International Journal of Forecasting 30 (2014) 43–54

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

# Combining forecasts: An application to elections

Andreas Graefe [a,*], J. Scott Armstrong [b,c], Randall J. Jones Jr. [d], Alfred G. Cuzán [e]

[a] LMU Munich - Department of Communication Science and Media Research, Munich, Germany
[b] University of Pennsylvania, Wharton School, Philadelphia, PA, United States
[c] Ehrenberg-Bass Institute, University of South Australia, Adelaide, Australia
[d] University of Central Oklahoma - Department of Political Science, Edmond, OK, United States
[e] University of West Florida - Department of Government, Pensacola, FL, United States

ARTICLE INFO

*Keywords:*
Election forecasting
Combining
Prediction markets
Polls
Econometric models
Expert judgment

ABSTRACT

We summarize the literature on the effectiveness of combining forecasts by assessing the conditions under which combining is most valuable. Using data on the six US presidential elections from 1992 to 2012, we report the reductions in error obtained by averaging forecasts within and across four election forecasting methods: poll projections, expert judgment, quantitative models, and the Iowa Electronic Markets. Across the six elections, the resulting combined forecasts were more accurate than any individual component method, on average. The gains in accuracy from combining increased with the numbers of forecasts used, especially when these forecasts were based on different methods and different data, and in situations involving high levels of uncertainty. Such combining yielded error reductions of between 16% and 59%, compared to the average errors of the individual forecasts. This improvement is substantially greater than the 12% reduction in error that had been reported previously for combining forecasts.


## 1. Introduction

Combining has a rich history, and 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 by Clemen, 1989). In using photographic equipment to combine portraits of people, Galton (1879, p. 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". In 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 which, among them, use all of the information available, and then average them. Zajonc (1962) summarized the related literature in psychology, which dates from the early 1900s. Note that these early applications of combining all related to estimation problems, rather than forecasting.

In more recent years, researchers have adopted combining as a simple and useful approach to reducing forecast error. Armstrong (2001) reviewed the literature in order to provide an assessment of the gains in accuracy that can be achieved by combining two or more numerical forecasts. Across thirty studies, the average forecast had 12% less error than the typical component forecast. In addition, the combined forecasts were often more accurate than even the most accurate component forecast.

One intuitive explanation as to why combining improves the accuracy is that it enables forecasters to use more information, and to do so in an objective manner. Moreover, bias exists both in the selection of data and in the forecasting methods that are used. Often the bias is unique to the data and the method, so that when various methods using different data are combined to make a forecast, the biases tend to cancel out in the aggregate.

The research interest in combining forecasts has increased since the publication of the frequently-cited

* Corresponding author.
*E-mail address:* a.graefe@lmu.de (A. Graefe).

0169-2070/$ – see front matter 
http://dx.doi.org/10.1016/j.ijforecast.2013.02.005

paper by Bates and Granger (1969). Numerous studies have demonstrated the value of combining and have tested many alternative proposed methods of weighting the components (for example, based on their historical accuracy), rather than using simple equal-weight averages. However, in an early review of more than two hundred published papers, Clemen (1989) concluded that using equal weights provides a benchmark that is difficult for more sophisticated approaches to beat.

In 2004, we started the www.PollyVote.com project, to test the benefits of combining forecasts of US presidential elections. Forecasts of election outcomes, produced by the following methods, were all collected and processed: polls, prediction markets, experts' judgment, and quantitative models. We expected large gains in forecast accuracy, since the availability of forecasts from such diverse methods and data sets provided ideal conditions for combining (Armstrong, 2001). We had no strong prior evidence as to the relative performances of the various methods, so we decided to combine the forecasts using equal weights. This approach provided additional benefits, including simplicity of calculation and the resulting potential appeal to a broad audience.

In the following sections, we briefly discuss why and how combining works, and outline the conditions under which it is most useful. We then report the results from our combination forecasts for six US presidential elections, three of which were predicted *ex ante*. The results reveal that combining forecasts under ideal conditions yields large gains in accuracy, much larger than those previously estimated by Armstrong (2001).

## 2. Why combining reduces forecast error

In this section, we explain the terms used to describe the mechanism of combining that is employed in this study, namely to calculate simple averages of forecasts.

### 2.1. A note on terms: typical error, combined error, bracketing

The error that is derived by averaging the absolute deviation of a set of $N$ numerical forecasts $F_i$ from the actual value $A$ is termed the "typical error":

$$\frac{\sum_{i=1}^{N} |F_i - A|}{N}.$$

Thus, the typical error is the error that one could expect from a random selection of an individual forecast from a given set of forecasts. In mathematical terms, it is similar to the expected value.

By comparison, the "combined error" is the error that is determined by first averaging the $N$ forecasts $F_i$, and then comparing that average with the outcome $A$:

$$\left| \frac{\sum_{i=1}^{N} F_i}{N} - A \right|.$$

When one forecast is higher than the actual score that was predicted, and one is lower, "bracketing" occurs (Larrick & Soll, 2006). That is, the value to be predicted lies within the range of a set of forecasts. In this situation, the combined error will invariably be lower than the typical error. When bracketing does not exist, the typical error and the combined error will be of the same magnitude. In that case, combining will not improve the accuracy, but neither will it diminish it.

### 2.2. An example from the 2012 election

In the 2012 election, President Obama won 52.0% of the two-party popular vote. Several months before the election, Abramowitz's (2012) "time for change" model predicted that Obama would receive 50.6% of the two-party vote for president, which was 1.4 percentage points lower than the actual result. Near the same time, Klarner's (2012) model predicted that Obama would garner 51.3% of the vote, which was 0.7 percentage points too low. Since both models under-predicted the outcome, no bracketing occurred, and hence, the typical error was equal to the combined error: 1.1 percentage points. That is, combining did as well as randomly picking one of the forecasts. In addition, combining did avoid the risk of picking the forecast model that incurred the largest error. However, it also prevented one from picking the most accurate forecast.[1]

Now, consider a situation in which two forecasts lie one on either side of the true value, bracketing it. The 2012 forecast of the Erikson and Wlezien model (2012a) was 52.6%. Thus, the typical error of the two models of Abramowitz and Erikson & Wlezien was 1.0 percentage points. However, the average of the two forecasts (51.6%) missed the true value by only 0.4 percentage points. In this situation, combining the forecasts of the two models reduced the error of the typical individual model by 60%. In addition, the combined forecast was more accurate than either of the individual forecasts.

## 3. Conditions in which combining is most useful

Combining is applicable to many estimation and forecasting problems. The only exception is when strong prior evidence exists that one method is best *and* the likelihood of bracketing is very low. Armstrong (2001) proposed *ex ante* conditions under which the gains in accuracy that result 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

Accuracy gains that result from combining are most likely to occur when forecasts from many evidence-based

[1] In most real-world forecasting situations, however, it is difficult to identify the most accurate forecast from among a set of forecasts (see Section 3.3).

methods are combined. By "evidence-based" forecasts, we mean forecasts that are generated using methods that adhere to accepted forecasting procedures for the given situation. (A useful tool in making this assessment is the Forecasting Audit at forprin.com.)

When combining, Armstrong (2001) recommended using at least five forecasts. Adding more forecasts may improve the accuracy, 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 varying numbers of estimates. The size of the error shrank as more estimates were included in the combination, although, again, at a diminishing rate. Jose and Winkler (2008) provided similar results for combinations of five, seven, and nine forecasts.

### 3.2. Use of forecasts that draw upon different methods and data

Combining forecasts is most valuable when the individual forecasts are diverse in the methods used and the theories and data upon which they are based. The reason for this is that such a set of forecasts is likely to include different biases and random errors, and, thus, should lead to bracketing and low error correlations.

Batchelor and Dua (1995) analyzed combinations of 22 US 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 larger the differences in the underlying theories or methods of the component forecasts, the greater the extent and probability of error reduction through combining. For example, when combining the real GNP forecasts of two forecasters, combining the five percent of forecasts that were most similar in their underlying theory reduced the error of the typical forecast by 11%. By comparison, combining the five percent 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. The 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 Clemen (2004) reached a similar conclusion. In their laboratory experiment, they asked each participant to use six different strategies to generate six different solutions to an estimation task. Then, the authors analyzed the relative accuracies of different combining approaches. The results showed that combining estimates across participants was generally more accurate than combining different estimates by the same participant. On average, combining a single estimate each from two participants was more accurate than combining four estimates from the same participant.

### 3.3. Uncertainty about the best forecast

Rather than combining forecasts, some analysts argue that it is better to simply pick the most accurate forecast. This objection seems to be of little practical relevance. Although a method's past performance may be an indication of its future performance, there is no assurance that a method will continue to be as accurate in the future as it has been in the past.[2] Under such uncertainty, there is little likelihood that one can identify the most accurate method for future forecasts.

A study, conducted to examine the strategies people use to make decisions based upon two sources of advice, provided experimental evidence: instead of combining the advice, the majority of participants tried to identify the most accurate source — and thereby reduced the accuracy (Soll & Larrick, 2009). In most real-world forecasting situations, it cannot be known beforehand that the selected forecast will be the most accurate. As a result, when picking a single forecast, one takes the risk of choosing a poor forecast. The prudent forecaster, therefore, may want to minimize this risk by combining, even though a particular forecast could eventually prove to be more accurate than the combination.

Research by Hibon and Evgeniou (2005) supports this approach. The authors compared the relative risks associated with two strategies for predicting the 3003 time series used in the M3-competition based on forecasts from fourteen methods: choosing an individual forecast or relying on various combinations of forecasts. Risk was measured as the incremental error that resulted from failing to identify the best individual forecast. When compared to picking an individual forecast at random, choosing a random combination of all possible combination forecasts reduced the risk by 56%.

Turning to the opposing argument, assume that the forecaster has very good evidence that a given forecast method will be more accurate than others. Even in this situation, combining may still improve the accuracy. Herzog and Hertwig (2009) and Soll and Larrick (2009) illustrate the conditions under which combining is better than picking a single forecast, even when one has complete knowledge as to the particular forecast that will be the most accurate. For example, the mean absolute error of the average of two forecasts is lower than the error of the best individual forecast if two conditions are met: (1) the two forecasts bracket the actual score being predicted, and (2) the absolute error of the less accurate forecast does not exceed three times the absolute error of the most accurate forecast.

## 4. The value of weighting components equally

As was noted previously, Clemen (1989) reviewed the literature on combining forecasts and concluded that equal weighting of the individual forecasts is often the best course of action when combining. More than twenty years later, these results are still valid.

[2] Election forecasting is no exception. Holbrook (2010) analyzed the relative accuracies of nine established econometric models for the elections from 1996 to 2004. He found that the models' accuracies varied considerably both within and across elections, and that no single model was always the most accurate.

In a recent study, Genre, Kenny, Meyler, and Timmermann (2013) analyzed various sophisticated approaches to combining forecasts from the European Central Bank's Survey of Professional Forecasters. Although some of the complex combining methods outperformed the simple averages at times, no approach was consistently more accurate over time, across target variables, and across time horizons. Stock and Watson (2004) arrived at similar results when analyzing the relative performances of several combining procedures for economic forecasts, using a seven-country data set over the period from 1959 to 1999. Sophisticated combination methods, which relied heavily on historical performances for weighting the component forecasts, performed worse than a simple average of all available forecasts.

Stock and Watson coined the term "forecast combination puzzle" when referring to the repeated empirical finding that the simple average often outperforms more complex approaches (p. 428). The authors explained their results as being a consequence of the instability of the individual forecasts, since the performances of the individual forecasts varied widely over time, depending on external effects such as economic shocks or political factors. In other words, a good performance in one year or country did not predict a good performance in another, which limits the value of differential weights (see also Section 3.3).

Smith and Wallis (2009) provided a formal explanation of the forecast combination puzzle, showing that the reason is estimation error. Based on results from a Monte Carlo study of combinations of pairs of forecasts, and a reappraisal of a published study on different combinations of multiple forecasts of US output growth, they found that a simple average of forecasts is expected to be more accurate than estimated optimal weights if (a) the optimal weights are close to equality and (b) large numbers of forecasts are combined. The reason for this is that, in such a situation, each forecast has a small weight, and the simple average provides an efficient trade-off against the error that arises from the estimation of weights.[3]

In summary, a large body of analytical and empirical evidence supports the use of equal weights when combining forecasts. In addition to their accuracy, simple averages have another major benefit: they are easy to describe, understand, and implement.

However, this is not to say that equal weights will always provide the best results. For example, estimated weights might be useful if one faces a limited number of forecasts that differ widely in accuracy, and one has access to a large sample that allows for the estimation of robust weights. In addition, there are useful and accessible alternatives to simple averages that do not require the estimation of weights, such as trimmed and Winsorized means. These measures eliminate the most extreme data points when calculating averages, and thus, can provide more robust estimates than the simple average. Jose and Winkler (2008) analyzed the relative performances of simple averages, trimmed means, and Winsorized means using datasets from the M3 Competition and the Survey of Professional Forecasters of the Federal Reserve Bank of Philadelphia. The authors found that trimmed and Winsorized means were slightly more accurate than the simple average, particularly when there was a high level of variability among the individual forecasts. In general, the research available suggests that the performances of different combination methods depend on the conditions faced by the forecaster. Thus, forecasters may find that tailoring rules for combining to specific forecasting problems may improve accuracy (Collopy & Armstrong, 1992).

Regardless of the selected combining approach, a general rule is to specify the combination procedure prior to analyzing the data, as this ensures objectivity. Without prior specification, the combined forecasts can be manipulated for political purposes or simply to make them fit with what the forecaster desires, an effect that might not even be apparent to the forecaster.

## 5. Evidence from a study of election forecasting

In this section we combine forecasts of the two-party popular vote shares in US presidential elections. There are several valid methods that are commonly used to predict election outcomes. These include polls, experts' judgment, quantitative models, and prediction markets. Each of these methods uses a different approach and draws upon data from different and varied sources. Election forecasts using these methods, therefore, are well suited to an assessment of the value of combining. The analysis includes the six elections from 1992 to 2012.

### 5.1. Combining procedure

Our approach to combining presidential election forecasts was to weight all of the component methods equally. Given the importance of combining across methods, we combined first *within* and then *across* component methods. In other words, we used equal weighting of all forecasts within each component method, then equal weighting across these forecasts from different methods. The rationale behind choosing this procedure was to equalize the impact of each component method, regardless of whether a component included many forecasts or only a few. For example, while only one suitable prediction market was available, there were forecasts from several quantitative models. In such a situation, a simple average of all available forecasts would over-represent models and under-represent prediction markets, which we expected would harm the accuracy of the combined forecast.

We do not suggest that this approach will generate 'optimal' forecasts, nor do we attempt to include all available forecasts. We describe the general procedure that was used, which was guided by the recommended principle of defining the combining procedures *a priori* (Armstrong,

[3] These results conform to a large body of evidence on the use of weights in linear models. These studies found that the relative performance of unit (or equal) weights compared to differential weights increases with smaller samples, larger numbers of predictor variables, and higher correlations among predictor variables (Dawes, 1979; Einhorn & Hogarth, 1975; Graefe & Armstrong, 2011).

2001).[4] We provide full disclosure of our data in the hope that other researchers will build upon our work. All data will be made publicly available on the IJF website.

### 5.1.1. Combining within methods

In the following subsections we describe the four forecasting methods that were used in this analysis, and explain our approach to combining forecasts within each method. Predictions from polls, models, and the Iowa Electronic Markets (IEM) were available for all six elections in our study, 1992–2012. In addition, we also conducted our own expert surveys for the three elections from 2004 to 2012. The results from combining the forecasts from these methods will be presented in Section 5.2.

*5.1.1.1. 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, rather than providing predictions, polls are 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.

Campbell and Wink (1990) analyzed the accuracy of Gallup trial heat polls for the eleven presidential elections from 1948 to 1988. The use of raw polls to forecast presidential elections produced large errors, which were greater, the longer the time until the election. Other research has shown that different polls conducted by reputable survey organizations at about the same time often reveal considerable levels of variation in their results. Errors caused by sampling problems, non-responses, inaccurate measurements, and faulty processing diminish the accuracy of polls, and the quality of surveys more generally (e.g. Erikson & Wlezien, 1999; Wlezien, 2003).

A simple approach to increasing poll accuracy is to combine polls that are conducted by different organizations at nearly the same time. 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 four electoral votes. They also forecasted Obama to win over McCain in 2008 with an error of only two electoral votes. In both elections, the median statistical approach missed the winner in only one state. Simply aggregating polls has also become popular in the news media. Well-known poll aggregators such as realclearpolitics.com and the Huffington Post Pollster update combined polls on an almost daily basis.

A more sophisticated approach to increasing poll accuracy is to calculate "poll projections", as we term them. Poll projections take the historical record of the polls into account when making predictions of the election outcome. For example, assume that the incumbent is leading the polls by 20 points in July. In analyzing historical polls conducted at around the same time, along with the respective election outcomes, one can derive a formula for translating the July polling figures into an estimate of the incumbent's expected final vote share. This is commonly done by regressing the incumbent's share of the vote on his polling results during certain time periods before the election. Prior research has found that such poll projections are much more accurate than treating raw polls as forecasts (Campbell, 1996; Campbell & Wink, 1990; Erikson & Wlezien, 2008).

In the present study, we adopted an approach for combining and damping polls that is similar to that of Erikson and Wlezien (2008). For each of the 100 days prior to a presidential election, starting with 1952, we averaged the incumbent party candidate's two-party support from all polls that were released over the previous seven days. When no polls were released on a given day, the most recent poll average available was used. Then, for each of the 100 days before the election, we regressed the incumbent's actual two-party share of the popular vote on the poll value for that day. This process produced 100 vote equations (and thus poll projections) per election year. These *ex ante* poll projections were calculated using successive updating. That is, when generating poll projections for the 1992 election, only historical data from the elections from 1952 to 1988 were used. When calculating poll projections for the 2012 election, all polls through 2008 were used. Polling data were obtained from the *iPoll databank* of the *Roper Center for Public Opinion Research*.

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

Each expert can be expected to use a different approach and rely on at least somewhat different data sources when generating 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 US presidential elections, which we used to re-calculate 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.[5] In 2000, the average forecast

[4] For the past three elections, from 2004 to 2012, we provided *ex ante* forecasts, which were updated continuously throughout the campaigns and were posted at www.pollyvote.com. In the present study, we report all forecasts as if they were calculated *ex post*. As a result, the combining procedure described here may differ slightly from the actual calculation of *ex ante* forecasts that was performed for these elections. However, for reasons of simplification and consistency, the present manuscript describes an identical approach to combining across all elections. The actual specifications of the PollyVote in each of these years are described in recap pieces for each election, which were published in *Foresight — The International Journal of Applied Forecasting* (Cuzán, Armstrong, & Jones, 2005; Graefe, Armstrong, Cuzán, & Jones, 2009; Graefe, Armstrong, Jones, & Cuzán, 2013).

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

was 72% more accurate than the typical forecast from fifteen experts.[6]

For the three elections from 2004 to 2012, we formed a panel of experts and contacted them periodically to obtain their estimates of the incumbent's share of the two-party popular vote on Election Day. Most of these experts were academic specialists in elections, though a few were analysts at think tanks, commentators in the news media, or former politicos. We deliberately excluded all election forecasters who had developed their own models, because that method was represented as a separate component in our combined forecast (see Section 5.1.1.3.). The number of respondents in each of the three surveys conducted in 2004 ranged from twelve to sixteen. For the four surveys in 2008, the number of respondents ranged from ten to thirteen. For the eleven surveys conducted in 2012, the number of respondents ranged from twelve to sixteen. Our combined expert forecast was the simple average of the forecasts made by the individual experts.[7] Because our panelists did not meet in person, the possibility of biases due to the influences of strong personalities or individual status was eliminated.

*5.1.1.3. Quantitative 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 outcomes of US presidential elections. Most models include between two and five variables, and they typically combine indicators of economic conditions and of public opinion to measure the incumbent's performance. For example, the models of Abramowitz (2012), Campbell (2012), Erikson and Wlezien (2012a), and Lewis-Beck and Tien (2012) all include a variable measuring opinion (presidential approval or support for the incumbent candidate), along with economic data. For descriptions of early election forecasting models (and other methods), see Campbell and Garand (2000), Jones (2002), and Lewis-Beck and Rice (1992). For overviews of the variables used in the most popular models, see Holbrook (2010) and Jones and Cuzán (2008).

Since the 1990s, the forecasts of competing models have been being published regularly at around Labor Day of the election year. For the past five elections, the forecasts of leading models have been being published in *American Politics Research*, 24(4), and *PS: Political Science and Politics*, 34(1), 37(4), 41(4) and 45(4). Most of the models predicting presidential elections have produced forecasts using data available near the end of July in the election year. Models have usually predicted the election winner correctly, albeit with varying levels of accuracy as to the candidates' vote shares. The forecast errors for a single model can vary widely across elections, and the structures of some of the models have changed over time, so it is difficult to identify the most accurate models.

Prior research has demonstrated that combining predictions from election forecasting models improves the forecast accuracy. Bartels and Zaller (2001) used various combinations of structural variables that are included in prominent presidential election models to construct 48 different models. The variables included six indicators of economic performance, a measure of the relative ideological moderation of the candidates, a measure of the length of time for which the incumbent party has held the White House, and a dummy for war years. We re-calculated the typical error of the 48 models for predicting the 2000 election from their data (Bartels & Zaller, 2001, Table 1), finding it to be 3.0 percentage points. In comparison, the error of the combined forecasts of all models was 2.5 percentage points. That is, combining reduced the error of the typical model by 17%. In a response to Bartels and Zaller's study, Erikson, Bafumi, and Wilson (2001) showed that creating models that combine structural variables with public opinion increases the accuracy further. The authors added presidential approval to the 48 models as an additional variable, thus doubling the number of models to 96. The sum of the absolute errors for their averaged models was 32% lower than that for the averaged Bartels and Zaller models.

Montgomery, Hollenbach, and Ward (2012) combined the forecasts from six established econometric models based on their past performances and uniqueness, using an approach called Ensemble Bayesian Model Averaging (EBMA). Across the nine elections from 1976 to 2008, the error of the combined EBMA forecast was 34% lower than that of a typical individual model. However, as was shown by Graefe (2013), the error of the EBMA forecast was 18% higher than the error of the simple average.

In the present study, we used forecasts from six models in 1992, eight in 1996, nine in 2000, ten in 2004, sixteen in 2008, and twenty-two in 2012. As noted, the forecasts for most models were released by late July, with some of these being updated once, or more often, as revised data became available. Whenever changes occurred, we recalculated the model averages. All of the models were developed by academics and either published in academic journals or presented at academic conferences.[8]

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

7 In 2004, we used the Delphi survey method, though from 2008 onward we eliminated the feedback step and the opportunity to modify the initial estimates, since the experts rarely changed their first estimates.

8 Model forecasts by Abramowitz (2012), Campbell (2012), Erikson and Wlezien (2012a), and Fair (2009) were available for all six elections. Forecasts by Holbrook (2012), Lewis-Beck and Tien (2012), Lockerbie (2012), and Norpoth and Bednarczuk (2012) were available for the five elections from 1996 to 2012. Forecasts by Cuzán (2012) were available for the four elections from 2000 to 2012. Forecasts by Hibbs (2012) were available for the three elections from 2004 to 2012. Forecasts by DeSart and Holbrook (2003), Graefe and Armstrong (2012b), Jerôme and Jerôme-Speziari (2012), Klarner (2012), and Lichtman (2008) were available for 2008 and 2012. Forecasts by Lewis-Beck and Rice (1992) and Sigelman (1994) were available for the 1992 election. A forecast by Haynes and Stone (2008) was available for the 2008 election. Forecasts from Armstrong and Graefe (2011), Campbell's (2012) convention bump model, Berry and Bickers (2012), Graefe (2012), Graefe and Armstrong (2012a), Lewis-Beck and Rice's (1992) proxy model, and Nate Silver's FiveThirtyEight.com were available for the 2012 election.

*5.1.1.4. Prediction markets.* Betting on election outcomes has a long history, and has been recognized as a useful means of forecasting election outcomes. Rhode and Strumpf (2004) studied historical markets that existed for the fifteen presidential elections from 1884 through 1940, and concluded that these markets "did a remarkable job forecasting elections in an era before scientific polling" (p. 127).

These markets were the precursors of today's online prediction markets, with the oldest being the *Iowa Electronic Markets* (IEM) established at the University of Iowa in 1988. In this study, we used prices from the IEM vote-share market as predictions of the vote. In comparing forecasts from the IEM with 964 polls for the five presidential elections from 1988 to 2004, Berg, Nelson, and Rietz (2008) determined that the IEM forecasts were closer to the actual election result than polls conducted on the same day 74% of the time. However, Erikson and Wlezien (2008) found *poll projections* to be more accurate than IEM forecasts.

Prediction market forecasts can be affected negatively by unexpected spikes in prices due to information cascades, which occur when people buy or sell shares simply because of the observed actions of other market participants (Anderson & Holt, 1997). We expected that combining market forecasts over a given time period could moderate these short-term disruptions in market prices, and thus, we combined IEM forecasts by calculating the 7-day rolling average of daily prices of the vote-share contract for the incumbent party candidate. The effect on the forecast accuracy of combining IEM prices was determined by comparing the 7-day average to the daily IEM average.

### 5.1.2. Combining across methods

Although some previous research has assessed the value of combining election forecasts within methods (e.g. Montgomery et al., 2012), we are not aware of any prior research that has combined forecasts both *within and across* methods, which is the approach presented here. 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 in techniques and assumptions, in the types of data used, and in data sources.

For each day in the forecast horizon, we calculated a simple average across the combined component forecasts: poll projections, experts, models, and IEM. We refer to this overall combined forecast as the *PollyVote*.[9]

## 5.2. Results

All of the forecasts reported refer to the two-party popular vote share of the candidate of the incumbent party. All analyses are conducted across the last 100 days prior to Election Day. That is, for the six elections from 1992 to 2012, we calculated daily forecasts and the corresponding errors for each of the 100 days prior to Election Day. Thus, we obtained 600 daily forecasts each from polls, models, and the IEM. Our own expert forecasts were only available for the three elections in 2004, 2008, and 2012, with a total of 296 daily forecasts.[10]

### 5.2.1. A note on error measures

We used the absolute error as a measure of the accuracy (that is, the difference between the predicted and actual vote shares, regardless of whether the error was positive or negative). In presenting the gains achieved through combining, we report the 'error reduction' in percentages. By this, we mean the extent to which the combined error is smaller than the typical error of a set of forecasts:

$$\frac{AE_{typical} - AE_{combined}}{AE_{typical}}.$$

For example, the combined error of the 2012 election forecasts from Abramowitz (2012) and Erikson and Wlezien (2012a) was 0.4 percentage points, compared to 1.0 percentage points for the typical error (see Section 2.2). Thus, the error reduction derived through combining was 60%. When analyzing the accuracy across time periods such as days or years, we report the mean error reduction (MER). The MER for a particular election year is determined by averaging the typical and combined errors across the 100-day period before calculating the error reduction. The MER across years is the simple average of the error reduction of each particular year.[11]

### 5.2.2. Accuracy gains from combining within methods

In Table 1, the section labeled "within component combining" shows the MER over the 100-day forecast horizon that is achieved by combining forecasts within a method category. On average across the six elections, combining poll projections yielded the largest error reductions (39%), even though the approach produced less accurate forecasts than individual polls in 2008.[12] The error reductions were also substantial when combining within the remaining methods: models (30%), expert forecasts (12%), and the IEM (10%). Calculating 7-day averages of IEM prices resulted in more accurate forecasts than the original IEM in each election year except for 1992.

### 5.2.3. Accuracy gains from combining across components

The "across component combining" section of Table 1 shows the MER of the PollyVote forecast compared to

[9] PollyVote stands for "many" and "politics". On our website, we playfully adopted a parrot as a mascot, because the method does little beyond repeating and combining what it borrows (or "hears") from others.

[10] In 2004, the first expert forecast was not available until 96 days prior to Election Day.

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

[12] The poor performance of poll projections in 2008 is probably attributable to the economic crisis that hit in mid-September of that year, less than two months before Election Day. Following this event, the gap in the polls increased decisively in favor of Obama, an effect that was detrimental to the accuracy of the damped poll projections. See Campbell (2010) for a discussion of the decisive impact of the economic crisis on the 2008 election outcome.

**Table 1**
Accuracy gains from combining (in %): across 100 days prior to election day.

| | 1992 | 1996 | 2000 | 2004 | 2008 | 2012 | Avg. |
|---|---|---|---|---|---|---|---|
| **Within component combining** | | | | | | | |
| Poll projections vs. typical poll | 71 | 62 | 53 | 52 | −40 | 39 | **39** |
| Model average vs. typical model | 6 | 43 | 0 | 5 | 51 | 75 | **30** |
| Combined experts vs. typical expert | na | na | na | 23 | 10 | 3 | **12** |
| 7-day IEM average vs. original IEM | −1 | 17 | 18 | 21 | 4 | 3 | **10** |
| **Across components combining: PollyVote vs.** | | | | | | | |
| Poll projections | −26 | 30 | −3 | 51 | 49 | 63 | **27** |
| Model average | 44 | 9 | 64 | 86 | −39 | 37 | **34** |
| Experts | na | na | na | 70 | 4 | 72 | **49** |
| IEM (7-day average) | 27 | −32 | −19 | 24 | −30 | 74 | **7** |
| **Within and across combining: PollyVote vs.** | | | | | | | |
| Typical individual poll | 61 | 73 | 52 | 77 | 14 | 77 | **59** |
| Typical individual model | 47 | 48 | 64 | 87 | 20 | 84 | **58** |
| Typical individual expert | na | na | na | 77 | 14 | 73 | **55** |
| Original IEM | 27 | −19 | −2 | 40 | −27 | 75 | **16** |

the error of the combined forecasts of the component methods. Across the six elections, the PollyVote provided more accurate forecasts than any one of its components. On average, the PollyVote forecast was 49% more accurate than the combined experts, 34% more accurate than the combined models, 27% more accurate than the poll projections, and 7% more accurate than the IEM 7-day average.

*5.2.4. Accuracy gains from combining within and across components*

The section of Table 1 labeled "within and across combining" shows the MER of the PollyVote forecast compared to the typical (uncombined) forecasts of each component method. The gains in accuracy compared to the typical individual poll (59%), the typical model (58%), and the typical expert (55%) were large. In each case, combining reduced the error by more than half. Compared to the original IEM, the PollyVote reduced the error by 16% on average, which is higher than Armstrong's (2001) earlier estimate of 12% on the benefits of combining.[13]

*5.2.5. Accuracy gains for different combinations of component methods*

Table 2 shows the percentage of days when bracketing occurred, and the MER compared to the typical component method for each of the three elections from 2004 to 2012.[14] As expected, the percentage of days with bracketing rose with the number of components included in the forecast.

*5.2.5.1. Combinations of two component methods.* On average, combining across two methods led to a 23% reduction in error relative to the typical component forecast. Combinations of IEM and expert forecasts yielded the largest gains in accuracy (error reduction: 29%). On the other hand, the gains from combining models and poll projections were the smallest (17%). One reason for the low rate of bracketing for models and poll projections could be that many models already use information from polls for measuring public opinion. In contrast, models are limited when it comes to incorporating information about the specific context of a particular election; this might be the reason why high rates of bracketing occur when models are being combined with methods that incorporate human judgment, such as expert forecasts or the IEM. The gains in accuracy were also relatively small when combining poll projections with the IEM forecasts. This conforms to the results of Erikson and Wlezien (2012b), who showed that prediction market forecasts mostly follow the polls.

*5.2.5.2. Combinations of three component methods.* On average, the combinations of three components led to error reductions of 37% relative to the typical forecast. The error reductions were the largest if the model forecasts were combined with human judgment from experts and the IEM (48%). The error reductions were smallest — although still at a substantial level of 31% — for the combination of models, polls, and the IEM.

*5.2.5.3. Combinations of four component methods.* The combination of four methods led to an error reduction of 48% relative to the typical forecast. In nearly three-quarters of cases (72%), combining the forecasts from all four component methods produced bracketing.

*5.2.6. Benefits of combining forecasts under uncertainty*

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

[13] The "hit rate" provides additional insights on the relative accuracies of the PollyVote and the IEM. The hit rate refers to the frequency with which forecasts of a given method predict the popular vote winner correctly, expressed as a percentage of all available forecasts of that method. Thus, the hit rate measures a method's capability to answer the question that is probably of most interest to the regular consumer of election forecasts: who will win (rather than what a candidate's share of the vote will be)? Based on the hit rate, the PollyVote outperformed the original IEM in four of the six elections, with two ties. On average, the PollyVote predicted the correct election winner on 97% of all 600 days in the forecast horizon, compared to a hit rate of 80% for the IEM.

[14] The reason for limiting this analysis to three elections is that these were the only elections for which forecasts from all four component methods were available.

**Table 2**
Bracketing and mean error reductions for different combinations of component methods (2004–2012).

| Combinations based on | % of days with bracketing | MER to typical component (in %) |
|---|---|---|
| **Two component methods** | | |
| IEM & experts | 43 | 29 |
| Models & IEM | 60 | 28 |
| Poll projections & experts | 35 | 23 |
| Poll projections & IEM | 41 | 20 |
| Models & experts | 32 | 20 |
| Models & poll projections | 33 | 17 |
| **Mean** | **41** | **23** |
| **Three component methods** | | |
| Models & IEM & experts | 68 | 48 |
| Poll projections & IEM & experts | 59 | 35 |
| Models & poll projections & experts | 50 | 32 |
| Models & poll projections & IEM | 67 | 31 |
| **Mean** | **61** | **37** |
| **Four component methods** | **72** | **48** |

*5.2.6.1. Uncertainty due to disagreement among forecasts.* If the forecasts derived from different methods agree, certainty about the situation usually increases. In contrast, high levels of disagreement among forecasts indicate a greater uncertainty. Disagreement among forecasts is often used as a conservative *ex ante* measure of uncertainty. For example, in analyzing 2787 observations for inflation and 2342 observations for GDP forecasts from the Survey of Professional Forecasters, Lahiri and Sheng (2010) confirmed the evidence from earlier research that disagreement within a given method tends to underestimate the level of uncertainty.

Table 3 shows the MER of the PollyVote compared to the typical component for different levels of uncertainty, calculated across all 600 days in the dataset. Uncertainty was measured as the range between the highest and lowest component forecasts on any given day. For example, a situation in which the lowest component forecast predicts the incumbent to gain 50% of the vote, and the highest component forecast predicts him to gain 52%, would represent a range of two percentage points. As Table 3 shows, the range between the component forecasts was between two and four percentage points on nearly half of all days (285 out of 600). In these situations, the PollyVote reduced the error of the typical forecast by about 43%. In general, the MER of the PollyVote compared to the typical component increased as the uncertainty increased. That is, the benefits from combining were larger when the disagreement among component forecasts, and in effect the chance of bracketing, was higher.

*5.2.6.2. 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. Fig. 1 shows the MER, calculated across all six 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 high early in the campaign, with a mean error reduction of nearly 1.5 percentage points. Subsequently, the gains from combining decrease as the election nears, which suggests that the forecasts from the different components tend to converge as the uncertainty decreases. Interestingly, the gains from combining increase again in the period from one month to two weeks before Election Day, which is around the time when the presidential debates are usually held. It is left for future research to clarify what is going on late in the campaign, for example, whether the results are driven by a particular forecasting component.

**Table 3**
Mean error reduction of the PollyVote compared to the typical component, depending on the range between the highest and lowest component forecasts.

| Range | *N* | ER (in %) |
|---|---|---|
| [0, 1] | 42 | 0 |
| ]1, 2] | 84 | 6 |
| ]2, 3] | 122 | 36 |
| ]3, 4] | 163 | 48 |
| ]4, 5] | 76 | 53 |
| ]5, 6] | 63 | 40 |
| ]6, 7] | 28 | 37 |
| ]7, 8] | 12 | 60 |
| ]8, 9] | 4 | 49 |
| ]9, 10] | 1 | 67 |
| ]10, 11] | 3 | 77 |
| ]11, 12] | 2 | 84 |

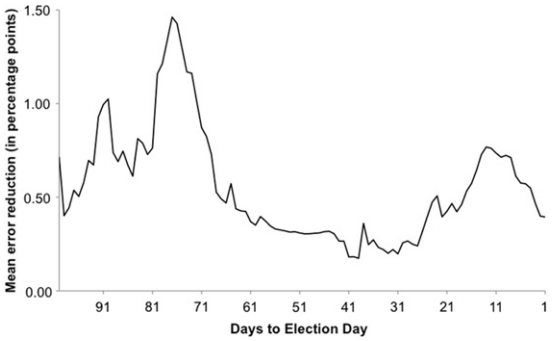


**Fig. 1.** Mean error reduction of the PollyVote forecast compared to the forecast of the typical component over the last 100 days across the six elections, 1992–2012.

## 6. Discussion

In applying a two-step approach by combining forecasts within and across four methods for forecasting US presidential elections, we achieved large gains in accuracy. Compared to forecasts from a randomly chosen poll, model or expert, the PollyVote forecast reduced the error by 55% to 59%. Compared to the original IEM, which is essentially a sophisticated approach for aggregating and combining dispersed information, the PollyVote reduced the error by 16%. Across the six elections, the PollyVote provided forecasts which were more accurate than those of any of its components. While combining is useful under all conditions, it is especially valuable in situations involving high levels of uncertainty.

These gains in accuracy were achieved by combining the forecasts using equal weights. Equal weights seemed to

be an appropriate and pragmatic choice, as there was a lack of prior knowledge on how to weight the methods, as well as insufficient data for analyzing the effects of differential weights. In addition, equal weights are simple to use and easy to understand. That being said, further improvements might be possible if additional knowledge could be gained about the relative performances of the different methods and their historical track records under certain conditions, such as their accuracies at different stages of an election cycle.

Combining should also be applicable to the prediction of other elections, and can be applied more generally in many other contexts as well. Given the various methods available to forecasters, combining is one of the most effective and reliable ways to improve the forecast accuracy and prevent large errors. Of course, the gains in accuracy from adding additional methods accrue at a diminishing rate, so there is a point at which the costs exceed the benefits.

## 7. 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) claimed 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 US companies who responded, 20% "usually combined", 19% "frequently combined", 29% "sometimes combined", and 32% "never combined". We suspect, however, that the survey respondents were referring to informal methods of combining, such as weighting the individual forecasts based on unaided judgment. Such approaches to combining do not conform to the procedures described in this paper.

We believe that combining, as properly defined and implemented, is rarely used today. A number of possible explanations for the low usage of formal combining exist.

1. *A lack of knowledge about the research on combining* is likely to be a major barrier to the use of combining in practice. The benefits of combining are not intuitively obvious, and people are unlikely to learn them through experience. In a series of experiments with MBA students at INSEAD, the majority of participants thought that an average of estimates would result in only an average performance (Larrick & Soll, 2006).

2. *Combining seems too simple.* Hogarth (2012) reported results from four studies showing that simple models often predict complex problems better than more complex models, but that in each case, people had difficulty accepting the findings from simple models. There is a strong belief that complex models are necessary for solving complex problems. Similarly, people might perceive the principle of combining as being "too easy to be true".

3. *Forecasters might seek extreme forecasts 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 the forecasting institutions. 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.

4. *Forecasters may think that 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. In practice, managers often use their unaided judgment to assign differential weights to individual forecasts. Such an *informal* approach to combining is likely to be harmful, as managers may select a forecast that suits their biases.

5. *People mistakenly believe that they can identify the most accurate forecast.* Soll and Larrick (2009) conducted experiments to examine the strategies that 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 actually reduced the accuracy.

One goal of the PollyVote.com project is to help people overcome these barriers by using the high-profile application of forecasting US presidential election outcomes to demonstrate the benefits of combining. Software providers could also contribute by including combining as a default. That is, software solutions should require users to actively opt out of combining if desired, after considering its applicability to the current situation.

## 8. Conclusions

Combining forecasts requires that the procedures be specified and fully disclosed prior to the preparation of the forecasts. This allows for the use of a variety of sources of information in a way that helps to control for bias. In short, combining must be objective.

We have estimated the improvement in accuracy that can be achieved by combining US presidential election forecasts within and across methods. The results are consistent with prior research on combining, but the potential gains are much larger than previously estimated. Under ideal conditions, forecasting errors can be reduced by more than half. Thus, the simple method of combining is one of the most useful procedures in a forecaster's toolkit.

If it is possible to use a number of evidence-based forecasting methods and alternative sources of data, combining forecasts should be considered for all situations that involve uncertainty. Combining forecasts has been shown to be much more useful as the uncertainty increases. For important forecasts, the costs of combining are likely to be trivial relative to the potential gains.

## 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 were summarizing their research accurately, and we thank all who replied. Jennifer Kwok, Kelsey Matevish and Nathan Fleetwood helped to edit the paper.

## References


Abramowitz, A. I. (2012). Forecasting in a polarized era: the time for change model and the 2012 presidential election. *PS: Political Science and Politics*, *45*, 618–619.

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* (pp. 417–439). Norwell: Kluwer.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting*, *23*, 321–327.

Armstrong, J. S., & Graefe, A. (2011). Predicting elections from biographical information about candidates: a test of the index method. *Journal of Business Research*, *64*, 699–706.

Bartels, L. M., & Zaller, J. (2001). Presidential vote models: a recount. *PS: Political Science and 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.

Berry, M. J., & Bickers, K. N. (2012). Forecasting the 2012 presidential election with state-level economic indicators. *PS: Political Science and Politics*, *45*, 669–674.

Campbell, J. E. (1996). Polls and votes: the trial-heat presidential election forecasting model, certainty, and political campaigns. *American Politics Quarterly*, *24*, 408–433.

Campbell, J. E. (2010). The exceptional election of 2008: performance, values, and crisis. *Presidential Studies Quarterly*, *40*, 225–246.

Campbell, J. E. (2012). Forecasting the presidential and congressional elections of 2012: the trial-heat and the seats-in-trouble models. *PS: Political Science and Politics*, *45*, 630–634.

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.

Collopy, F., & Armstrong, J. S. (1992). Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations. *Management Science*, *38*, 1394–1414.

Cuzán, A. G. (2012). Forecasting the 2012 presidential election with the fiscal model. *PS: Political Science and Politics*, *45*, 648–650.

Cuzán, A. G., Armstrong, J. S., & Jones, R. J. (2005). How we computed the PollyVote. *Foresight: The International Journal of Applied Forecasting*, *1*, 51–52.

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

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.

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.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, *13*, 171–192.

Erikson, R. S., Bafumi, J., & Wilson, B. (2001). Was the 2000 presidential election predictable? *PS: Political Science and Politics*, *34*, 815–819.

Erikson, R. S., & Wlezien, C. (1999). Presidential polls as a time series: the case of 1996. *Public Opinion Quarterly*, *63*, 163–177.

Erikson, R. S., & Wlezien, C. (2008). Are political markets really superior to polls as election predictors? *Public Opinion Quarterly*, *72*, 190–215.

Erikson, R. S., & Wlezien, C. (2012a). The objective and subjective economy and the presidential vote. *PS: Political Science and Politics*, *45*, 620–624.

Erikson, R. S., & Wlezien, C. (2012b). Markets vs. polls as election predictors: an historical assessment. *Electoral Studies*, *31*, 532–539.

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.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. (2013). Combining expert forecasts: can anything beat the simple average? *International Journal of Forecasting*, *29*, 108–121.

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

Graefe, A. (2012). Issue and leader voting in US presidential elections. *Electoral Studies*, http://dx.doi.org/10.1016/j.electstud.2013.04.003.

Graefe, A. (2013). *Conditions of ensemble Bayesian model averaging for political forecasting*. Working Paper. Available at http://ssrn.com/abstract=2266307.

Graefe, A., & Armstrong, J. S. (2011). Conditions under which index models are useful: reply to bio-index commentaries. *Journal of Business Research*, *64*, 693–695.

Graefe, A., & Armstrong, J. S. (2012a). Predicting elections from the most important issue: a test of the take-the-best heuristic. *Journal of Behavioral Decision Making*, *25*, 41–48.

Graefe, A., & Armstrong, J. S. (2012b). Forecasting elections from voters' perceptions of candidates' ability to handle issues. *Journal of Behavioral Decision Making*. http://dx.doi.org/10.1002/bdm.1764.

Graefe, A., Armstrong, J. S., Cuzán, A. G., & Jones, R. J. (2009). Combined forecasts of the 2008 election: the Pollyvote. *Foresight: The International Journal of Applied Forecasting*, *12*, 41–42.

Graefe, A., Armstrong, J. S., Jones, R. J., & Cuzán, A. G. (2013). Combined forecasts of the 2012 election: the Pollyvote. *Foresight: The International Journal of Applied Forecasting*, *28*, 50–51.

Haynes, S., & Stone, J. A. (2008). A disaggregate approach to economic models of voting in US 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. (2012). Obama's reelection prospects under "bread and peace" voting in the 2012 US presidential election. *PS: Political Science and Politics*, *45*, 635–639.

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. (2012). When simple is hard to accept. In P. M. Todd, G. Gigerenzer, & The ABC Research Group (Eds.), *Ecological rationality: intelligence in the world* (pp. 61–79). Oxford: Oxford University Press.

Holbrook, T. M. (2010). Forecasting US presidential elections. In J. E. Leighley (Ed.), *The Oxford handbook of American elections and political behavior* (pp. 346–371). Oxford: Oxford University Press.

Holbrook, T. M. (2012). Incumbency, national conditions, and the 2012 presidential election. *PS: Political Science and Politics*, *45*, 640–643.

Jerôme, B., & Jerôme-Speziari, V. (2012). Forecasting the 2012 US presidential election: lessons from a state-by-state political economy model. *PS: Political Science and Politics*, *45*, 663–668.

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 US presidential elections: a brief review. *Foresight: The International Journal of Applied Forecasting*, *10*, 29–34.

Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: some empirical results. *International Journal of Forecasting*, *24*, 163–169.

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

Klarner, C. (2012). State-level forecasts of the 2012 Federal and Gubernatorial elections. *PS: Political Science and Politics*, *45*, 655–662.

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. (2012). Election forecasting for turbulent times. *PS: Political Science and Politics*, *45*, 625–629.

Lichtman, A. J. (2008). The keys to the White House: an index forecast for 2008. *International Journal of Forecasting*, *24*, 301–309.

Lockerbie, B. (2012). Economic expectations and election outcomes: the presidency and the house in 2012. *PS: Political Science and Politics*, *45*, 644–647.

Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving predictions using ensemble Bayesian model averaging. *Political Analysis*, *20*, 271–291.

National Industrial Conference Board (1963). *Forecasting sales*. Studies in business policy, No. 106. New York.

Norpoth, H., & Bednarczuk, M. (2012). History and primary: the Obama reelection. *PS: Political Science and Politics*, *45*, 614–617.

PoKempner, S. J., & Bailey, E. (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*, 14–15.
Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*, 331–355.
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.
Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430.
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.
Zajonc, R. B. (1962). A note on group judgments and group size. *Human Relations*, *15*, 177–180.

**Andreas Graefe** is currently a research fellow at LMU München, Germany. Before that, he worked as Senior Manager Resource Management, Forecasting & Planning at the pay TV company Sky Deutschland. He has done validation work on several forecasting methods such as the index method and prediction markets. He has published a number of research articles in journals including the *International Journal of Forecasting*, the *Journal of Behavioral Decision Making*, and the *Journal of Business Research*. He also serves as Prediction Market Editor for *Foresight* and Associate Editor of Research Methods in Business for the *Journal of Business Research*.

**J. Scott Armstrong** is Professor of Marketing at the Wharton School, University of Pennsylvania. He is a founder of the *Journal of Forecasting*, the *International Journal of Forecasting*, and the International Symposium on Forecasting. He is the creator of forecastingprinciples.com and editor of *Principles of Forecasting* (Kluwer, 2001), an evidence-based summary of knowledge on forecasting. In 1996, he was selected as one of the first "Honorary Fellows" by the International Institute of Forecasters. One of Wharton's most prolific scholars, his current research involves the application of scientific forecasting methods to climate change, the effectiveness of learning at universities, and the use of the index method to make predictions in situations with many variables and much knowledge. His book, *Persuasive Advertising* (2010), summarizes the evidence-based knowledge on persuasion.

**Randall J. Jones, Jr.** is Professor of Political Science at the University of Central Oklahoma, where his work includes teaching election forecasting and conducting research in that field. He is author of the textbook, *Who Will Be in the White House? Predicting Presidential Elections* (Longman, 2002), and is a founder and former officer of the Political Forecasting Group of the American Political Science Association.

**Alfred G. Cuzán** is Professor of Political Science at the University of West Florida. He is the author of many papers on American presidential elections and Latin American politics. In collaboration with Richard J. Heggen and Charles M. Bundrick, he developed the fiscal model of presidential elections. Also, in 2004, he joined J. Scott Armstrong and Randall J. Jones, Jr., to design the first application of the combination principle to the prediction of of a presidential election outcome.