Combining Forecasts: An Application to Elections

Andreas Graefe
Department of Communication Science and Media Research
LMU Munich, Germany
a.graefe@lmu.de

J. Scott Armstrong
Wharton School
University of Pennsylvania, Philadelphia, PA, USA
armstrong@wharton.upenn.edu

Randall J. Jones
Department of Political Science
University of Central Oklahoma, Edmond, OK, USA
ranjones@uco.edu

Alfred G. Cuzan
Department of Government
University of West Florida, Pensacola, FL, USA
acuzan@uwf.edu

January 28th, 2013

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 U.S. Presidential elections from 1992 through 2012, we then report the reduction 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 on average more accurate than each of the component methods. The gains in accuracy from combining increased with the number of forecasts used, especially when these forecasts were based on different methods and different data, and in situations involving high uncertainty. Combining yielded error reductions ranging from 16% to 59%, compared to the average errors of the individual forecasts. This improvement is substantially greater than the 12% reduction in error that had been previously reported for combining forecasts.

1. Introduction

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." 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 that, among them, use all the information available, and then average them. Zajonc (1962) summarizes related literature in psychology, which dates from the early 1900s. Note that these early applications of combining related to estimation problems, rather than forecasting.

In more recent years, researchers have adopted combining as a simple and useful approach to reduce forecast error. Armstrong (2001) reviewed the literature 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 forecasts. In addition, the combined forecasts were often more accurate than the most accurate component forecast.

One intuitive explanation as to why combining improves accuracy is that it enables 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 cancel out in the aggregate.

Research interest in combining forecasts has increased since publication of a frequently-cited paper by Bates and Granger (1969). Numerous studies have demonstrated the value of combining and tested many 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 to beat by more sophisticated approaches.

In 2004, we started the PollyVote.com project to test the benefits of combining forecasts of U.S. presidential elections. Forecasts for predicting election outcomes, produced by the following methods, were collected and processed: polls, prediction markets, experts' judgment, and quantitative models. We expected large gains in forecast accuracy, since forecasts using such diverse methods and data provided ideal conditions for combining (Armstrong, 2001). We had no strong prior evidence as to the relative performance of each method. For this reason, 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 results from combining forecasts for six U.S. 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 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 was employed in this study, which was 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 deviations of a set of N numerical forecasts F_i from the actual value A is termed the "typical error":

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

The typical error is thus the error that one can expect by randomly selecting 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 accuracy, but it will not diminish accuracy either.[1]

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

[1] Note that the benefits of combining are limited to numerical forecasts and do not apply to categorical data. The reason is that for categorical forecasts, bracketing is not possible. In such cases, combining can harm forecast accuracy (see Armstrong et al., 2013).

before the election Abramowitz's "time for change" model (2012) predicted that Obama would receive 50.6% of the two-party vote for president, which was 1.4 percentage points lower than the result. Near the same time, the model by Klarner (2012) predicted that Obama would garner 51.2% of the vote, which was 0.8 percentage points too low. Since both models under-predicted the outcome, no bracketing occurred; 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, combining also prevented one from picking the most accurate forecast.[2]

Now, consider a situation in which two forecasts lie on either side of the true value, bracketing it. The 2012 forecast of the Erikson & Wlezien model (2012a) was 52.6%. Thus, the typical error of the two models by 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 both models reduced the error of the typical individual model by 60%. In addition, the combined forecast was more accurate than each of the individual forecasts.

3. Conditions when 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 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 accuracy, though at a diminishing rate of improvement. Nine of the thirty 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

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

methods. Vul and Pashler (2008) plotted the errors for combinations of a varying number 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 methods used and in the theories and data upon which they are based. The reason 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.

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 larger the differences in the underlying theories or methods of the component forecasts, the higher the extent and probability of error reduction through combining. For example, when combining 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. 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 for generating six different solutions to an estimation task. Then, the authors analyzed the relative accuracy 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 from two participants was more accurate than combining four estimates from the same participant.

3.3. Uncertainty about the best forecast

Rather than combine 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 the method will

continue to be as accurate as in the past.[3] Under such uncertainty, there is little likelihood that one will be able to determine which method will be most accurate in the future.

A study that was 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 accuracy (Soll & Larrick, 2009). In most real-world forecasting situations, there is no assurance 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 risk associated with two strategies for predicting the 3,003 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 randomly picking an individual forecast, choosing a random combination of all possible combination forecasts reduced risk by 56%.

Turning to the opposing argument, assume that the forecaster does have very good evidence that a given forecast method will be the most accurate. Even in this situation, combining, nevertheless, may improve accuracy. Herzog and Hertwig (2009) and Soll and Larrick (2009) illustrate when combining is better than picking a single forecast, even when one has complete knowledge about which individual forecast is the most accurate. For example, the average of two forecasts is more accurate than 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 noted previously, Clemen (1989) reviewed the literature on combining forecasts and concluded that equally weighting the individual forecasts is often the best course of action when combining. More than twenty years later, these results still remain valid.

In a recent study Genre *et al.* (2013) analyzed various sophisticated approaches to combining forecasts from the European Central Bank's Survey of Professional Forecasters. Although at times some of

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

the complex combining methods outperformed the simple averages, 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 performance of several combining procedures for economic forecasts, using a seven-country data set over the time period from 1959 to 1999. Sophisticated combination methods, which relied heavily on historical performance for weighing the component forecasts, performed worse than a simple average of all available forecasts.

Stock and Watson have 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 a consequence of the instability of individual forecasts, since the performance of individual forecasts varied widely over time, depending on external effects such as economic shocks or political factors. In other words, good performance in one year or country did not predict good performance in another, which limits the value of differential weights (see also Section 3.3).

Smith and Wallis (2009) provided a formal explanation for the forecast combination puzzle, showing that the reason is estimation error. Based on results from a Monte Carlo study of combinations of two 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 if (b) a large number of forecasts are combined. The reason 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.[4]

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.

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 can rely on a large sample that allows for estimating robust weights. In addition, there are useful and accessible alternatives to simple averages that do not require estimating 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 performance of simple averages, trimmed, and Winsorized means for using datasets from the M3 Competition and the Survey of Professional Forecasters of the Federal Reserve Bank of Philadelphia. The

[4] These results conform to a large body of evidence on the use of weights in linear models. These studies found the relative performance of unit (or equal) weights compared to differential weights increases with small samples, a large number of predictor variables, and high correlation among predictor variables (Dawes, 1979; Einhorn & Hogarth, 1974; Graefe & Armstrong, 2011).

authors found that trimmed and Winsorized means were slightly more accurate than the simple average, in particular when there was large variability among the individual forecasts. In general, the research available suggests that the performance of different combination methods depends on the conditions faced by the forecaster. Forecasters might want to use different rules for combining, depending on the conditions of the forecasting problem (Collopy & Armstrong, 1992).

Regardless of the selected combining approach, a general rule is to specify the procedure for how to combine 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 might desire, 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 U.S. presidential elections. Several valid methods 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 for assessing 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 component methods equally. Given the importance of combining across methods, we first combined *within* and then *across* component methods. In other words, we used equal weighting of all forecasts within each component method, then equal weighting across 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 that used a similar method and similar information. 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 to define the combining procedures *a priori* (Armstrong, 2001).[5] We provide full

[5] For the past three elections in from 2004 and to 2012, we provided *ex ante* forecasts, which were continuously updated throughout the campaigns

disclosure of our data in the hope that other researchers will build upon our work. All data will be made publicly available at the IJF website.[6]

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 to 2012. In addition, we conducted our own expert surveys for the three elections from 2004 to 2012. The results of combining 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, polls do not provide predictions but rather 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 as the time before the election was longer. 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, inaccurate measurement, 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 near 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 forecast Obama to win over McCain in 2008 with an error of only two electoral votes. In both elections, the median statistics approach missed the winner in only one

---

and 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 slightly differ from the calculation of *ex ante* forecasts that was actually performed in 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 of each election, which were published in *Foresight – The International Journal of Applied Forecasting* (Cuzán *et al.*, 2005; Graefe *et al.*, 2009, 2013).

[6] For now, the links to the data files can be accessed at: http://dl.dropbox.com/u/3662406/Data/PollyVote/Links_PollyVote_data.pdf

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 into account the historical record of the polls when making predictions of the election outcome. For example, assume that the incumbent leads the polls by 20 points in July. In analyzing historical polls conducted 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 & Wink 1990; Campbell 1996; Erikson & Wlezien 2008).

In the present study, we adopted an approach for combining and damping polls that is similar to 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. Successive updating was used to calculate *ex ante* poll projections. That is, when generating poll projections of the 1992 election, only historical data from the elections from 1952 to 1988 were used. When calculating poll projections of 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). 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 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, from which we re-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.[7] In 2000, the average forecast was 72% more accurate than the typical forecast from fifteen experts.[8]

For the three elections from 2004 to 2012, 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 politicos. 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 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 forecasts made by the individual experts.[9] Because our panelists did not meet in person, the possibility of bias due to the influence 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 outcome of U.S. presidential elections. Most models include two to five variables and typically combine indicators of economic conditions and public opinion to measure the incumbent's performance. For example, models by Abramowitz (2012), Campbell (2012), Lewis-Beck and Tien (2012), and Erikson and Wlezien (2012a) 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 Lewis-Beck and Rice (1992), Campbell and Garand (2000), and Jones (2002). For overviews of the variables used in the most popular models see Jones and Cuzán (2008) and Holbrook (2010).

Since the 1990s, forecasts of competing models have been regularly published near Labor Day of the election year. For the past five elections, the forecasts of leading models were published in *American*

---

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

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

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

Politics Research, 24(4) and *PS: Political Science and Politics*, 34(1), 37(4), 41(4) and 45(4). Most models predicting presidential elections have produced forecasts using data available near the end of July in the election year. Usually models have correctly predicted the election winner, albeit by varying accuracy as to candidates' vote shares. Forecast errors for a single model can vary widely across elections, and the structure of some of the models has changed over time, so it is difficult to identify the most accurate models.

Prior research demonstrated that combining predictions from election forecasting models is beneficial to 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 for how long 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, 2000, Table 1), which was 3.0 percentage points. By comparison, the combined error for 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, Erikson *et al.* (2001) showed that creating models that combine structural variables with public opinion further increases accuracy. The authors added presidential approval as an additional variable to the 48 models, thus doubling the number of models to 96. The sum of the absolute errors for their averaged models was 32% lower than for the averaged Bartels and Zaller models.

Montgomery *et al.* (2012) combined the forecasts from six established econometric models based on their past performance 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 the error of a typical individual model. However, as 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, forecasts for most models were released by late July, and some were 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.[10]

[10] Model forecasts by Abramowitz (2012), Campbell (2012), Fair (2009), and Erikson & Wlezien (2012a) were available for all six elections. Forecasts by Holbrook (2012), Lewis-Beck and Tien (2012), Lockerbie (2012), and Norpoth & 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 Lichtman (2008), Graefe and Armstrong (2012), Jerôme & Jerôme-Speziari (2012), DeSart and Holbrook (2003), and Klarner (2008) were available for 2008 and 2012. A forecast by Lewis-Beck and Rice (1992) and Sigelman (1994) was available for the 1992 election. A forecast by Haynes and Stone (2008) was available for the 2008 election. A forecasts by

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, the oldest being the *Iowa Electronic Markets* (IEM), which were 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 et al. (2008) determined that 74% of the time the IEM forecasts were closer to the actual election result than polls conducted on the same day. However, Erikson and Wlezien (2008) found *poll projections* to be more accurate than IEM forecasts.

Prediction market forecasts can be negatively affected 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. We thus 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 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 technique and assumptions, in the types of data used, and in data sources. We recognized that the demonstrated accuracy of the IEM and poll projections might diminish the gains from combining across methods. We also were aware that the impact of a dominant method tends to fade as the number of component methods increases.

For each day in the forecast horizon, we calculated a simple average across the combined

---

Armstrong & Graefe (2011), Campbell's (2012) convention bump model, Berry & Bickers (2012), Graefe (2012), Graefe & Armstrong (2012), Lewis-Beck and Rice's (2012) proxy model, and Nate Silver's FiveThirtyEight.com was available for the 2012 election.

component forecasts: poll projections, experts, models, and IEM. We refer to this overall combined forecast as the *PollyVote.*[11]

5.2. Results

All of the reported forecasts 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 from polls, models, and the IEM. Our own expert forecasts were available only for the three elections in 2004, 2008, and 2012, for a total of 296 daily forecasts.[12]

5.2.1. A note on error measures

We used the absolute error as a measure of accuracy (that is, the difference between the predicted and actual vote shares, regardless whether the error was positive or negative). In presenting the gains achieved through combining, we report the "error reduction" in percent. 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 by Abramowitz (2012) and Erikson & 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 accuracy across time periods such as days or years, we report mean error reduction (MER). The MER for a particular election year is determined by averaging the typical and combined errors across the 100-day time-period before calculating the error reduction. The MER across years is the simple average of the error reduction of each particular year.[13]

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

[11] PollyVote stands for "many" and "politics." On our website, we playfully adopted a parrot as a mascot because the method does little else than repeat and combine what it borrows (or "hears") from others.

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

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

approach produced less accurate forecasts than individual polls in 2008.[14] 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 the error of the combined forecasts of component methods. Across the six elections, the PollyVote provided more accurate forecasts than each 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.

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

	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**

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. Gains in accuracy

[14] The poor performance of poll projections in 2008 can likely be attributed to the economic crisis that hit in mid-September of that year, less than two months before Election Day. With 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.

were large compared to the typical individual poll (59%), the typical model (58%), and the typical expert (55%). In each case, combining reduced 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 the benefits of combining of 12%.[15]

5.2.5. Accuracy gains for different combinations of component methods

Table 2 shows the percentage of days in which bracketing occurred and the MER compared to the typical component method for the each of the three elections from 2004 to 2012.[16] As expected, the percent 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% error reduction 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, gains from combining models and poll projections were smallest (17%). A possible reason for the low rate of bracketing for models and poll projections might be that many models already include information from polls to measure 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 combining models with methods that incorporate human judgment, such as expert forecasts or the IEM. Gains in accuracy were also relatively small when combining poll projections and the IEM forecasts. This conforms to results by 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. Error reductions were largest if the model forecasts were combined with human judgment from experts and the IEM (48%). The error reductions were smallest – although still at the substantial level of 31% – for the combination of models, polls, and the IEM.

---

[15] The "hit rate" provides additional insight on the relative accuracy of the PollyVote and the IEM. Hit rate refers to the frequency with which forecasts of a given method correctly predict the popular vote winner, expressed as a percent of all available forecasts of that method. The hit rate thus measures a method's capability to answer the question that is probably most interesting to the regular consumer of election forecasts: who will win (rather than what will a candidate's share of the vote 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.

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

Table 2: Bracketing and mean error reduction for different combinations of component methods (2004 to 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**

5.2.5.3. 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 out of four cases (72%), combining the forecasts from all four component methods produced bracketing.

5.2.6. Benefits of combining forecasts under uncertainty

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

5.2.6.1. Uncertainty due to disagreement among forecasts

If forecasts derived from different methods agree, certainty about the situation usually increases. In contrast, high disagreement among forecasts indicates high uncertainty. Disagreement among forecasts is often used as a conservative *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 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 forecast at 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 shown in Table 3, on nearly half of all (285 out of 600) days, the range between the component forecasts was within two to four percentage points. 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 uncertainty increased. That is, the benefits from combining were larger when disagreement among component forecasts, and in effect the chance of bracketing, was higher.

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

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

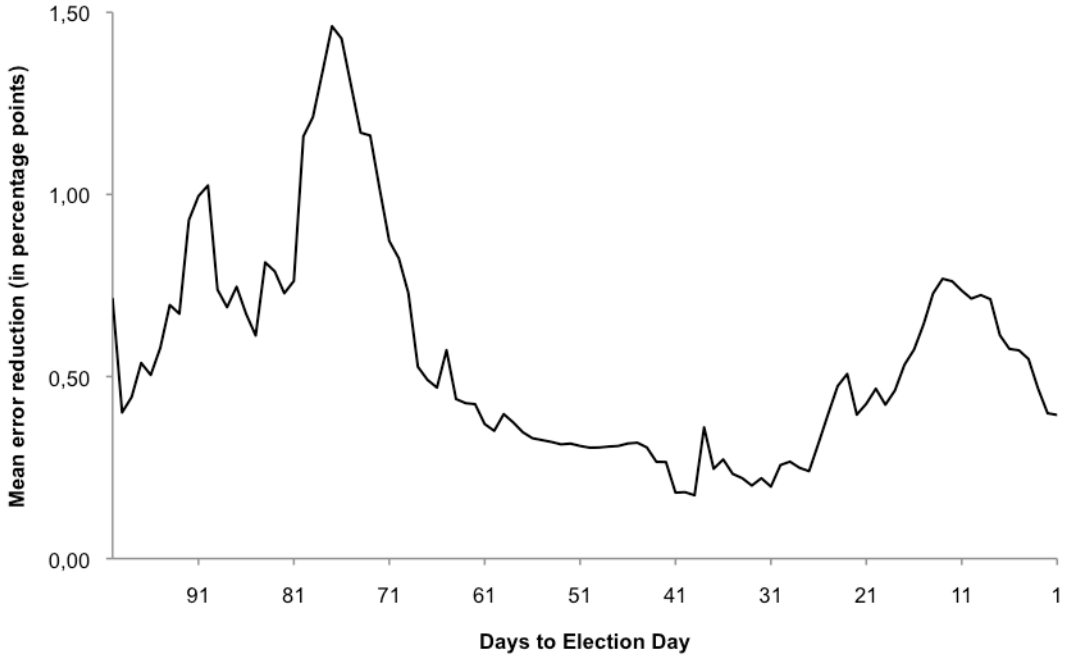
5.2.6.2. Uncertainty due to long time horizons

Uncertainty usually increases with the time horizon of the forecast. Accordingly, combining should be more helpful early in a campaign. Figure 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 uncertainty decreases. Interestingly, the gains from combining increase again in the period from one month to two weeks before Election Day, which is about the time when the presidential debates are usually held. It is up to future research to clarify what is going on late in the campaign, for example, whether the results are

driven by a particular forecasting component.

Figure 1: Mean error reduction of the PollyVote forecast compared to the forecast of the typical component over last 100 days across the six elections from 1992-2012



6. Discussion

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, the PollyVote forecast reduced error by 55% to 59%. Compared to the original IEM, essentially a sophisticated approach for aggregating and combining dispersed information, the PollyVote reduced error by 16%. Across the six elections, the PollyVote provided more accurate forecasts than each of its components. While combining is useful under all conditions, it is especially valuable in situations involving high uncertainty.

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 being 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 points in time 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 various methods available to forecasters, combining is one of the most effective and reliable ways to improve 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 costs exceed 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 U.S. companies responding, 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 individual forecasts based on unaided judgment. Such approaches to combining do not conform to the procedures as described in this paper.

We believe that combining, properly defined and implemented, is in little use today. A number of possible explanations exist for the low usage of formal combining:

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 this through experience. In a series of experiments with MBA students at INSEAD, a majority of participants thought that an average of estimates would reflect only average performance (Larrick & Soll 2006).

Combining seems too simple. Hogarth (2012) 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 from simple models. 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 the forecasting institutions. Forecasters face a general trade-off between accuracy and attention. More extreme forecasts usually gain more attention, and the media are 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. 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 may select a forecast that suits their biases.

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 reduced accuracy.

One goal of the PollyVote.com project is to help people to overcome these barriers by using the high-profile application of forecasting U.S. Presidential Election outcomes to demonstrate the benefits of combining. Software providers might also contribute by including combining as a default. That is, software solutions should require users to actively opt out of combining 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 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 U.S. 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 was shown to be much more useful as uncertainty increases. For important forecasts, the costs of combining forecasts are likely to be trivial relevant 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 accurately summarized their research, and we thank all who replied. 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 & 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. (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.

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.

Armstrong, J. S., Du, R., Graefe, A., Green, K. C. & House, A. (2013). Predictive validity of evidence-based advertising principles, Working paper, Available at: http://advertisingprinciples.com/images/stories/PredictivevalidityofEBAv84-clean.docx.pdf.

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.

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

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

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

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. & 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 & 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,* Winter 2005, 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. & Wlezien, C. (2012a). The objective and subjective economy and the presidential vote, *PS: Political Science & Politics*, 45, 620–624.

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

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. (1999). Presidential polls as a time series: The case of 1996, *Public Opinion Quarterly*, 63, 163-177.

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

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(1), 108-121.

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

Graefe, A. (2013). Contributions of the index method to election forecasting, *Working paper*.

Graefe, A. (2012). Issue and leader voting in U.S. presidential elections, *2012 APSA Annual Meeting Paper*, Available at ssrn.com/abstract=2110794.

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

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., Jones, R. J. & Cuzán, A. G. (2013). Combined forecasts of the 2012 election: The Pollyvote. *Foresight: The International Journal of Applied Forecasting,* Winter 2013 (in press).

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,* Winter 2009, 41-42.

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. (2012). Obama's reelection prospects under “Bread and Peace” voting in the 2012 US presidential election, *PS: Political Science & 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*. Oxford: Oxford University Press, pp. 61-79.

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

Holbrook, T. M. (2010). Forecasting US presidential elections. In J. E. Leighley (Ed.), *The Oxford Handbook of American Elections and Political Behavior*, Oxford: Oxford University Press, pp. 346-371.

Jerôme, B. & Jerôme-Speziari, V. (2012). Forecasting the 2012 US presidential election: Lessons from a state-by-state political economy model, *PS: Political Science & 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 U.S. presidential elections: A brief review. *Foresight: The International Journal of Applied Forecasting,* Summer 2008, 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. (2008). Forecasting the 2008 U.S. House, Senate and Presidential Elections at the district and state level, *PS: Political Science & 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. & Tien, C. (2012). Election forecasting for turbulent times, *PS: Political Science & Politics,* 45, 625-629.

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

Quarterly Press.

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 & 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 & Politics,* 45, 614-617.

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.

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.