

Combining Forecasts: An Application to American Presidential Elections

Andreas Graefe, LMU University of Munich, Germany

J. Scott Armstrong, The Wharton School, University of Pennsylvania, USA

Randall J. Jones, Jr., University of Central Oklahoma, USA

Alfred G. Cuzán, University of West Florida, USA

January 30, 2012

Abstract. Combining forecasts is simple, inexpensive, and effective, yet few organizations use it. This occurs because the benefits of combining are counter-intuitive and because people are unaware of the empirical evidence. Here we summarize the literature on the effectiveness of combining by assessing the conditions under which it is most effective. Using data on U.S. Presidential elections from 1992 through 2008, we then examine the error reduction obtained by averaging forecasts within and across four groups of methods (polls, the Iowa Electronic Markets, quantitative models, and expert judgment). The gains in accuracy from combining increased with the number of component forecasts used, especially when these forecasts were based on different methods and different data. Combining yielded error reductions ranging from 5% to 64% compared to the average errors of the uncombined individual forecasts; this improvement is substantially greater than the 12% that had been previously reported for combining. We also show combining is especially effective in situations involving high uncertainty.

1. On the history of combining forecasts

Combining has a rich history, not only in forecasting. In 1818, Laplace wrote "in combining the results of these two methods, one can obtain a result whose probability law of error will be more rapidly decreasing" (as cited in Clemen, 1989). In using photographic equipment to combine portraits of people, Galton (1879, 135) found that "all composites are better looking than their components, because the averaged portrait of many persons is free from the irregularities that variously blemish the look of each of them." For the field of population biology, Levins (1966) noted that, rather than striving for one master model, it is often better to build several simple models that, among them, use all the information available, and then average them. These early applications of combining related to estimation problems rather than forecasting. Such studies were common in psychology in the early 1900s; Zajonc (1962) summarizes this literature.

As with estimation problems, research in forecasting has shown that combining improves accuracy. Interest in this stream of research has been growing, particularly since publication of a prominent paper by Bates and Granger (1969). Much of the early interest was in how to best weigh the components, and many statisticians proposed schemes for doing so. But after reviewing 209 published papers, Clemen (1989) concluded that averaging (i.e., using equal weights) is often the best course of action when combining. This study became one of the most frequently cited papers in forecasting; as of 2011, it had attracted more than 1,000 citations in Google Scholar. Armstrong (2001) also reviewed the literature on combining forecasts. In his meta-analysis of 30 studies, he estimated that, on average, combined forecasts 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. In effect, the relatively simple method of combining is one of the most effective forecasting techniques available to forecasters. In this paper, we apply the technique to presidential elections forecasting.

2. Why combining reduces forecast error

Mathematically, the error obtained by averaging two or more forecasts will never be larger than the average of the components' 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*. The typical error is the error that one can expect if one would randomly pick the forecast of a single component. If the true value lies within the range of individual forecasts – the phenomenon of *bracketing* (Larrick & Soll, 2006) – the average forecast will always be more accurate than the typical forecast. Over all conditions, combining forecasts protects against selecting

the worst method and thus helps avoid large errors.

Here is an example from American presidential elections that illustrates the benefits of combining. In 2008, the incumbent candidate, John McCain, received 46.3% of the two-party vote. Abramowitz's "time for change" model (2008) predicted McCain to take 45.7%, which yielded an error of 0.6 percentage points. The model by Holbrook (2008) forecasted a vote-share of 44.3%, missing the actual outcome by 2.0 percentage points. Thus, the average of the errors of the two models, what we call the typical error, was 1.3 percentage points. Both models under-predicted the outcome, so no bracketing occurred; hence, the typical individual forecast was as accurate as the average of both forecasts (which was 45.0%). Combining did not increase accuracy, but it did not harm it either, and it did reduce the likelihood of picking the model that incurred the largest error.

Now, imagine a situation in which two forecasts lie on either side of the true value, bracketing it. The 2008 forecast made by Erikson & Wlezien with their model (2008b) was 47.8%, which corresponds to an error of 1.5 percentage points. Thus, the typical error of the two models by Holbrook and Erikson & Wlezien is 1.8 percentage points ($= [2.0 + 1.5] / 2$). Although this error is larger than in the previous example, the average of the two forecasts (i.e., 46.1%) misses the true value only by 0.2 percentage points.

Combining forecasts can help even when one knows, in advance, which method is most accurate. It is well known that Abramowitz's model has been consistently among the most accurate models. But the average of the models by Abramowitz and Erikson & Wlezien (i.e., 46.8%) yields an error of 0.5 percentage points, which is smaller than the error incurred with Abramowitz's model. In general terms, the average of two forecasts that bracket the truth is always more accurate than the best individual forecast, if the error of the less accurate forecast does not exceed three times the error of the most accurate forecast (measured as absolute error). See Herzog and Hertwig (2009) or Soll and Larrick's (2009) PAR model for an illustration as to when combining is better than picking a single forecast, even if one has complete knowledge about which forecast is the most accurate.

One intuitive explanation as to why combining helps is that it allows forecasters to use more information and to do so in an objective manner. Moreover, bias exists in the selection of data and in the forecasting methods that are used. Often the bias is unique to the data and to the method, so that when various methods using different data are combined in making a forecast, bias tends to be diffused and thus reduced.

Combining is applicable to almost all estimation and forecasting problems. The only exception would be where strong evidence exists that one method is best *and* the likelihood of bracketing is very low. In the next sections, we discuss conditions as to when combining is most

useful, and we report findings from an application of combining 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.

3. Conditions when combining is most useful

Armstrong (2001) proposed certain *ex ante* conditions under which accuracy gains from combining are expected to be highest: (1) a number of evidence-based forecasts can be obtained, (2) the forecasts draw upon different methods and data, and (3) there is uncertainty about which forecast is most accurate.

3.1 Use of a number of evidence-based forecasts

By “*evidence-based forecasts*,” we mean forecasts used for combining that are generated with methods that are consistent with proper forecasting procedures for the given situation. (A useful tool in making this assessment is the Forecasting Audit at forprin.com.) Armstrong (2001) recommended using at least five forecasts. Adding more forecasts helps, though at a diminishing rate of improvement. Nine of the 30 studies in his meta-analysis 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. The size of the error shrank as more estimates were included in the combination, although, again, at a diminishing rate. A study by Yaniv and Milyavsky (2007, Exp. 1) revealed similar findings: Given three estimates--the participants’ initial estimate and two sources of advice--the error of the combined forecast was 9% smaller than that of the typical participant’s final estimate. Given five estimates (four sources of advice), error reduction through averaging amounted to 15%. Further advice did not improve the gains from combining.

3.2 Use of forecasts that draw upon different methods and data

Differences among forecasts can be obtained by selecting forecasts from different methods that draw upon different data. The underlying assumption is that such a set of forecasts is likely to include different biases and random errors and thus should lead to bracketing and low correlations of errors.

For example, Batchelor and Dua (1995) analyzed combinations of 22 U.S. economic forecasts that differed in their underlying 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 larger the differences in the underlying theories or methods 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 the most similar methods was 2%, compared to 21% for combinations of forecasts derived from the most diverse methods. Winkler and Clement (2004) reached a similar conclusion.

3.3 Uncertainty about the best forecast

Identifying the most accurate forecast among a set of forecasts is a difficult task. In most real world situations, analysts do not know about the relative accuracy of individual sources at the time they make a forecast. The greater the uncertainty, the more difficult it is to identify the most accurate forecast, and the more likely 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 that 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 independent rankings from experts. The Delphi method might be used to obtain these judgments. Another possibility is to use prior research on which method works best in the situation. However, as described earlier, combining can increase accuracy even if one knows in advance which component forecast is most accurate.

4. Evidence from a study of election forecasting

Several valid methods are commonly used for predicting election outcomes. These include polls, experts' judgment, quantitative models, and prediction markets. Each of these methods uses a different approach and draws upon different data. This renders the field ripe for exploiting the advantages of combining.

4.1 Combining procedure

Our initial approach to combining presidential election forecasts was to weigh all methods equally. In a two-step procedure, forecasts were combined both *within* and then *across* component methods. Combining *within* each method entailed averaging forecasts that were based on similar approaches or similar information. This procedure equalized the impact of each method, whether a method had generated many forecasts or only a few. For example, while only one suitable prediction market was available, there were forecasts from several econometric models that used a similar method and similar information. A simple average of all available forecasts would over-represent models and under-represent prediction markets. After combining forecasts within each method, we then averaged the combined forecasts from the four component methods.

4.1.1 Within component combining

Predictions from polls, models, and the IEM were available for the five elections held between 1992 and 2008. In addition, we conducted our own expert surveys for the two most recent elections.¹ We expected error reductions from combining forecasts that use similar methods and/or draw upon similar information (i.e., within component combining) to be similar to the 12% reported in prior research (Armstrong 2001).

4.1.1.1 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, polls are a common means of forecasting election outcomes. Scholars, the news media, and the public commonly interpret polls as forecasts and project the results to Election Day (Hillygus, 2011).

Campbell and Wink (1990) analyzed the accuracy of Gallup trial-heat polls for the 11 U.S. Presidential Elections from 1948 to 1988. Their findings suggested that using raw polls to forecast presidential elections yields large errors, particularly before the fall campaign begins. Other research has shown that polls conducted by reputable survey organizations at about the same time often reveal considerable variation in results. Errors caused by sampling problems, non-responses,

¹ For the past two elections in 2004 and 2008, we provided *ex ante* forecasts that were continuously updated throughout the campaigns and posted at www.pollyvote.com. In the present study, we report all forecasts as if they were calculated *ex post*.

inaccurate measurement, faulty processing, and house effects impact the accuracy of polls and the quality of surveys more generally (Donsbach & Traugott, 2008; Wlezien, 2003).

Combining polls with structural variables increases forecast accuracy. Models by Abramowitz (2008), Campbell (2008), Lewis-Beck and Tien (2008), and Wlezien and Erikson (2008b) all include a variable measuring opinion (presidential approval or support for the incumbent candidate) along with economic data. These models, as well as others, are considered in Section 4.1.1.3.

Using the median of all state-level polls taken within a month of the presidential election, Gott and Colley (2008) correctly predicted Bush's victory over Kerry in 2004 with an error of only 4 electoral votes. They also forecast Obama to win over McCain in 2008 with an error of only 2 electoral votes. In both elections, the median statistics approach missed the winner in only one state.

Simply aggregating polls has become popular in the news media and reduces potential biases noted previously. Well-known poll aggregators include realclearpolitics.com and the Huffpost Pollster (formerly Pollster.com), which provide combined poll projections on an almost daily basis.

Hillygus (2011) has noted several approaches to combining polls, and at this point finds no one best practice for doing so. In this study, we combined polls by calculating rolling averages of all polls released over a 7-day period for each of the last 100 days prior to Election Day. Across all five elections, we used results from 1809 polls: 250 in 1992, 545 in 1996, 467 in 2000, 351 in 2004 and 196 in 2008. All polls were obtained from the *iPoll databank* of the *Roper Center for Public Opinion Research*.

4.1.1.2 Experts

Before the emergence of polls in the 1930s, judgments from political insiders and experienced observers were commonly used for forecasting (Kernell, 2000). They still are. Expert analysts are assumed to be independent when making predictions, and they have experience in reading and interpreting polls, assessing their significance during campaigns, and estimating the effects of recent or expected events on their results.

Experts can be expected to use different approaches and rely on various data sources when generating their forecasts. Thus, combining experts' judgments should increase forecast accuracy. We were unable to find prior studies on the gains from combining expert forecasts of election results. However, we did locate two expert surveys 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 forecast of the typical individual expert.² In 2000, the average forecast was 72% more accurate than the typical forecast from 15 experts.³

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 popular vote on Election Day. Most experts were academic specialists in elections, though a few were analysts at think tanks, commentators in the news media, or former politicians. We deliberately excluded election forecasters who developed their own models because that method was represented as a separate component in our combined forecast (see Section 4.1.1.3). The number of respondents in each of the three surveys conducted in 2004 ranged from 12 to 16. For the four surveys in 2008, the number of respondents ranged from 10 to 13. Our combined expert forecast was the simple average of forecasts made by the individual experts. Because our panelists did not meet in person, the possibility of bias due to the influence of strong personalities or individual status was eliminated.

4.1.1.3 Models

A common explanation of electoral behavior is that elections are referenda on the incumbent party's performance during the term that is ending. For more than three decades, scholars have amplified and tested this theory, most commonly by developing econometric models, usually to predict the outcome of U.S. presidential elections. Most models include two to five variables and typically include indicators of economic conditions and public opinion to measure the incumbent's performance (Jones & Cuzán, 2008). For descriptions of early election forecasting models and other methods, see Lewis-Beck and Rice (1992), Campbell and Garand (2000), and Jones (2002).

Since the 1990s, forecasts of competing models have been regularly published near Labor Day of the election year. For the past four elections, the forecasts of leading models were published in *American Politics Research*, 24(4) and *PS: Political Science and Politics*, 34(1), 37(4), and 41(4).

Most of the models have produced forecasts using data available near the end of July of the election year. Usually, the models have correctly predicted the election winner, albeit by varying accuracy concerning candidates' vote shares. Moreover, forecast errors for a single model can vary widely across elections, and the structure of the some of the models has changed over time. It is,

² *The Washington Post*. Pundits' brew: How it looks; Who'll win? Our fearless oracles speak, November 1, 1992, p. C1, by David S. Broder.

³ *The Hotline*. Predictions: Potpourri of picks from pundits to professors, November 6, 2000.

therefore, difficult to identify the most accurate model.

Bartels and Zaller (2001), using various combinations of variables from prominent presidential election models, constructed 48 models to generate ex post forecasts of the 2000 election outcome. By using the Bartels and Zaller data, we determined the forecast error for the typical model within the group of 48 to be 3.0%. By comparison, the average forecast error for all models was 2.5%. This is an error reduction of 17% gained by combining. Montgomery *et al.* (2011) used *ensemble Bayesian model averaging (EMBA)* to combine the forecasts from six established econometric models based on their past performance and uniqueness. For the nine elections from 1976 to 2008, the error of the combined EMBA forecast was 34% lower than the error of a typical individual model.⁴

We recalculated model averages whenever new or updated individual model forecasts became available. Forecasts were available from 6 models in 1992, 8 in 1996, 9 in 2000, 10 in 2004, and 16 in 2008. Forecasts for most models were available in July and August, and some were updated once or more often as revised data became available. All models were developed by academics and either published in scientific journals or presented at academic conferences.⁵

4.1.1.4 Prediction markets

Prediction (or betting) markets have a long history in election forecasting. Rhode and Strumpf (2004, 127) studied historical markets that existed for the 15 presidential elections from 1884 through 1940 and concluded that these markets “did a remarkable job forecasting elections in an era before scientific polling”. In comparing forecasts from the *Iowa Electronic Markets (IEM)* with 964 polls for the five presidential elections from 1988 to 2004, Berg et al. (2008) found that 74% of

⁴ The six models were the models by Abramowitz (2008), Campbell (2008), Erikson & Wlezien (2008b), Fair (2009), Hibbs (2000), and Lewis-Beck & Tien (2008).

⁵ Model forecasts by Abramowitz (2008), Campbell (2008), Fair (2009), and Erikson & Wlezien (2008b) were available for all five elections. Forecasts by Holbrook (2008), Lewis-Beck and Tien (2008), Lockerbie (2008), and Norpoth (2008) were available for the four elections from 1996 to 2008. Forecasts by Cuzán and Bundrick (2008) were available for the three elections from 2000 to 2008. Forecasts by Hibbs (2000) were available for the elections in 2004 and 2008. Forecasts by Lichtman (2008), Graefe and Armstrong (2011), Jérôme and Jérôme (2011), Haynes and Stone (2008), DeSart and Holbrook (2003), and Klarner (2008) were available for the 2008 election. A forecast by Lewis-Beck and Rice (1992) and Sigelman (1994) was available for the 1992 election.

the time the IEM forecasts were closer to the actual election results than polls conducted the same day. However, Erikson and Wlezien (2008a) found polls to be more accurate than the IEM forecasts when they were discounted to control for lead time before the election.

Prediction market forecasts can be negatively affected by unexpected positive or negative spikes in prices due to information cascades. Information cascades occur when people defer their private information but rely on the information publicly revealed by others (Anderson & Holt, 1997). We expected that combining market forecasts over a certain period of time could moderate such short-term disruptions of market prices. We combined IEM forecasts by calculating 7-day rolling averages of the daily average of the two-party vote-share contract for the incumbent party's candidate. We then assessed the accuracy of combined IEM forecasts by comparing the 7-day average to the daily IEM average.

4.1.2 *Across component combining*

Although some previous research has assessed the value of combining election forecasts within methods, we are not aware of any prior research that has combined forecasts both *within and across* methods, which is the approach we took. Each of the four component methods in our study could be expected to produce valid forecasts, but we anticipated that the most significant gains in accuracy would come from combining across the methods. This is because the four methods differ significantly in technique and assumptions, in the types of data used, and in data sources. We recognized that the demonstrated accuracy of the Iowa Electronic Markets in predicting elections might diminish the gains from combining across methods. We also were aware, however, that the impact of a dominant method tends to diminish as the number of components increases.

For each day in the forecast horizon, we calculated a simple average across the combined component forecasts: polls, experts, models, and IEM. We refer to this overall combined forecast as the *PollyVote*.

4.2 **Results**

All forecasts reported refer to the two-party popular vote share of the candidate of 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 the error was positive or negative).⁶ For the five elections from 1992 to 2008, we calculated daily forecasts for each of the 100 days prior to

⁶ We report only effect sizes and avoid statistical significance. For an explanation, see Armstrong (2007).

Election Day. Thus, we obtained 500 daily forecasts from polls, models, and the IEM. Our own expert forecasts were available only for the two elections in 2004 and 2008, for a total of 196 daily forecasts.⁷ All data and calculations are publicly available from the IJF website.

4.2.1 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 10%. Gains were higher when combining within models (16%) and expert forecasts (16%). Calculating 7-day averages of IEM prices reduced the error of the original IEM by 6%; this combining procedure yielded more accurate forecasts than the original IEM in each election year except for 1992.

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

	1992	1996	2000	2004	2008	Total
Within component combining						
Model average vs. typical model	6	43	0	5	51	16
Combined experts vs. typical expert	na	na	na	23	10	16
Poll average vs. typical poll	1	3	0	27	36	10
7-day IEM average vs. original IEM	-1	20	11	17	2	6
Across components combining: PollyVote vs.						
Poll average	84	52	39	67	27	60
Model average	77	-39	51	86	-6	52
Experts	na	na	na	69	38	22
IEM (7-day average)	70	-64	-57	19	7	-2
Within and across combining: PollyVote vs.						
Typical individual poll	84	54	38	76	53	64
Typical individual model	78	7	51	87	48	60
Typical individual expert	na	na	na	76	44	34
Original IEM	70	-55	-52	33	9	5
Average of all individual forecasts	39	25	-4	27	3	21

4.2.2 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 60% more accurate than the combined poll average, 52% more accurate than the

⁷ In 2004, the first expert forecast was not available before 96 days prior to Election Day.

combined model average, and 22% more accurate than the combined experts. However, the PollyVote did not improve upon the 7-day IEM forecasts (-2%).

4.2.3 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 poll (64%) and model (60%) and the typical expert (34%). Compared to the original IEM, the PollyVote reduced the error by 5% on average. In addition, the average error of the PollyVote was 21% lower than the simple average of all individual forecasts (i.e., without combining within component methods first). The results demonstrate the benefit of combining within and across components.

The error reduction of 5% compared to the original IEM is rather small and one might question whether the gains from combining are worth the cost. The hit rate provides additional insight on the relative accuracy of both methods. The hit rate is an important criterion for assessing the accuracy of election forecasts, as it measures the frequency with which a forecast correctly predicted election winner. As shown in Table 2, the PollyVote predicted the correct election winner on 96% of all days, compared to a hit rate of 81% for the IEM.

Table 2: Hit rate (in %) of the PollyVote and the original IEM forecasts

	1992	1996	2000	2004	2008	Total
PollyVote	100	100	79	100	100	96
Original IEM	65	97	55	90	100	81

4.2.4 Accuracy gains for different combinations of component methods

Table 3 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 for the 196 daily forecasts from the two elections in 2004 and 2008, for which forecasts from all four component methods were available. As expected, a higher incidence of bracketing yielded larger error reductions.

4.2.4.1 Combinations of two component methods

On average, combining across two methods led to a 22% error reduction relative to the typical forecast. Combinations that included the model forecasts yielded the largest gains in accuracy, especially when this method was paired with the experts (error reduction: 30%). The

probabilities of bracketing (25%) and MER (11%) were smallest for the combination of IEM and polls, which suggests that the IEM incorporates much information from the polls. Gains in accuracy were also small when combining experts and the IEM. A possible explanation for this might be the similarity of both methods as they allow for incorporating all available information. Another explanation is that the experts may have consulted the IEM throughout the campaign, including at the time when making their forecasts.

4.2.4.2 Combinations of three component methods

On average, the combinations of three components led to error reductions of 36% relative to the typical forecast. Again, the error reductions were largest (up to 44%) if the combined forecast included information from the models. The error reductions were smallest – although still at the substantial level of 23% – for the combination of polls, experts, and the IEM.

4.2.4.3 Combinations of four component methods

The combination of four methods led to an error reduction of 46% relative to the typical forecast. In two out of three cases, combining the forecasts from all four component methods yielded bracketing.

Table 3: 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 & IEM	52	23
Models & polls	50	27
Models & experts	49	30
Polls & experts	36	22
IEM & experts	29	17
Polls & IEM	25	11
Mean	40	22
Three component methods		
Models & polls & experts	67	44
Models & IEM & experts	65	43
Models & polls & IEM	63	33
Polls & IEM & experts	44	23
Mean	60	36
Four component methods	69	46

4.2.5 Benefits of combining forecasts under uncertainty

There are many reasons for uncertainty in forecasting, such as high disagreement among forecasts, long lead times, or the challenge of identifying an accurate forecast. In the following discussion, we analyze the benefits of combining under these conditions.

4.2.5.1 Uncertainty due to disagreement among forecasts

If forecasts derived from different methods agree, certainty about the situation increases. Vice versa, high disagreement among forecasts indicates high uncertainty. Disagreement among forecasts is often used as an *ex ante* measure for uncertainty. For example, 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 within a given method tends to underestimate the level of uncertainty.

Table 4 shows 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 500 daily forecasts from the five elections in our data set. Low (high) uncertainty refers to the forecasts in the lower (upper) quartile. Medium uncertainty refers to the forecasts in the interquartile range.

Table 4: Mean error reduction (in %) of combining under uncertainty

Uncertainty	Range	MER
Low	<= 3.1	21
Medium	3.1 to 6.7	40
High	> 6.7	67

As expected, the MER of the PollyVote compared to the typical component increased as uncertainty increased. In situations with low uncertainty, the PollyVote reduced the error of the typical component forecast by 21%. In situations with high uncertainty, combining yielded a 67% error reduction. In sum, the benefits from combining were larger when disagreement among component forecasts, and thus the chance of bracketing, was higher.

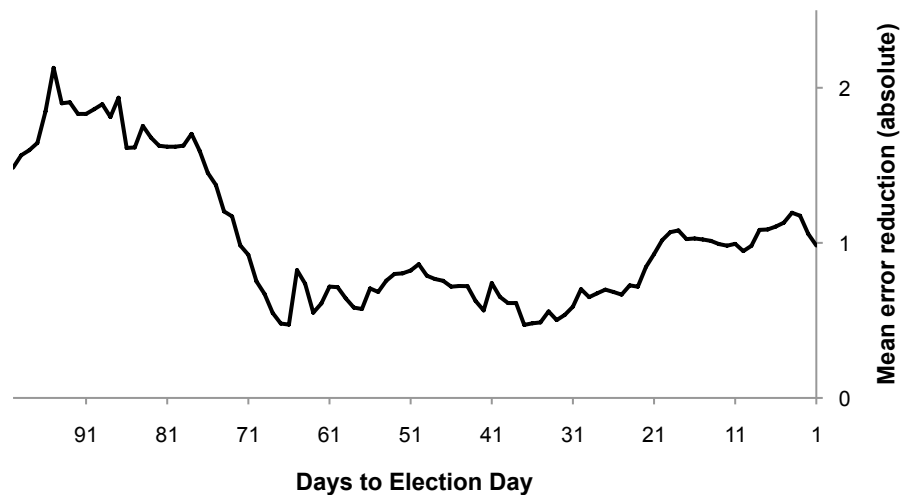
4.2.5.2 Uncertainty due to long time horizons

Uncertainty usually increases with the time horizon of the forecast. Accordingly, combining should be more helpful early in a campaign. Figure 1 shows the MER, calculated across all five elections, of the combined PollyVote forecast compared to the forecast of the typical component for

the last 100 days prior to Election Day.

As expected, the gains from combining are particularly high early in the campaign, with a mean absolute error reduction of up to 2 percentage points. Subsequently, the gains from combining decrease until around one month prior to Election Day, when the mean absolute error reduction is about half of a percentage point. Interestingly, from there the gains from combining increase again, with an average error reduction of around 1 percentage point during the last three weeks before Election Day. An explanation might be the presidential debates, which are held during that period and introduce much uncertainty regarding the evaluations of candidates, which is reflected in volatile polls and IEM forecasts.

Figure 1: Mean error reduction of the PollyVote forecast compared to the forecast of the typical component over time



4.3 Summary of results

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 34% to 64%. Compared to the original IEM, essentially an approach that aggregates and combines dispersed information, the PollyVote reduced error by 5%.

While combining is useful under all conditions, it is especially valuable in situations involving high uncertainty. In situations with high disagreement among the component forecasts,

the PollyVote reduced the error of the typical component on average by 67%.

These gains in accuracy were achieved by using equal weights for combining the forecasts. Equal weights seemed to be an appropriate and pragmatic choice as there is a lack of prior knowledge on how to weight the methods as well as insufficient data to analyze the effects of differential weights. In addition, equal weights are simple to use and easy to understand. That said, further improvements might be possible if additional knowledge is gained about the relative performance of the different methods and their historical track record under certain conditions, such as their accuracy during different point in times in an election cycle.

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 one of the most effective and reliable ways to improve forecast accuracy and to prevent large errors. We expect that the more methods that are combined, the more accurate the forecasts. Of course this would occur at a diminishing rate, so there is a point at which costs exceed benefits.

5. Barriers to combining

Over the past half-century, practicing forecasters have advised firms to use combining. For example, the National Industrial Conference Board (1963) and Wolfe (1966) recommended combined forecasts. PoKempner and Bailey (1970) concluded that combining was a common practice among business forecasters. Dalrymple's (1987) survey on the use of combining for sales forecasting revealed that, of the 134 U.S. companies responding, 20% "usually combined", 19% "frequently combined," 29%, "sometimes combined," and 32% "never combined". We suspect however, that the respondents are referring to informal method of combining, such as weighting individual forecasts based on unaided judgment. Such an approach to combining does not conform to the definition in this paper. There are a number of possible explanations for the low usage of formal 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).

Combining seems 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 difficulty accepting 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 did he find 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.

6. Conclusion

Although combining forecasts has long been known, it is seldom used because it is counter-intuitive and because people are unaware of the substantial benefits. 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, which combined a number of valid forecasting methods that use different data sources. This approach 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 6% to 16%. Compared to the typical uncombined individual forecast, the combining procedure yielded mean error reductions ranging from 5% to 64%.

7. 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 *annual meeting of the American Political Science Association*. 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 we thank all who replied. Kelsey Matevish helped to edit the paper.

8. References

- Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model, *PS: Political Science and Politics*, 41, 691–69.
- Anderson, L. R. & Holt, C. A. (1997). Information cascades in the laboratory, *American Economic Review*, 87, 847-862.
- Armstrong, J. S. (2007). Significance tests harm progress in forecasting, *International Journal of Forecasting*, 23, 321-327.
- 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.
- Bates, J. M. & Granger, C. W. J. (1969). The combination of forecasts, *Operational Research Quarterly*, 20, 451-468.
- Berg, J. E., Nelson, F. D. & Rietz, T. A. (2008). Prediction market accuracy in the long run, *International Journal of Forecasting*, 24, 285-300.
- Campbell, J. E. (2008). The trial-heat forecast of the 2008 presidential vote: Performance and value considerations in an open-seat election, *PS: Political Science and Politics*, 41, 697–701.
- Campbell, J. E. & Garand, J. C. (2000). *Before the Vote. Forecasting American National*

Elections, Thousand Oaks, CA: Sage Publications.

Campbell, J. E. & Wink, K. A. (1990). Trial-heat forecasts of the popular vote, *American Politics Quarterly*, 18, 251-269.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting*, 5, 559-583.

Cuzán, A. G. & Bundrick, C. M. (2008). Forecasting the 2008 presidential election: A challenge for the fiscal model, *PS: Political Science and Politics*, 41, 717-722.

Dalrymple, D. J. (1987). Sales forecasting practices: Results from a United States survey, *International Journal of Forecasting*, 3, 379-391.

DeSart, J. A. & Holbrook, T. M. (2003). Statewide trial-heat polls and the 2000 presidential election: A forecast model, *Social Science Quarterly*, 84, 561-573.

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

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

Erikson, R. S. & Wlezien, C. (2008b). Leading economic indicators, the polls, and the presidential vote, *PS: Political Science and Politics*, 41, 703-707.

Fair, R. C. (2009). Presidential and congressional vote-share equations, *American Journal of Political Science*, 53, 55-72.

Galton, F. (1879). Composite portraits, made by combining those of many different persons into a single resultant figure. *Journal of the Anthropological Institute of Great Britain and Ireland*, 8, 132-144.

Gott, J. R. & Colley, W. N. (2008). Median statistics in polling. *Mathematical and Computer Modelling*, 48, 1396-1408.

Graefe, A. & Armstrong, J. S. (2011). Forecasting elections from voters' perceptions of candidates' ability to handle issues, *Under review*, Paper presented at the *Bucharest Dialogues on Expert Knowledge, Prediction, Forecasting: A Social Science Perspective*, Bucharest 2010. Available at: <https://dl.dropbox.com/u/3662406/Articles/PollyIssues.pdf>

Haynes, S. & Stone, J. A. (2008). A disaggregate approach to economic models of voting in U.S.

presidential elections: forecasts of the 2008 election, *Economics Bulletin*, 4, 1-11.

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

Hibbs, D. A. (2000). Bread and peace voting in U.S. presidential elections, *Public Choice*, 104, 149–180.

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

Hillygus, D. S. (2011). The evolution of election polls in the United States, *Public Opinion Quarterly*, 75, 962-981.

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.

Holbrook, T. M. (2008). Incumbency, national conditions, and the 2008 presidential election, *PS: Political Science and Politics*, 41, 709-712.

Jerôme, V. & Jérôme, B. (2011). Forecasting the 2012 US presidential election: What can we learn from a state level political economy model? Paper presented at the *Annual Meeting of the American Political Science Association*, Seattle, 2011. Available at: <http://ssrn.com/abstract=1902853>

Jones, R. J. (2002). *Who Will Be in the White House? Predicting Presidential Elections*, New York: Longman Publishers.

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.

Kernell, S. (2000). Life before polls: Ohio politicians predict the 1828 presidential vote, *PS: Political Science and Politics*, 33, 569-574.

Klarner, C. (2008). Forecasting the 2008 U.S. House, Senate and Presidential Elections at the district and state level, *PS: Political Science and Politics*, 41, 723-728.

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.

Levins, R. (1966). The strategy of model building in population biology, *American Scientist*, 54, 421-431.

Lewis-Beck, M. S. & Rice, T. W. (1992). *Forecasting Elections*, Washington, DC: Congressional Quarterly Press.

Lewis-Beck, M. S. & Tien, C. (2008). The job of president and the jobs model forecast: Obama for '08? *PS: Political Science and Politics*, 41, 687-690.

Lichtman, A. J. (2008). The keys to the white house: An index forecast for 2008, *International Journal of Forecasting*, 24, 301-309.

Lockerbie, B. (2008). Election forecasting: The future of the presidency and the house, *PS: Political Science and Politics*, 41, 713-716.

Montgomery, J. M., Hollenbach, F. & Ward, M. D. (2011). *Improving predictions using ensemble Bayesian model averaging*, Working paper. Available at: <http://montgomery.wustl.edu/Papers/EBMASingleSpaced.pdf>

National Industrial Conference Board (1963). *Forecasting Sales*. Studies in Business Policy, No. 106. New York.

Norpoth, H. (2008). On the razor's edge: The forecast of the primary model, *PS: Political Science and Politics*, 41, 683-686.

PoKempner, S. J. & E. Bailey (1970). *Sales Forecasting Practices*. New York: The Conference Board.

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

Sigelman, L. (1994). Predicting the 1992 election, *Political Methodologist*, 5 (2), 14-15.

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.

Wlezien, C. (2003). Presidential Elections Polls in 2000: A Study in Dynamics. *Presidential Studies Quarterly*, 33, 172-186.

Wolfe, H. D. (1966). *Business Forecasting Methods*. New York: Holt, Rinehart and Winston.

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

Zajonc, R.B. (1962). A note on group judgments and group size, *Human Relations*, 15, 177-180.