Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 24 (2008) 259–271
international journal of forecasting
www.elsevier.com/locate/ijforecast

Evaluating U.S. presidential election forecasts and forecasting equations

James E. Campbell

Department of Political Science, University at Buffalo, SUNY, Buffalo, NY 14260, United States

Abstract

This article examines four problems with past evaluations of presidential election forecasting and suggests one aspect of the models that could be improved. Past criticism has had problems with establishing an overall appraisal of the forecasting equations, in assessing the accuracy of both the forecasting models and their forecasts of individual election results, in identifying the theoretical foundations of forecasts, and in distinguishing between data-mining and learning in model revisions. I contend that overall assessments are innately arbitrary, that benchmarks can be established for reasonable evaluations of forecast accuracy, that blanket assessments of forecasts are unwarranted, that there are strong (but necessarily limited) theoretical foundations for the models, and that models should be revised in the light of experience, while remaining careful to avoid data-mining. The article also examines the question of whether current forecasting models grounded in retrospective voting theory should be revised to take into account the partial-referendum nature of non-incumbent, open-seat elections such as the 2008 election.

Keywords: Evaluating forecasts; Data mining; Econometric models; Forecasting criticism; Voting; Presidential incumbency; Open-seat elections

It is commonplace to lament the sad state of political forecasting. Moreover, suspicions that the entire enterprise is intellectually bankrupt have only been fortified by the most recent forecasting fiasco; the unanimous declaration by quantitative modelers of presidential elections at the American Political Science Association in August 2000 that we could ignore the frantic rhetorical posturing of the next few months. Election campaigns are tales of sound and fury but of no significance because of the offsetting effects of each side's propaganda broadsides. The die had been cast: Gore would defeat Bush by decisive, even landslide, margins.– Philip E. Tetlock (2005, p. 25)

The approach [of the election forecasters] is totally inductive....[T]hese empirical exercises would work better if they were supported by some model or hypothesis about the likely relationship between relevant variables... In successive exercises, changes in the set and measurement of variables are driven by trial and error with statistical coefficients... A reasoned model is still missing for why the impacts

E-mail address: jcampbel@buffalo.edu.

0169-2070/$ - see front matter © 2008 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2008.03.001

of relevant variables should be additive rather than multiply together or interact in still other ways.– Josep M. Colomer (2007, pp. 140–141)

In its relatively short history, election forecasting has managed to attract an inordinate amount of criticism. While some has been reasonable and fair, much has not been. Though criticism is part of the research process and may help the collective endeavor to progress by identifying errors and suggesting corrections, unreasonable criticism may divert attention away from real problems and unjustly undermine confidence in the enterprise. The purpose of this article is twofold. The first is to examine and respond to some criticisms that have been leveled at presidential election forecasters and their models. Most of this criticism has been of little value. It has often been plainly wrong, misleading, and generally not constructive. The second purpose of this article is to offer some constructive criticism of election forecasting models, in order that they may be improved by recognizing the distinction between incumbent and open-seat elections. This is an especially relevant criticism as we approach the open-seat election of 2008.

1. Evaluating election forecasting criticisms

Though a number of valid criticisms of election forecasting can be made, past criticism of the enterprise has been largely off-target and not constructive. Virtually every aspect of the Tetlock and Colomer epigrams above, for instance, are erroneous in some way (as will shortly be explained). Some of the problems with other forecast criticisms are idiosyncratic, often dealing just with the plain facts of what forecasters have or have not done. For example, after observing Ray Fair's use of out-of-sample testing of his equation (Fair, 2002), Rebecca Morton (2006, p. 373) noted that "political science models have less ability to do such out-of-sample predictions, given the data limitations they have to begin with (Norpoth being an exception)." In fact, most political science equations routinely perform out-of-sample tests, and have done so for many years (e.g., Campbell, 2004b; Campbell & Wink, 1990; Holbrook, 2004; Lewis-Beck & Tien, 2004; Lockerbie, 2004; Wlezien & Erikson, 2004; see also Lewis-Beck, 2005, p. 153; and Campbell, 2000, p. 175).

Apart from particular problems with critiques of forecasting equations, such as the one just addressed, there have been four more general problems with critiques of election forecasting. Critics have had problems in appraising forecasting equations, in evaluating the accuracy of both forecasting models generally and forecasts of individual elections, in identifying the theoretical foundations of forecasts, and in distinguishing between data-mining and learning in the revising of models.

1.1. Evaluating forecasting models

Like any effort with a common objective, there is an element of competition in forecasting. Which forecasting equation is best? While different forecasters and observers use different values in making such an overall evaluation of the models, the most notable and systematic of these efforts has been Michael Lewis-Beck's. Lewis-Beck (Lewis-Beck, 1985, 2005; Lewis-Beck & Tien, 2007) identifies four dimensions of forecast model quality: accuracy (A), parsimony (P), reproducibility (R), and lead time (L). He then scores each model on each dimension with a score of between 0 and 2, and combines these scores using the following formula to arrive at a quality index (Quality 1) for the forecasting equation:

$$\text{Quality 1} = ((3A + P + R)L)/M, \tag{1}$$

where M is the maximum value of the numerator, so that the index ranges from 0 to 1.

This index has several problems, relating to both the rough scoring of the components and their relative weighting in the index.[1] Though the four components of the quality index are justified, there is no justification for the particular formula chosen to combine them.

[1] The coding for some components appears somewhat ad hoc. Consider the scoring of the lead-time component. In the 2004 election, Norpoth's (2004) model was the earliest, made 278 days before the election. His equation was scored the maximum value of 2 for lead time. The Abramowitz (2004) model, producing a forecast 184 days later, received the same score of 2. Thirty days later, Holbrook's equation produced a forecast, and was scored 1.5 for lead time. Making a forecast just a week later, my forecast was scored 1.0 for lead time. To recap: a 184 day break and no score change, a 30 day break and a drop of half a point, a seven day break (and still two months before Election Day) and another half point drop (Lewis-Beck, 2005, p. 157).

Table 1
Quality indices of forecasting equations

Model	Accuracy	Parsimony	Reproduce	Lead	Quality 1	Quality 2
Abramowitz	1.5	1	2	2.0	0.75 (1)	0.63 (3)
Campbell	2.0	2	2	1.0	0.50 (6)	0.83 (1)
Holbrook	1.5	1	2	1.5	0.56 (5)	0.56 (4)
Lockerbie	1.0	2	2	2.0	0.70 (2)	0.50 (5)
Lewis-Beck/Tien	2.0	1	2	1.5	0.68 (3)	0.75 (2)
Norpoth	1.5	1	1	2.0	0.65 (4)	0.50 (5)
Wlezien/Erikson	1.0	2	1	1.5	0.45 (7)	0.38 (7)

Note: except for the Quality 2 index, the data are from Lewis-Beck (2005, Table 1). The Quality 2 index is constructed by the author.
Quality 1 is Lewis-Beck's index $(3A+P+R)L/M$, where M is the maximum of the numerator.
Quality 2 is an alternative index $(L+P+R)3A/M$, where M is the maximum of the numerator.

Since these four components could be combined in an infinite number of ways, the decision to use the particular formula of Eq. (1) for the index, without any clear justification, is arbitrary. To illustrate, using the same scoring of the components that Lewis-Beck uses (2005, 157), another formula for quality (Quality 2), emphasizing accuracy rather than lead time, is:

$$\text{Quality 2} = ((L+P+R)(3A))/M \qquad (2)$$

The differences between the two quality indices, using the same components scored in precisely the same way, are displayed in Table 1. The table presents the two quality scores for the seven models evaluated by Lewis-Beck (2005). As can readily be observed from the models' rankings (in parentheses), the two quality indices produce very different and barely correlated scores ($r=0.12$). The fact that one obtains such different scores from the two composites suggests that neither is reliably grounded — and this is without raising questions about the idiosyncratic scoring of the components that are entered into the indices, or whether these are the full set of components that should be considered (e.g., model stability could be also added as a component). Without some established criteria for combining the index components in a particular way, there is no basis for using one quality index over any other. Before we can claim one index of quality, we must first defend the particular mix of qualities that make one model better than another.

1.2. Evaluating the accuracy of forecasts

The ultimate standard for any forecast or any forecasting model must be its accuracy. Lead time, parsimony, and reproducibility cannot compensate for the inaccuracy of a model's forecasts. Though one would think that accuracy could be gauged easily, critics have had substantial difficulties in assessing the accuracy of both forecasting models and individual election forecasts. Forecasts are commonly judged to have been either right or wrong (as the epigrams illustrate) without benchmarks, using some unstated, vague, and unsubstantiated metric. As if this did not make the forecasting enterprise frustrating enough, forecasters are regularly subjected to a "Catch 22" of election forecasting. After an election, critics pronounce the forecasts to have been either accurate but painfully obvious or inaccurate and the failed product of ivory-tower dwelling political numerologists. It is a no-win situation.

What makes the utter arbitrariness of these evaluations more than breathtaking is that these off-the-cuff judgments of the rightness or wrongness of forecasts are often applied to matters that the forecasts never predicted. For instance, the forecasting models in the 2000 election were roundly criticized (see the Tetlock epigram) for having wrongly predicted that Democrat Al Gore would defeat Republican George W. Bush. Although George W. Bush won a majority of the electoral vote and was thus elected president, Al Gore received a majority of the two-party popular presidential vote. None of the major forecast models predicted the electoral vote, or who would receive a majority of these votes. They predicted the popular vote and, in this dichotomous sense, each of the major forecasting models could be said to have been "right" in predicting a popular vote plurality for Gore in 2000 (Campbell, 2001a).

The main point, however, is that these forecasts should not be judged as simply right or wrong. Almost without exception, forecasting equations of U.S. presidential elections offer predictions of the two-party vote percentage for the major party presidential candidates.[2] They do not directly forecast which candidate is expected to win the popular or electoral vote majority. Since the forecast of the vote is an interval measure, the rightness or wrongness of these vote forecasts is only reasonably discussed as a matter of degree, not as a dichotomy. To understand how the accuracy of election forecasts should be judged fairly, we must first understand the limits of what can be expected in the prediction of the vote percentage. From a very practical standpoint, unless random or unpredicted events fall very luckily into place, every forecast of the vote can be expected to be "wrong," to differ from the actual vote percentage by some margin. None of the forecasters claim that their models are perfect, exactly anticipating the ramifications of all of the actions of voters and candidates months before the election. They certainly do not claim that political observers should, as Philip Tetlock (2005) puts it, "ignore the frantic rhetorical posturing" of the general election campaign. Nor do the forecasters claim that the ingredients, the predictor variables, used in generating the forecasts are without error. We know, for example, that the measures of the pre-election economy used in many forecasting models are refined and improved years and sometimes decades after the election, and that public opinion measures of approval and voter preferences contain sampling and other measurement errors.

However, the fact that election forecast errors are matters of degree and are to be expected does not mean that forecasters are "off the hook" (Campbell, 2004a, p. 734). Interval forecasts are not beyond evaluation. It only means that the evaluation of the forecasts is not a simple either-or, right-or-wrong verdict, and that the degree of rightness or wrongness requires that we have some bearing on what can be reasonably expected. To obtain some bearing on the accuracy of forecasts, and whether the extent of an error is relatively small or large, requires some benchmarks for comparison. Elsewhere (Campbell, 2005, p. 23) I have suggested that reasonable benchmarks for forecast comparison should be obtained from the information available at the time of the forecasts, including the candidates' standings in the preference polls. This essentially assesses the extent of the error reduction that the forecast equations offer compared to forecasts that could have been generated at the time without the equations.

There are three obvious benchmarks for assessing the forecasting equations (or, for that matter, any other forecasting technique such as the market approach). The first and easiest benchmark to surpass, easiest because it is made without any contemporary information about the election, is the error associated with a random guess of a 50–50 vote, or the slightly more informed guess of the mean in-party vote in past elections. A second benchmark, setting a more difficult standard, is the error associated with polls conducted around the time of the forecast. The third and most demanding benchmark is the error associated with the polls conducted just prior to the election.

Although these three benchmarks yield different errors in different election years for differently timed models, their average errors over a series of elections can provide some guidance in determining how good or how poor a forecast was. In the fifteen elections from 1948 to 2004, the mean absolute error of naively guessing either a tie vote or the mean in-party vote (the null model) has been 4.5 percentage points. Over this same set of elections, the mean absolute error in presidential preference polls (Gallup Polls) conducted after the conventions and up to Labor Day (around the time that the fall campaign gets underway and when most of the forecasts are made) has been about 4.0 percentage points — only about half a percentage point better than the no-contemporary-information prediction. The third benchmark, preference polls conducted in early November and just before the election, have had an average error since 1948 of about 2.3 percentage points. To put the demanding nature of this standard in perspective, the mean absolute error in the reported two-party presidential vote of the American National Election Studies from 1952 to 2004, a vote measure collected *after* the election, has been 2.2 percentage points (Campbell, 2008, Table

[2] Though it is no longer used, Lewis-Beck and Rice (1992) constructed a model to forecast the electoral vote. Although not a quantitative forecaster, Allan Lichtman (1996) also has a qualitative model that predicts the winner of the presidential election rather than the vote share for the in-party candidate.

Table 2
Evaluating presidential vote forecasts relative to three benchmarks

Benchmarks	Mean absolute error from vote	Accuracy evaluation
	<2.3%	Quite accurate
November/pre-election day	2.3%	
polls	2.3 to 3.1%	Reasonably accurate
	3.2 to 4.0%	Fairly accurate
Post-convention/Labor	4.0%	
Day polls	4.0 to 4.5%	Inaccurate
Random split/mean	4.5%	
in-party vote	>4.5%	Quite inaccurate

Note: the polls and votes are two-party divisions. The mean absolute errors in the polls are based on the two-party division of Gallup preference polls in elections from 1948 to 2004.

3.3) — only a tenth of a percentage point less than the final poll standard.

How might these benchmarks provide some bearing on evaluating forecast accuracy? Table 2 puts the benchmarks in perspective. The evaluation is clear at the extremes. A forecast can fairly be judged as "very inaccurate" if its error is larger than what could be achieved by a naive model. Errors of 4.5 percentage points or greater should be deemed very inaccurate. At the other extreme, a forecast can be judged as "quite accurate" if it is more accurate than the polls around Election Day. Errors of 2.3 percentage points or less deserve this praise. The cut-points between these extreme benchmarks are not as clear-cut. Accuracy between that of the Election Day polls and the post-convention/Labor Day polls could be classified as "reasonably accurate" if they are closer to the accuracy of the election day polls (errors of 2.3 to 3.1 percentage points) or "fairly accurate" if they are closer to the accuracy of the post-convention/Labor Day polls (3.2 to 4.0 percentage points). Forecast errors greater than those of the post-convention/Labor Day polls but less than naive guesses (errors of 4.0 to 4.5 percentage points) might be best classified as "inaccurate."

Of course, these are the mean errors of the benchmarks, and not an indication of forecast accuracy in any particular election. The mean errors of the benchmarks might be considered more suitable for evaluating the general accuracy of a forecasting model, rather than a particular forecast for a particular election. However, there is a problem with using benchmarks for particular elections. The problem is that there is no reliable order of the accuracy standards of the benchmarks in each election. While the naive forecast in general has the largest error of the benchmarks, it could yield the smallest error in any given election (e.g., the 2000 election). Suppose there is an election in which the naive model is 50% for the in-party candidate, the Labor Day polls indicate 55%, the Election Day polls indicate 56%, and the actual vote is 52%. How should a forecast of 54% be evaluated? The good news for the forecast is that it is more accurate than the contemporary poll (2 points off rather than 3) and much more accurate than the Election Day polls (2 points off rather than 4). Looks like a good forecast. On the other hand, the forecast is no more accurate than the naive guess (both 2 points off). Aside from indicating the hazards of evaluating forecasts in close elections (where random forecasts appear to do quite well), the bottom line is that the only reasonable standards for determining the accuracy of election forecasts are the averages of the three benchmarks, and not the particular values of the benchmark in any one election year.

Critics of forecasting have not only had a difficult time in evaluating the accuracy of election forecasting models, but have often not bothered to evaluate the accuracy of individual forecasts, effectively regarding all forecasts as a monolith. Despite the wide variety of forecasting equations and their varied forecasts of the in-party vote in different elections, evaluation verdicts have often been made *en masse*.

The mistake of pronouncing a group verdict on all forecasts was most clearly in evidence in the aftermath of the 2000 election. The Tetlock epigram is a good example of painting all forecasts with the same broad brush, but a number of other examples exist as well (e.g., Johnston, Hagen, & Jamieson, 2004, p. 101). Writing in *The Chronicle of Higher Education* shortly after the 2000 election, D.W. Miller (2000) wrote that "Political scientists who tried to forecast the voting in the presidential race also took it on the chin." Bill Mayer offered a similar verdict in observing that political scientists were "likely to remember 2000 as the Year when the Forecasting Models went Crash" (2003, p. 153).

Did a single, negative verdict fit the varied forecasts of the 2000 election, or were these critics off-base in evaluating all of the forecasts as a "fiasco" of the

models that "went crash"? The actual two-party popular vote for in-party candidate Al Gore in the 2000 election was 50.3 percentage points. The forecasts ranged from 50.8 to 60.3 percentage points. The errors, thus, ranged from 0.5 to 10.0 percentage points — a huge 9.5 percentage point spread in the forecast models' errors. While five of the forecasts would be classified as very inaccurate by the standards discussed above, three would be evaluated far more favorably. Fair's forecast of 50.8% for Gore (Fair, 2002), an error of just half a percentage point, would have to be judged as quite accurate. My forecast of 52.8% (Campbell, 2001b), an error of 2.5 percentage points, and Abramowitz's forecast of 53.2% (Abramowitz, 2001), an error of 2.9 points, fall in the reasonably accurate category. Certainly none of these three forecasts could be fairly labeled a "fiasco", and none of these three models "went crash" in 2000, which was a fairly tough election for everyone.[3] If reasonable criteria are applied and the facts about the forecasts are examined, there is simply no good reason to "lament the sad state of political forecasting" (Tetlock, 2005, p. 25).

1.3. Locating theoretical foundations

Critics of forecasting have made mistakes in judging not only the outputs of the forecasts, but their origins as well. Some critics labor under the mistaken impression that election forecasting is atheoretical, pseudo-scientific data-mining (Eisenhower & Nelson, 2000). This may be what lies behind what Philip Tetlock's "suspicions that the entire enterprise is intellectually bankrupt." Josep Colomer (2007) makes the atheoretical charge bluntly: "the approach [of the election forecasters] is totally inductive." If one did not know better, you might think from what the critics have written, that the forecasts were based on the quality of the election year's Beaujolais Nouveau or the league of the team winning baseball's World Series.

The idea that the major presidential election forecasting models lack theoretical foundations is absolute nonsense. Most, if not all, of the forecasting models are firmly rooted in the theory of retrospective voting generally, and the theory of economic voting more particularly. The theory of retrospective voting, that voters evaluate the record of the in-party, has a long intellectual history. As Walter Lippmann wrote long ago in *The Phantom Public*: "To support the Ins when things are going well; to support the Outs when they seem to be going badly, this, in spite of all that has been said about Tweedledum and Tweedledee, is the essence of popular government" (1925, p. 126). V.O. Key (1966) in *The Responsible Electorate*, and later Morris P. Fiorina (1981) in *Retrospective Voting in American National Elections*, further developed and tested the retrospective voting theory. Retrospective voting theory provides a theoretical foundation for including presidential approval ratings in the models, as well as a foundation for using early presidential preference polls (since evaluations of past performance permit early preferences to be developed). Beyond retrospective voting theories, the models draw on theories of economic voting, partisanship, incumbency, and campaign effects. Edward Tufte (1978), Michael Lewis-Beck (1988), Robert Erikson (1989) and scores of others (see, for example, Lewis-Beck, 2006; Lewis-Beck & Stegmaier, 2000) have fleshed out theories of economic voting, and most models include some broad-based economic indicator on this basis. Recently, several forecasting models have been adapted to reflect advances in theoretical research about the muted effects of partial responsibility for the economy (Nadeau & Lewis-Beck, 2001; Norpoth, 2002; Whitten & Palmer, 1999).

Though less frequently noted, many of the forecasting equations also have a firm theoretical foundation in both theories of partisanship and theories of presidential campaign effects. Theories of the impact of voter partisanship, in the tradition of *The American Voter* (Campbell, Converse, Miller, & Stokes, 1960) and the normal vote (Converse, 1966a, b), suggest a good deal of aggregate stability in the vote division. The direct and indirect effects of long-term considerations, such as partisanship, on voter decision-making provides a basis for thinking that

[3] Three of the forecasts (Cuzan & Bundrick, 2005; Lewis-Beck & Tien, 2004; Wlezien & Erikson, 2004) were in the quite accurate category in 2004. Two (Abramowitz, 2004; Campbell, 2004b) were in the reasonably accurate category, and one (Norpoth, 2004) was in the fairly accurate category (though the exceptionally long lead-time for this model might be reason to be a bit more generous in evaluating its accuracy). Three (Fair, 2004; Holbrook, 2004; Lockerbie, 2004) were in the quite inaccurate category in 2004.

public opinion before the campaign might be a good indicator of the public's eventual vote division. Theories about campaign effects, in the tradition of *The People's Choice* (Lazarsfeld, Berelson, & Gaudet, 1944), also suggest a good deal of voter stability after the nominating conventions (see also Gelman & King, 1993; Holbrook, 1996). In *The American Campaign* (Campbell, 2008), originally published in 2000, I proposed and tested the theory of the predictable campaign that provides the basis for the trial-heat and economy forecasting equation. This theory explicitly rejects Philip Tetlock's contention that the forecasting equations suggest that "Election campaigns are tales of sound and fury but of no significance because of the offsetting effects of each side's propaganda broadsides." Campaigns are necessary to convert the fundamentals or raw materials of politics into votes, and the general parity of candidate campaigns does not lead to insignificant campaign effects, but to a narrowing effect of campaigns.[4]

Contrary to Colomer's statements, presidential election forecasting research is steeped in electoral theory.[5] This is not to say, however, that theory can offer very specific guidance about model specification and measurement issues. Contrary to the idea that forecasting enterprise can be a purely "theory-driven process" (Lewis-Beck, 2005, p. 154), there are a host of very practical decisions (the choice of indicators, the timing of lags, coding decisions for the undecided and don't knows in survey questions, etc.) that must be made in assembling a forecasting model. Very different models coming out of these decisions would comport equally well with the same theoretical foundations.

Forecasting cannot be based entirely on electoral theory. Perhaps nothing demonstrates this point more clearly than the considerable differences in forecasts of the same election produced by different models with the same theoretical heritage. In 2000, both Holbrook's (2001) and Abramowitz's (2001) forecast equations had the same three components in common: presidential approval, the general economy, and the number of terms that the presidential party had held office. The indicators of the first two components, however, differed. Despite the common theoretical basis for the equations, the two forecasts differed by *more than seven percentage points*. In-party candidate Al Gore was predicted by Abramowitz's model to receive 53.2% of the two-party vote, and 60.3% of the two-party vote by Holbrook's model. Guided by the same theory, the models produced quite different forecasts. Clearly theory only takes us so far in structuring forecasting models. The Bayesian averaging of models approach of Bartels and Zaller (2001) and Sidman, Mak and Lebo (2008-this volume) is essentially a recognition of the limits of theory in forecasting model construction.

I will go further than suggesting that theory is limited practically in the guidance it can provide for forecasting: the construction of forecasting models *should not* necessarily be driven exclusively by explanatory theory (Campbell, 2000, p. 182). The purpose of electoral theory is to have a deep understanding of what causes the vote. The purpose of forecasting is to accurately anticipate what the vote will be. There is reason to believe that these purposes are related, but they are definitely not the same thing. For example, knowing that vote intentions at a time well before the election are likely to remain stable to the time of the election is important information for forecasting purposes, but is not very illuminating from an explanatory standpoint. Theory is unquestionably important to forecasting, but concern for theory should not impede forecasting accuracy.

1.4. Distinguishing data-mining from learning

Since electoral theories provide a great deal of latitude in the specification of forecasting equations, there is reason for critics to be concerned that decisions

[4] Tetlock is not the first, and unfortunately will probably not be the last, to make the mistake of thinking that the models assume no campaign effects. Jay P. Greene in *The American Prospect* wrote that "These models share not only a methodology but also a political assumption: campaigns do not significantly affect election outcomes" (2002).

[5] Colomer also contends that "A reasoned model is still missing for why the impacts of relevant variables should be additive rather than multiply together or interact in still other ways" (2007, p. 141). Unfortunately, he does not indicate which variables should be specified as interactions. He also does not consider the limitations imposed on the models by the small number of available observations. Finally, Colomer states that "almost all regression results are given in tabular form, not even as equations" (2007, p. 140). Given the equivalence of the equation and tabular content, it is unclear what Colomer reads into the different formats in which the equations' coefficients are presented.

about model specifications may be made only with an eye to reducing in-sample and even out-of-sample errors. Modifications made to equations immediately after an election (or in the weeks leading into an election season) might be especially suspicious on these grounds. The suspicion is that if the model did not work especially well in one election, that the forecaster would simply adjust the model so that it would have worked well. This data-mining orientation to model specification and respecification is certainly something to be concerned about (Morton, 2006, p. 373). As Colomer states, "In successive exercises, changes in the set and measurement of variables are driven by trial and error with statistical coefficients" (2007, p. 141).

The critics have one foot on firm ground here. Certainly forecasters could be tempted to revise their models to produce the strongest fit (or retrofit) to past elections (or a forecast they would like politically in the coming election). Forecasters ought to be wary of too easily making changes in their models. As I have argued elsewhere, "model stability (the constancy of model specification from one election to the next) must be a goal of election forecasting along with prediction accuracy and lead time before the election" (Campbell, 2004a, p. 735). It could easily be added to the list of quality components identified by Lewis-Beck (1985) in constructing his quality index. The credibility of the forecasts depends in no small part on their track records, and without some model stability there can be no meaningful track record. Consumers and critics of forecasting should be assured that models are not recast willy-nilly to retrofit the equation to the peculiarities of the past election, or to fit the equation to the particular issues looming in the next election.

While I agree with the concerns of the critics in this respect, it is also important to understand that forecasters are dealing with limited data and a small number of elections, and that there is a great deal of room to learn about the possibilities and limitations of forecasting equations. The idea that model stability should be a goal of election forecasting does not mean that forecasters should not learn "from errors in recent trials" (Lewis-Beck, 2005, p. 155). Nothing, including concerns about both model stability and theory, should supersede the goal of greater forecast accuracy in constructing or revising forecasting models. As well as offering a temptation for data-mining, each passing election offers an opportunity to learn about forecasting specifications that can improve the future accuracy of the models.

Forecasting experience from the 2000 election is again instructive. While there was good reason to focus on the effects of Al Gore's unusual prospective-oriented strategy that year, there was also more to learn in 2000 about open-seat presidential elections. In the post-1948 series of presidential elections, the series that most forecasters use to estimate their models, there had only been four previous open-seat contests (1952, 1960, 1968, and 1988). This offered little data on which to base distinctions between open-seat and incumbent elections.[6] The 2000 election added valuable insights about open seat presidential elections, and drew attention to theoretical developments (Nadeau & Lewis-Beck, 2001; Norpoth, 2002; Whitten & Palmer, 1999) that should improve the models. I will elaborate on this shortly.

The point is that model construction and revision must be undertaken in a way that avoids both data-mining and curve fitting, while at the same time allowing for model adaptations that reflect what has been learned from the additional information provided by recent elections. How far forecasters should go in revising an equation or resisting the impulse to revise is a difficult judgment call. On the one hand, curve-fitting should be avoided; but on the other hand, learning from experience should be welcomed. One solution to this tension may be to keep both the unrevised and the revised model alive for several elections until testing under actual forecasting conditions empirically resolves the issue of which specification is superior. This might be considered a more limited and directed form of the multiple model Bayesian averaging approach to forecasting (Bartels & Zaller, 2001). In the 2004 election, with the parties holding an unusually late second convention, I decided to estimate a second equation (with a convention bump variable) that took this development into account. While the choice between these two equations was not resolved by the 2004 election, the dual equation option

[6] Colomer's inquiry into the lack of interaction terms is relevant here. While one would suppose that presidential approval and the economy would be less indicative of the in-party vote when the incumbent was not seeking reelection, there was simply insufficient data to obtain reliable estimates of the difference in these effects with only four open seat elections.

avoided the charge of model-cooking and still allowed adaptations to new information and circumstances.

2. Constructive criticism and the 2008 election

It is clear that forecasters have learned a good deal in their relatively brief history, both from experience in forecasting and from theoretical developments. Virtually all of the current models have been tweaked here or there, some more than others. Adaptations have ranged from adding variables (e.g., the jobs variable in Lewis-Beck and Tien's model (2004)), to discounting economic conditions in elections without an incumbent (Lewis-Beck and Tien's model and my model (Campbell, 2004b)), to coding decisions about "don't knows" (Abramowitz's model (2004)), and whether to use registered voter or likely voter preference polls (my model). Still, there is much more yet to learn. As the forecasting field moves into another election season, what might the field have learned from past experience and electoral research? In particular, what should the field have learned that would be helpful in forecasting the 2008 presidential election?

2.1. Is retrospective voting conditional?

A number of possible improvements could be made in the models: averaging indicator measurements (polls) where possible, using consistent timing of economic measurements in the estimation of the model and the production of the forecast, and maintaining multiple models when considering model revisions. While each of these might improve election forecasting at the margins, there is one larger problem lurking in the background for most of the models; a problem that is especially relevant for the 2008 election, and for forecasting models most heavily grounded in retrospective voting theory. The problem is that many of the models are better suited to predicting races in which the incumbent is running than those lacking an incumbent; and the 2008 election is not only an open-seat election, it will be the first election since 1952 in which neither major party's presidential candidate had previously served as either president or vice president.

There are several reasons to suspect that the models are better at predicting an incumbent's vote than an in-party successor's vote. First, most of the elections used to estimate most forecasting equations have involved incumbent races rather than open seat contests. Of the fifteen elections since 1948, ten involved incumbent candidates and only five were open seat races. If parameter estimates are calculated largely on the basis of one type of election, it is likely that they would not do as well in predicting another type.

Table 3
Presidential incumbency and election margins, 1868–2004

Size of the popular vote for the winning candidate	Incumbent was in the race	No incumbent in the race	All presidential elections
Near dead-heat (less than 51.5%)	3 (14%)	6 (46%)	9 (26%)
Competitive (51.5% to 57.0%)	10 (48%)	5 (38%)	15 (44%)
Landslide (greater than 57.0%)	8 (38%)	2 (15%)	10 (29%)
Total	21 (100%)	13 (100%)	34 (100%)

Note: vote percentages are of the two-party vote. The 1912 election is excluded because of the unprecedented third-party candidacy of Theodore Roosevelt.

Second, as has already been noted, at least with respect to the economy, research suggests that an open-seat election involves less of a retrospective judgment of the incumbent's record than an election with an incumbent (Nadeau & Lewis-Beck, 2001; Norpoth, 2002). Successor candidates are not accorded the full rewards or punishments that apply to incumbents. This logic suggests a discounting of both the economic record and the overall record, as evaluated by the president's approval rating. Some models have taken this into account in their economic measures, but none have discounted presidential approval ratings. It seems unlikely that voters will punish whoever is the Republican successor to President Bush to the same degree that they would him, if he were running in 2008.

Third, the outcomes of open-seat presidential elections have been systematically different from the outcomes of incumbent elections. Open-seat presidential elections have historically been much closer than incumbent elections. Table 3 presents the record of the thirty-four elections since the end of the Civil War. The 1912 election has been excluded because of the unprecedented candidacy of former Republican President Theodore Roosevelt as a third-party candidate.

The elections are categorized by the closeness of the national two-party popular vote as "near dead heat" elections, in which the winning presidential candidate received less than 51.5% of the vote, "competitive" elections, in which the winner received between 51.5 and 57% of the vote, and "landslide" elections, in which more than 57% of the vote was cast for the winning candidate. The table then presents the distributions of the twenty-one presidential elections since 1868 in which an incumbent ran, the thirteen in which the incumbent did not run, and all thirty-four elections.

As is clear from the table, near dead heat elections have been more than three times as likely in open seat elections than when an incumbent was running. Nearly half of all open-seat elections have been near dead-heats, while only about one out of every seven elections with an incumbent has been this close. The mean winning two-party vote in elections since 1868 (1912 excluded) with an incumbent in the race was 56.0%. Without an incumbent running, the mean winning vote was only 53.1%. Put differently, winning candidates in open seat contests typically won with half the margin of winning candidates in elections in which the incumbent was a candidate.

The problem is that many forecasting models do not fully take this difference between open-seat and incumbent elections into account. Some have properly discounted responsibility for the economy, but others have not. Some also include a third-term variable that may partly reflect the open-seat difference. However, none of the forecasting models that use presidential approval as a predictor discount its efficacy in open seat contests. There is no question that presidential approval is a good indicator of the likely vote for the in-party, whether it is the incumbent or a would-be successor. However, there would seem to be little doubt that presidential approval is not an equally good predictor for the two types of in-party candidates. Bush in 1988 was not Reagan to voters and Gore in 2000 was not Clinton. Both could benefit from associations with their popular predecessor, but it certainly seems plausible to believe that neither could effectively claim the full credit.

2.2. Successor versions of two models

As an initial test of the notion that open-seat presidential contests are only partially retrospective referenda on the in-party elections, I compared both my original Labor Day trial-heat and economy model (Campbell, 2004b) and Abramowitz's "time-for-a-change" model (Abramowitz, 2004) with successor variants of these models. In effect, one version of each model treats successor candidates like incumbents, while a second version awards successor candidates only half the credit or blame for the economy and, in the case of the "time-for-a-change" model, the president's approval rating. Ideally, rather than prespecifying the partial credit for successor candidates as half the credit or blame accorded incumbents, we would allow an interaction term to estimate the degree of credit or blame attributed to successor candidates. Because of the small number of open-seat elections in the post-1948 election series being estimated, it is unrealistic to expect an interaction term to yield stable estimates of the partial impact of

Table 4
Original and successor versions of the trial-heat and economy forecasting model, 1948–2004

Dependent variable: the two-party popular vote for the in-party's presidential candidate		
Predictor variables	Full credit economy for successors (1.)	Half credit economy for successors (2.)
Early September preference poll	0.466 (8.106)	0.446 (8.142)
Second-quarter growth rate for real GDP (annualized)	0.563 (4.713)	0.602 (5.223)
Constant	26.414	27.645
Adjusted R^2	0.896	0.909
Standard error of estimate	1.798	1.678
Durbin–Watson	1.873	1.971
Mean out-of-sample absolute error	1.586	1.460
Median out-of-sample absolute error	1.064	0.974
Largest out-of-sample absolute error	5.221	3.861
Elections with 3%+errors	2	2
Mean error with the incumbent running	1.498	1.361
Mean error without incumbent running	1.762	1.658

Note: $N=15$. The coefficients in parentheses are t-ratios. All coefficients are significant at $p<0.01$, one-tailed. The successor specification halved the GDP growth rate variable when an incumbent was not seeking election. This includes the five elections of 1952, 1960, 1968, 1988, and 2000. The mean errors with and without an incumbent in the race are out-of-sample errors.

Table 5
Original and successor versions of Abramowitz's "Time for Change" forecasting model, 1948–2004

Dependent variable: the two-party popular vote for the in-party's presidential candidate		
Predictor variables	Original specification (1.)	Successor specification (2.)
June approval difference	0.106 (3.763)	0.130 (4.448)
In-party term (0 for 1st term, 1 for more)	5.154 (3.754)	3.483 (2.718)
First-half growth rate for real GDP (annualized)	0.724 (3.519)	0.685 (3.624)
Constant	51.121	50.807
Adjusted R^2	0.832	0.845
Standard error of estimate	2.291	2.199
Durbin–Watson	1.898	2.160
Mean out-of-sample absolute error	1.902	1.698
Median out-of-sample absolute error	1.096	1.249
Largest out-of-sample absolute error	5.258	5.688
Elections with 3%+errors	3	3
Mean error with incumbent running	1.672	1.491
Mean error without incumbent running	2.362	2.110

Note: $N=15$. The coefficients in parentheses are t-ratios. All coefficients are significant at $p<0.01$, one-tailed. The successor specification halved both the approval difference and the GDP growth rate variables when an incumbent was not seeking election. This includes the five elections of 1952, 1960, 1968, 1988, and 2000. The mean errors with and without an incumbent in the race are out-of-sample errors.

the record for successor candidates. Since the theory suggests that credit is greater than zero and less than one, the specification of half-credit splits the difference, until additional open-seat elections allow more reliable discount rates to be estimated.

The results of the two sets of full and half credit forecast models are reported in Tables 4 and 5. As the overall summary statistics for the two pairs of models indicate, the half-credit successor versions of the two models produced a slightly stronger fit than specifying that successor candidates receive the full credit or blame. The adjusted R^2 values for the successor models are slightly higher in the successor versions (Eq. (2) in both tables), and the standard errors of the estimates and mean out-of-sample errors are smaller in these specifications.

As suggested above, the mean out-of-sample errors in open-seat elections are smaller in the successor versions of both models than in their original specifications.[7] In addition, the mean out-of-sample errors are also smaller in the successor versions of both models when the incumbent was running. As one might expect, given that the differences between the original and successor discounted versions differ only by a partial discounting of one or two variables, the differences in accuracy are not great. The differences are certainly not large enough to be considered definitive, but the successor specifications appear to be at least as strong as the original versions in both incumbent and open seat elections. Conservatively speaking, it appears that it is an open question, at the very least, as to whether models should be revised to address open seats.

2.3. The 2008 election

The successor specification has several clear implications for forecasts of the 2008 presidential election. Depending on the extent of economic growth in the second quarter of 2008, the forecast for the would-be Republican successor to President Bush (probably John McCain at the time of writing) will be predicted to receive a slightly lower vote than would have been the case if President Bush had been running with identical preference poll and economic numbers. While the coefficient for the economy in the successor version of the equation increased a bit, the increase does not offset the halving of the economic impact for successor candidates. The net differences in the forecasts are not great, in the order of half to one-and-a-half percentage points, but are potentially politically important in what is likely to be a close election.

Since the Abramowitz "time-for-change" model is more deeply rooted in retrospective voting theory, the change to a conditional retrospective specification in the successor version of the equation makes a potentially greater difference to its forecast for 2008. The extent of the difference depends on the President's approval

[7] When possible, the estimation of both models used GDP data that was released by the Bureau of Economic Analysis in August of the election year, data that would have been available when making a real forecast, rather than later revised data.

numbers leading into the campaign, as well as the state of the economy, but the potential magnitude of the forecast differences between the two versions is evident in the difference in coefficients for the in-party term variable. All things being equal, whereas the original specification predicts a penalty of about 4 percentage points for the in-party candidate seeking more than a second term for his party ((51.121 − 5.154) − 50), the successor version of the model predicts a penalty of only about 2.7 percentage points ((50.807 − 3.483) − 50). In examining contingent predictions based on the historical range of the predictor variables, the two versions of the equation can yield predictions that differ by as much as 1.8 percentage points, though two-thirds of the contingent predictions differed by one percentage point or less. Still, in a close election, which 2008 may well be, a prediction difference of even one percentage point may be politically critical.

The 2008 experience will probably not settle the matters of whether retrospective voting is conditional or whether forecasting models should incorporate the differences between incumbent and open-seat elections. One case can only reveal so much. However, each additional election offers forecasters a learning experience, an opportunity to carefully assess and prudently improve their models.

References

Abramowitz, A. I. (2001). The time for change model and the 2000 election. *American Politics Quarterly*, *29*, 279−282.

Abramowitz, A. I. (2004). When good forecasts go bad: the time-for-change model and the 2004 presidential election. *PS: Political Science and Politics*, *37*, 745−746.

Bartels, L. M., & Zaller, J. (2001). Presidential vote models: a recount. *PS: Political Science and Politics*, *34*, 8−20.

Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The American Voter.* New York: Wiley.

Campbell, J. E. (2000). The science of forecasting presidential elections. In James E. Campbell, & James C. Garand (Eds.), *Before the Vote: Forecasting American National Elections* (pp. 169−187). Thousand Oaks, CA: Sage Publications.

Campbell, J. E. (2001). Taking stock of the forecasts of the 2000 presidential election. *American Politics Research*, *29*, 275−278.

Campbell, J. E. (2001). The referendum that didn't happen: the forecasts of the 2000 presidential election. *PS: Political Science and Politics*, *34*, 33−38.

Campbell, J. E. (2004). Introduction: the 2004 presidential election forecasts. *PS: Political Science and Politics*, *37*, 733−735.

Campbell, J. E. (2004). Forecasting the presidential vote in 2004: placing preference polls in context. *PS: Political Science and Politics*, *37*, 763−767.

Campbell, J. E. (2005). Introduction: assessments of the 2004 presidential vote forecasts. *PS: Political Science and Politics*, *38*, 23−24.

Campbell, J. E. (2008). *The American Campaign, Second Edition: U.S. Presidential Campaigns and the National Vote.* College Station, TX: Texas A&M University Press.

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

Colomer, J. M. (2007). What other sciences look like. *European Political Science*, *6*, 134−142.

Converse, P. E. (1966). The concept of the 'normal vote'. In Angus Campbell, Philip E. Converse, Warren E. Miller, & Donald E. Stokes (Eds.), *Elections and the Political Order* (pp. 9−39). New York: Wiley.

Converse, P. E. (1966). Information flow and the stability of partisan attitudes. In Angus Campbell, Philip E. Converse, Warren E. Miller, & Donald E. Stokes (Eds.), *Elections and the Political Order* (pp. 136−157). New York: Wiley.

Cuzan, A. G., & Bundrick, C. M. (2005). Deconstructing the 2004 presidential election forecasts: the fiscal model and the Campbell collection compared. *PS: Political Science and Politics*, *38*, 255−262.

Erikson, R. S. (1989). Economic conditions and the presidential vote. *American Political Science Review*, *83*, 567−573.

Eisenhower, K., & Nelson, P. (2000). The phony science of predicting elections. *Slate.* http://www.slate.com/id/83375/#sb83381.

Fair, R. C. (2002). *Predicting Presidential Elections and Other Things.* Palo Alto, CA: Stanford University Press.

Fair, R. C. (2004). A vote equation and the 2004 election. http://fairmodel.econ.yale.edu/vote2004/index2.htm.

Fiorina, M. P. (1981). *Retrospective Voting in American National Elections.* New Haven, Conn.: Yale University Press.

Gelman, A., & King, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, *23*, 409−451.

Greene, J. P. (2002). Forecasting follies. *The American Prospect*, *4,* http://www.prospect.org/cs/articles?article=forecasting_follies.

Holbrook, T. M. (1996). *Do Campaigns Matter?* Thousand Oaks, Calif.: Sage Publications.

Holbrook, T. M. (2001). Forecasting with mixed economic signals: a cautionary tale. *PS: Political Science and Politics*, *34*, 39−44.

Holbrook, T. M. (2004). Good news for Bush? Economic news, personal finances, and the 2004 presidential election. *PS: Political Science and Politics*, *37*, 759−761.

Johnston, R., Hagen, M. G., & Jamieson, K. H. (2004). *The 2000 Presidential Election and the Foundations of Party Politics*. New York: Cambridge University Press.

Key, V. O. (1966). *The Responsible Electorate.* New York: Vintage Books.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The People's Choice.* New York: Columbia University Press.

Lewis-Beck, M. S. (1985). Election forecasts in 1984: how accurate were they? *PS: Political Science and Politics*, *18*, 53−62.

Lewis-Beck, M. S. (1988). *Economics and Elections.* Ann Arbor: University of Michigan Press.

Lewis-Beck, M. S. (2005). Election forecasting: principles and practice. *British Journal of Politics and International Relations*, *7*, 145−164.

Lewis-Beck, M. S. (2006). Does economics still matter? Econometrics and the vote. *Journal of Politics*, *68*, 208−212.

Lewis-Beck, M. S., & Rice, T. W. (1992). *Forecasting Elections.* Washington, D.C.: CQ Press.

Lewis-Beck, M. S., & Stegmaier, M. (2000). Economic determinants of economic outcomes. *Annual Review of Political Science*, *3*, 183−219.

Lewis-Beck, M. S., & Tien, C. (2004). Jobs and the job of president: a forecast for 2004. *PS: Political Science and Politics*, *37*, 753−758.

Lewis-Beck, M. S., & Tien, C. (2007, June 25). Forecasting presidential elections: when to change the model? *International Symposium on Forecasting, New York City.*

Lichtman, A. J. (1996). *The Keys to the White House, 1996.* Lanham, Maryland: Madison Books.

Lippmann, W. (1925). *The Phantom Public.* New York: Harcourt, Brace.

Lockerbie, B. (2004). A look to the future: forecasting the 2004 presidential election. *PS: Political Science and Politics*, *37*, 741−743.

Mayer, W. G. (2003). Forecasting presidential nominations or, my model worked just fine, thank you. *PS: Political Science and Politics*, *36*, 153−157.

Miller, D. W. (2000, November 17). Election results leave political scientists defensive over forecasting models. *The Chronicle of Higher Education*. http://chronicle.com/weekly/v47/i12/12a02401.htm. Accessed: February 11, 2008.

Morton, R. B. (2006). *Analyzing Elections.* New York: W.W. Norton.

Nadeau, R., & Lewis-Beck, M. S. (2001). National economic voting in U.S. presidential elections. *Journal of Politics*, *63*, 159−181.

Norpoth, H. (2002). On a short-leash: term limits and the economic voter. In Han Dorussen, & Michael Taylor (Eds.), *Economic Voting* (pp. 121−136). Oxford: Routledge.

Norpoth, H. (2004). From primary to general election: a forecast of the presidential vote. *PS: Political Science and Politics*, *37*, 737−740.

Sidman, A. H., Mak, M., & Lebo, M. J. (2008). Model averaging and presidential election forecasts: Lessons learned from 2000 and 2004. *International Journal of Forecasting*, *24*, 237−256 (this volume). doi:10.1016/j.ijforecast.2008.03.003.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know It?* Princeton: Princeton University Press.

Tufte, E. R. (1978). *Political Control of the Economy.* Princeton, N.J.: Princeton University Press.

Whitten, G. D., & Palmer, H. D. (1999). Cross-national analyses of economic voting. *Electoral Studies*, *18*, 49−67.

Wlezien, C., & Erikson, R. S. (2004). The fundamentals, the polls, and the presidential vote. *PS: Political Science and Politics*, *37*, 747−751.