 12 to 17. For the four surveys in 2008, the number of respondents ranged from 9 to 13. Our *experts'* forecast is the simple average of individual expert forecasts.

Model averages were recalculated whenever new or updated individual model forecasts became available. For the 2008 election, we averaged forecasts from 16 quantitative models. Ten models were used in 2004. For the retrospective analyses of the elections in 1992, 1996, and 2000, we included the forecasts of 4, 8, and 9 models, respectively. Forecasts for most models were available in July and August, and several were updated in mid- or late-October, as revised data became available.

We combined IEM forecasts by calculating 7-day rolling averages of the vote-share contract for the incumbent party's candidate. For the forecasts from 1992 to 2004 we used the average daily trading price. For 2008, we used the price of the last trade of the day.

Across component combining

For each day in the forecast horizon, we calculated simple averages across the already combined component forecasts (i.e., polls, experts, models, and IEM). We refer to this combined forecast as the *PollyVote*. The PollyVote forecasts were updated several times per week or even daily during the run-up to the 2004 and 2008 elections at pollyvote.com.

Expected accuracy gains from combining

It was expected that the gains from combining would depend on the combining procedure as well as the conditions. As noted previously, the primary goal of combining *within* methods was to group redundant forecasts in order to create a set of four different component forecasts. We expected error reductions from combining forecasts that use similar methods and/or draw upon similar information to be similar to the 12% reported in prior research (Armstrong 2001). But we expected large gains from combining forecasts when the component methods differed in their underlying method and data.

Combining polls

The information generated by any single poll is unreliable for forecasting the outcome of the election, especially early in the campaign. Polls conducted by reputable survey organizations at about the same time often reveal considerable variation in results. Common explanations for these discrepancies are sampling error, non-response bias, or response bias (Donsbach & Traugott 2008, pt. III). Combining was expected to mitigate the problem of variation among polls.

Combining experts' forecasts

Expert forecasts have been shown to be of some value in political forecasting. For predicting the outcome of two controversial ballot measures, Lemert (1986) found politicians' predictions to be more accurate than predictions made by members of the mass public.

We located two surveys of experts and pundits that were conducted shortly before the 1992 and 2000 U.S. presidential elections and calculated the gains from combining the individual predictions. In 1992, the average forecast of ten expert predictions was 4% more accurate than the typical forecast.[i] In 2000, the combined forecast was 72% more accurate than the typical forecast from 15 experts. [ii]

Because our panelists did not meet in person, the possibility of bias due to the influence of strong personalities or individual status was eliminated. Furthermore, different experts can be expected to use different approaches and rely on different data sources when generating their forecasts.

Combining models' forecasts

Most econometric models are usually able to predict the correct election winner far in advance. However, the individual track record in predicting the actual vote shares of these models is mixed and forecast errors for a single model can vary widely across elections. In such a situation, it is difficult to identify the most accurate model.

Bartels and Zaller (2001) used data from the 13 U.S. presidential elections from 1948 to 1996 and varying combinations of six economic and three political variables to estimate 48 regression models. The authors then analyzed the performance of these 48 models for the 2000 election. Using their data, we calculated an error of 3.0 percentage points for the typical model. By comparison, the average forecast of all models was off by 2.5 percentage points, which amounts to an error reduction of 17%.

Combining IEM forecasts

We expected 7-day averages to moderate overreactions of the market due to information cascades, which can cause unexpected positive or negative spikes in prices. Information cascades occur when people defer their private information but rely on the information publicly revealed by others. See Anderson and Holt (1997) who conducted a simple experiment to create information cascades in the laboratory. So the challenge for combining is whether the gain from controlling cascades is superior to the cost of not having the most up-to-date information.

Unlike for the other three components, accuracy of the combined IEM forecasts was not assessed by comparing the 7-day average to the error of the typical IEM forecast within this period. Instead, accuracy was compared to the most recent IEM forecast. Note that this approach does not ensure that the combined forecast will be at least as accurate as the most recent forecast.

Combining: across components

While each of the four component methods can be expected to produce valid forecasts*, a priori* it is unclear which method is most accurate. In addition, each component uses a different method and draws upon different data. In such situations, the gains from combining across components are expected to be high; the combined forecast might even be more accurate than the most accurate component forecast. However, this becomes less likely as the number of components increases.

Results

All forecasts refer to the popular two-party vote share of the candidate from the incumbent party. We used the absolute error as a measure of accuracy (i.e., the difference between the predicted and actual vote shares, regardless whether positive or negative) as a measure of accuracy. For the five U.S. presidential elections from 1992 to 2008, we calculated daily forecasts for each of the 93 days prior to Election Day. Thus, we obtained 465 daily forecasts from polls, models, and the IEM. Our own expert forecasts were available only for the two elections in 2004 and 2008 with a total of 186 additional forecasts.

Table 1: Accuracy gains from combining (Mean error reduction in %)

	1992	**1996**	**2000**	**2004**	**2008**	**Mean**
Within component combining						
Poll average vs. typical poll	6	11	15	18	22	**14**
Model average vs. typical model	5	45	0	8	47	**21**
Combined experts vs. typical expert	na	na	na	24	11	**18**
7-day IEM average vs. original IEM	-2	11	7	13	14	**9**
Across components combining: PollyVote vs.						
Poll average	85	52	30	62	19	**50**
Model average	82	-33	69	75	-31	**32**
Experts	na	na	na	55	31	**43**
IEM (7-day average)	69	-40	-23	4	-3	**1**
Within and across combining: PollyVote vs.						
Typical individual poll	86	57	40	69	37	**58**
Typical individual model	83	17	69	78	45	**58**
Typical individual expert	na	na	na	65	39	**52**
Original IEM	69	-32	-17	17	11	**10**

Accuracy gains from combining within components

The "Within component combining" section of Table 1 shows the mean error reduction (MER) achieved through combining within components over the whole forecast horizon compared to the typical component forecast. Overall, MER from combining within polls was 14%. Gains were high when combining within models (21%) and expert forecasts (18%). Calculating 7-day averages of IEM prices reduced the error of the original IEM by 9%; this combining procedure yielded more accurate forecasts than the original IEM except for 1992.

Accuracy gains from combining across components

The "Across components combining" section of Table 1 shows the MER of the PollyVote forecast compared to the error of the combined component forecasts. On average, the PollyVote forecast was 50% more accurate than the combined poll average, 32% more accurate than the combined model average, and 42% more accurate than the combined experts. Compared to the 7-day IEM forecasts, error reduction was small (1%). In addition, the 7-day IEM forecasts were more accurate than the PollyVote in three elections and less accurate in two.

Accuracy gains from combining within and across components

The "Within and across combining" section of Table 1 shows the MER of the PollyVote forecast compared to the typical (uncombined) component forecasts. Gains in accuracy were large compared to the typical individual model and poll (MER: 58%) and the typical expert (52%). Compared to the original IEM, the PollyVote reduced the error by 10% on average.

Accuracy gains for different combinations of component methods

Table 2 shows the percentage of days in which bracketing occurred and the MER compared to the typical component for all possible combinations of component methods. The numbers were calculated ex post from 186 daily forecasts from the two elections in 2004 and 2008, for which forecasts from all four methods were available. As expected, a higher incidence of bracketing yielded larger error reductions.

Combinations of two component methods

On average, combining across two methods led to a 21% error reduction relative to the typical forecast. Combinations that included the model forecasts yielded the largest gains in accuracy, in particular the combined forecast of models and experts (error reduction: 42%). The probabilities of bracketing (26%) and MER (9%) were smallest for the combination of IEM and expert forecasts. One explanation for this might be the similarity of both methods as they allow for incorporating all available information. Another explanation is that the experts were likely to have consulted the IEM when making their forecasts. Gains in accuracy were also small when combining polls and the IEM, because the IEM incorporates information from polls.

Combinations of three component methods

On average, the combinations of three models led to error reductions of 32% relative to the typical forecast. The error reductions were largest (42%) if the combined forecast included information from the models. The error reductions were smallest – although still at the substantial level of 21% – for the combination of polls, experts, and the IEM.

Table 2: Bracketing and mean error reduction for different combinations of component methods

Combinations based on		% of days with bracketing	MER to typical component (in %)
Two component methods			
Models & experts		50	42
Models & IEM		48	23
Models & polls		38	23
Polls & experts		37	18
Polls & IEM		32	9
IEM & experts		26	9
	Mean	**41**	**21**
Three component methods			
Models & IEM & experts		62	42
Models & polls & experts		62	35
Models & polls & IEM		59	29
Polls & IEM & experts		47	21
	Mean	**59**	**32**
Four component methods		**67**	**35**

Combinations of four component methods

The combination of four methods led to an error reduction of 35% relative to the typical forecast. In two out of three cases, combining the forecasts from all four component methods yielded bracketing.

Benefits of combining forecasts under uncertainty

Another benefit of combining is that it allows for assessing uncertainty. If forecasts derived from different methods agree, certainty about the situation increases. By comparison, high disagreement among forecasts indicates high uncertainty. Disagreement among forecasts has often been used as an *ex ante* measure for uncertainty. However, in analyzing 2,787 observations for inflation and 2,342 observations for GDP forecasts from the Survey of Professional Forecasters, Lahiri and Sheng (2010) confirmed evidence from earlier research showing that disagreement generally tends to underestimate the level of uncertainty.

Table 3 shows the MAE of the PollyVote and the typical component as well as the MER of the

PollyVote compared to the typical component for different levels of uncertainty. Uncertainty was measured as the range between the component forecasts of a certain day with the lowest and the highest forecast. We then calculated the quartiles for the ranges of the 186 daily forecasts from the two elections in 2004 and 2008. Low uncertainty refers to the forecasts in the lower quartile and vice versa. Medium uncertainty refers to the forecasts in the interquartile range.

Table 3: Error reduction of combining under uncertainty

Uncertainty	Range of forecasts	No. of days	MAE of the PollyVote	MAE of the typical component	Mean error reduction vs. typical component (%)
Low	0 to 1.7	47	0.9	1.0	11
Medium	1.7 to 3.9	93	1.0	1.4	28
High	3.9 to 5.5	46	0.7	1.7	61

The MER of the PollyVote compared to the typical component increased with increasing uncertainty. In situations with low uncertainty, the PollyVote reduced the error of the typical component forecast by 11%. In situations with high uncertainty, combining yielded a 61% error reduction. In sum, the benefits from combining were larger when disagreement among component forecasts was higher.

In applying a two-step approach of combining forecasts within and across four methods for forecasting U.S. presidential elections, we achieved large gains in accuracy. Compared to forecasts from a randomly chosen poll, model, or expert, our PollyVote forecast reduced error by 52% to 58%. Compared to the original IEM, essentially an approach for aggregating and combining dispersed information, the PollyVote reduced error by 10%. Combining is especially useful to avoid large errors and in situations involving high uncertainty.

Combining should be applicable to predicting other elections and, more generally, can be applied in many other contexts, as well. Given the methods available to forecasters, combining is an effective way to improve forecast accuracy and to prevent large errors. However, despite its usefulness and ease of use, combining is not often used.

Barriers to combining

Lack of knowledge about the research on combining is likely to be a major barrier for the use of combining in practice. The benefits of combining are not intuitively obvious and people are unable to learn this through their experience. In a series of experiments with highly qualified MBA students, a majority of participants thought that the average of estimates would reflect only average performance (Larrick & Soll 2006).

Is combining too simple? Hogarth (in press) reported results from four case studies showing that simple models often predict complex problems better than more complex ones. In each case, people had difficulties to accept the findings. There is a strong belief that complex models are necessary to solve complex problems. Similarly, people might perceive the principle of combining as "too easy to be true".

Forecasters might seek an extreme forecast in order to gain attention. Batchelor (2007) found long-term macroeconomic forecasts to be consistently biased as a result of financial, reputational, or political incentives of forecasting institutions. Only in short-term forecasting horizons he found individual forecasts to converge to the more accurate consensus forecast. Forecasters face a general trade-off between accuracy and attention: more extreme forecasts usually gain more attention and the media is more likely to report them.

Forecasters may think they are already using combining properly. Based on the findings from his meta-analysis, Armstrong (2001) recommended combining forecasts mechanically, according to a predetermined procedure. A general rule is to weight forecasts equally, unless there is strong prior evidence that supports differential weights. In practice, managers often use unaided judgment to assign differential weights to individual forecasts. Such an *informal* approach to combining is likely to be harmful as managers can select a forecast that suits their biases.

People mistakenly believe they can identify the most accurate forecast. Soll and Larrick (2009) conducted experiments to examine the strategies people use to make decisions based upon two sources of advice. Instead of combining the advice, the majority of participants tried to identify the most accurate source – and thereby harmed accuracy.

Conclusion

A meta-analysis by Armstrong (2001) estimated a 12% error reduction due to combining forecasts. Many of these studies involved combining forecasts from only two sources, and in most cases the sources were similar. Subsequent studies indicated that gains could be substantially larger when combining more forecasts, especially when the underlying methods and data differ.

Further evidence was obtained in the present study on forecasting U.S. presidential elections, a situation that involves a number of different valid forecasting methods that use different data sources. This situation also allowed for combining forecasts within and across component methods, a procedure that had not previously been tested.

Averaging within and across forecasts from four component methods improved the accuracy of election forecasts. Average gains from within component combining ranged from 9% to 21%. Combining across components yielded forecasts that were, on average, 1% to 50% more accurate than the four component forecasts. In addition, the combined forecast performed about as well as the best component forecast, a 7-day rolling average of IEM prices.

Compared to the typical uncombined individual forecast, the combining procedure yielded mean error reductions ranging from 10% to 58%.

Acknowledgments

Kesten Green and Stefan Herzog provided helpful comments. We also received suggestions when presenting earlier versions of the paper at the 2009 *International Symposium on Forecasting*, the 2010 *Bucharest Dialogues on Expert Knowledge, Prediction, Forecasting: A Social Sciences Perspective,* and the *2011 APSA Annual Meeting*. We sent drafts of the paper to all authors whose research was cited on substantive points to ensure that we accurately summarized their research and thank all those who replied. Kelsey Matevish helped to edit the paper.

References

Anderson, L. R. & Holt, C. A. (1997). Information cascades in the laboratory, *American Economic Review*, 87, 847-862.

Armstrong, J. S. (2001). Combining forecasts. In: J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Norwell: Kluwer, pp.417-439.

Bartels, L. M. & Zaller, J. (2001). Presidential vote models: A recount. *PS: Political Science & Politics*, 34, 9-20.

Batchelor, R. (2007). Bias in macroeconomic forecasts, *International Journal of Forecasting*, 23, 189-203.

Batchelor, R. & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, 41, 68-75.

Berg, J. E., Nelson, F. D. & Rietz, T. A. (2008). Prediction market accuracy in the long run, *International Journal of Forecasting*, 24, 285-300.

Donsbach, W. & Traugott, M. W. (2008). *Sage Handbook of Public Opinion Research,* Thousand Oaks, CA: Sage Publications.

Erikson R. S. & Wlezien, C. (2008). Are Political Markets Really Superior to Polls as Election Predictors? *Public Opinion Quarterly*, 72, 190-215.

Herzog, S. M. & Hertwig, R. (2009). The wisdom of many in one mind. Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20, 231-237.

Hibon, M. & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting*, 21, 15-24.

Hogarth, R. (in press). When simple is hard to accept. In P. M. Todd, G. Gigerenzer, & The ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world*. Oxford: Oxford University Press.

Jones, R. J. & Cuzán, A. G. (2008). Forecasting U.S. presidential elections: A brief review. *Foresight – The International Journal of Applied Forecasting,* Summer 2008, 29-34.

Lahiri, K., & Sheng, X. (2010). Measuring forecast uncertainty by disagreement: The missing link. *Journal of Applied Econometrics*, 25, 514-538.

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

Lemert, J. B. (1986). Picking the winners: Politician vs. voter predictions of two controversial ballot measures. *Public Opinion Quarterly*, 50, 208-221.

Rhode, P. W. & Strumpf, K. S. (2004). Historical presidential betting markets, *Journal of Economic Perspectives*, 18, 127-141.

Soll, J. B. & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 780-805.

Vul, E. & Pashler, H. (2008). Measuring the crowd within: probabilistic representations within individuals. *Psychological Science*, 19, 645-647.

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

Yaniv, I. & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, 103, 104-120.

[i] *The Washington Post*. Pundits' brew: How it looks; Who'll win? Our fearless oracles speak, November 1, 1992, Page C1, by David S. Broder.
[ii] *The Hotline*. Predictions: Potpourri of picks from pundits to professors, November 6, 2000.