

Thin-Slice Forecasts of Gubernatorial Elections

Daniel J. Benjamin

Dartmouth College and Institute for Social Research

Jesse M. Shapiro*

University of Chicago and NBER

First Draft: June 28, 2005

This Draft: October 23, 2006

Abstract

We showed 10-second, silent video clips of unfamiliar gubernatorial debates to a group of experimental participants and asked them to predict the election outcomes. The participants' predictions explain more than 20 percent of the variation in the actual two-party vote share across the 58 elections in our study, and their importance survives a range of controls, including state fixed effects. In a horse race of alternative forecasting models, participants' visual forecasts significantly outperform economic variables in predicting vote shares, and are comparable in predictive power to a measure of incumbency status. Adding policy information to the video clips by turning on the sound tends, if anything, to worsen participants' accuracy, suggesting that naïveté may be an asset in some forecasting tasks.

JEL classification: D72, J45, Z19

Keywords: thin slices, charisma, elections

*We are deeply indebted to the Taubman Center for State and Local Government and to KNP Communications for financial support. We thank Alberto Alesina, Nalini Ambady, Chris Chabris, James Choi, Stefano DellaVigna, Ray Fair, Luis Garicano, Matt Gentzkow, Ed Glaeser, David Laibson, Steve Levitt, Ulrike Malmendier, Hal Movius, Kevin M. Murphy, Emily Oster, Jane Risen, Emmanuel Saez, Bruce Sacerdote, Andrei Shleifer, Matt Weinzerl, Rick Wilson, and seminar participants at the University of Chicago, Harvard University, the Stanford Institute for Theoretical Economics, and the University of Michigan for helpful comments. Benjamin thanks the Program on Negotiation at Harvard Law School; the Harvard University Economics Department; the Chiles Foundation; the Federal Reserve Bank of Boston; the Institute for Quantitative Social Science; Harvard's Center for Justice, Welfare, and Economics; the National Institute of Aging, through Grant Number T32-AG00186 to the National Bureau of Economic Research; the Institute for Humane Studies; and the National Science Foundation for financial support. We are very grateful to Sujie Chang, Jonathan Hall, Ethan Lieber, Dina Mishra, Marina Niessner, Krishna Rao, and David Sokoler for outstanding research assistance, Robert Jacobs for excellent programming, and John Neffinger for generous assistance with the second and third rounds of the study. E-mail: daniel.benjamin@gmail.com, jmshapir@uchicago.edu.

1 Introduction

From 1988 to 2002, the standard deviation in the two-party vote share in U.S. gubernatorial elections was 11 percentage points, and the interquartile range was from 42 percent to 55 percent in favor of the Democratic candidate. Most economic analyses of the predictors of election outcomes focus on the impact of economic conditions (Fair, 1978; Alesina, Roubini and Cohen, 1997; Wolfers, 2002) and political circumstances (Levitt, 1994; Lee, 2001). Yet, these factors typically leave much of the overall variation in vote shares unexplained. In addition to the intrinsic value of econometric forecasts (Fair, 1996), understanding the sources of the remaining variation is important if we believe, as much evidence suggests, that the identity of the officeholder matters for the policies undertaken (Jones and Olken, 2005; Lee, Moretti, and Butler, 2004; Fiorina, 1999; Glaeser, Ponzetto, and Shapiro, 2005).

In this paper, we test an election forecasting tool based on the predictions of naïve experimental participants. In our laboratory study, participants saw 10-second, silent video clips from televised debates in 58 unfamiliar gubernatorial elections from 1988-2002, and guessed the winner of each election. The use of short selections of video takes advantage of the fact that judgments about other people from “thin slices”—exposures to expressive behavior as brief as a few seconds—tend to be highly predictive of reactions to much longer exposures (Ambady and Rosenthal, 1992). This fact makes it possible to obtain reliable ratings from a large number of participants without requiring lengthy laboratory sessions.

We first use our measure to assess the quality of participants’ forecasts of elections. The share of participants predicting a Democratic victory is highly related to actual election outcomes, and can account for more than 20 percent of the variation in two-party vote shares in our sample of elections. This result survives a wide range of controls, including race, height, and state fixed effects. A range of tests also confirm that familiarity with the candidates or election outcomes does not explain our findings.

After demonstrating the predictive validity of our measure, we compare the predictive power of participants’ forecasts to that of the economic and political factors typically included in econometric models of election outcomes. We find that, as a forecasting tool, participants’ ratings outperform a range of models that relate economic circumstances in the state to election outcomes. Turning to

a comparison with political variables, we find that the predictive power of participants' ratings is comparable to a measure of the incumbency status of the candidates. A combination of campaign spending and incumbency status outperforms our measure, although our laboratory-based index alone achieves more than half of the predictive power of a carefully specified multivariate model in predicting the vote shares in our sample of 58 elections.

We turn next to the question of why participants are able to accurately guess election outcomes. We first show that inferences about policy positions do not seem to be driving participants' success in predicting outcomes. Participants performed poorly in guessing the party affiliations of the two candidates, and when we allowed participants to hear the sound associated with the video clips, their ability to guess political positions improved, but their ability to guess election outcomes *worsened*. We also find that participants' ratings of candidate likability, physical attractiveness, and leadership are only weakly related to election outcomes, suggesting that participants' forecasts of election outcomes capture information beyond these factors. Finally, we investigate whether participants can detect differences in candidates' confidence about winning the election. We show that participants' predictions have predictive power even in cases where the election was judged to be close or a "dead heat" as of the time of the debate, suggesting that participants can predict the outcome with reasonable success even when there is little reason for one candidate to be especially confident.

One possible explanation for the accuracy of participants' guesses is that their reactions measure candidates' charisma or personal appeal, and that these characteristics affect voter behavior directly. Though we cannot make such definitive causal claims based on our evidence, this mechanism is consistent with a growing literature on the returns to beauty in labor markets (Hamermesh and Biddle, 1994; Biddle and Hamermesh, 1998; Mobius and Rosenblat, 2006). It also has potentially important implications for political competition, for the selection of individuals into political careers, and for the sorting of candidates across jurisdictions.

Our findings are consistent with an existing literature on the role of physical appearance in elections. Rosenberg, Bohan, McCafferty and Harris (1986) study the effects of candidate attractiveness by constructing campaign flyers for a hypothetical election. Hamermesh (2005) studies the role of attractiveness in American Economic Association elections using students' ratings of still photographs. Analyzing a single election, the multi-candidate 1996 Romanian presidential race,

Schubert et al (1998) found that electability ratings based on still photographs and brief video clips correlated with first-round voting outcomes. In the paper most closely related to our own, Todorov et al (2005) independently show that ratings of competence based on photographs of Congressional candidates predict election outcomes and vote shares. More recently, Berggren, Jordahl and Poutvaara (2006) find that ratings of physical attractiveness outperform ratings of competence in predicting Finnish election outcomes.

We make several contributions relative to this existing literature. Most importantly, our regression framework permits us to assess the incremental predictive power of personal appeal, after accounting for economic and political predictors of electoral success and to compare the relative predictive power of these factors. In addition, by manipulating the presence of sound in video clips, our methodology allows us to separate the predictive power of personal appeal from the role of other factors, such as party affiliation. We also show that participants' predictions about election outcomes deliver a more accurate forecast of actual results than do their ratings of candidate characteristics such as likability, physical attractiveness, and good leadership, suggesting that participants are able to incorporate predictive information from factors beyond those commonly included in existing studies. Additionally, our use of video clips from candidate debates allows us to control for image quality, which may confound studies that use candidate-supplied photographs.

Our finding that adding sound to the video clips tends to worsen participants' accuracy relates to psychological evidence that verbal information can interfere with more instinctive visual judgments (e.g., Etcoff et al, 2000), and that individuals have difficulty ignoring irrelevant information (Camerer, Loewenstein, and Weber, 1989). It may also help to explain why the election forecasts of highly informed experts often perform no better than chance in predicting outcomes (Tetlock, 1999).

Finally, our evidence relates to the literature on economic and political predictors of election outcomes in general (Fair, 1978; Alesina and Rosenthal, 1995), and to the literature on the predictors of gubernatorial election outcomes in particular (Peltzman, 1987 and 1992; Adams and Kenny, 1989; Chubb, 1988; Levernier, 1992; Kone and Winters, 1993; Besley and Case, 1995; Leyden and Borrelli, 1995; Niemi, Stanley and Vogel, 1995; Partin, 1995; Lowry, Alt, and Ferree, 1998; Wolfers, 2002). We show that naïve participants' intuitive predictions perform comparably or better than many of the variables emphasized in the literature. Moreover, while we do not conclusively demon-

strate that factors such as candidate charisma have a causal effect on voter behavior, the findings we present constitute suggestive evidence of a role for such factors in gubernatorial politics.

The remainder of the paper proceeds as follows. Section 2 describes the procedures for our laboratory survey and for the collection of economic and political predictors of election outcomes. Section 3 presents our findings on the accuracy of participants' predictions of electoral outcomes, and Section 4 presents our estimates of the relative strength of economic, political, and personal factors in determining the outcomes of gubernatorial elections. Section 5 discusses evidence on the factors driving participants' ratings. Section 6 concludes.

2 Laboratory Procedures and Data

In order to measure participants' forecasts of election outcomes, we showed them 10-second video clips of major party gubernatorial candidates. Participants rated the personal attributes of the candidates, guessed their party affiliation, and predicted which of the two candidates in a race would win.

To study the effects of additional information, we included three (within-subject) experimental conditions. Most of the clips were silent. Some of the clips had full sound. Finally, some of the clips had "muddled" sound, so that participants could hear tone of voice and other nonverbal cues but not make out the spoken words. These clips were generated by "content-filtering" the audio files, removing the sound frequencies above 600 Hz, a common procedure in psychological research (e.g., Rogers, Scherer, and Rosenthal 1971, Ambady et al 2002). The audio tracks on the processed files sound as though the speaker has his hand over his mouth.

We used clips from C-SPAN DVDs of gubernatorial debates.¹ By taking both candidates' clips from the same debate, we ensured that stage, lighting, camera, and sound conditions were virtually identical for the two candidates in a given election. We used a total of 68 debates from 37 states, with 58 distinct elections. In elections with more than two candidates, we focused on the main

¹The C-SPAN DVDs are drawn from debates aired by C-SPAN during the gubernatorial election season. We attempted to use every available C-SPAN DVD so as to avoid selection bias in the sample of elections we studied. Conversations with Ben O'Connell at C-SPAN on July 25, 2006, suggest that the primary factors involved in C-SPAN's selection of gubernatorial debates are the compliance of local TV stations with re-airing, and the importance of the election. While this latter factor would lead one to expect that more competitive races from larger states are more likely to be included in the C-SPAN collection, in unreported regression models we find no evidence that debates from more competitive races are more likely to be included, and only weak evidence that debates from larger states are more likely to be in our sample.

Democrat and the main Republican in the race.

2.1 Participants

Participants were 264 undergraduates (virtually all Harvard students), recruited through on-campus posters and e-mail solicitations. We promised students \$14 for participating in a one-hour experiment on “political prediction,” with the possibility to earn more “if you can correctly predict who won the election.” We held 11 sessions in a computer classroom during 3-4pm on May 7, 9, 10, 12, and 13, 2005; during 2-3pm on January 8, 9, 10, and 11, 2006; and during 2-3pm on March 2 and 4, 2006. We mailed checks to participants within a week of their participation.

2.2 Materials

The clips were generated by drawing random 10-second intervals of the debates during which the camera focused only on one of the two major candidates. We dropped clips in which the candidate’s name or party appeared, or in which the candidate stated his own or his opponent’s name or party. For each candidate in each debate, we used three clips, the first three clips that we did not drop. The computer randomly selected one of these three clips for a participant to see. For each of these three clips, we created a muddled version and a silent version by modifying the audio content.

Participants in different sessions viewed different numbers of each kind of clips. Because of initial concerns that the silent and muddled clips would be boring to watch, each participant in the first session saw 15 elections with full sound, 3 with muddled sound, and 3 with no sound. Informal interviews with participants after the session indicated that the concerns were unwarranted. In the subsequent sessions in May, each participant saw 7 elections each with full sound, muddled sound, and no sound. In the January and March 2006 sessions, participants saw 21 elections, all of them without sound.²

The informal interviews also suggested that, after watching a number of elections, some participants had difficulty recalling which candidate was which when answering the questionnaire. To address this issue, we created still shots of each candidate by taking the first frame of each clip. From the second (May 9) session onward, the computer displayed the relevant still shot while par-

²Statistical tests show no difference in participants’ ability to forecast election outcomes across the three rounds of sessions.

ticipants filled out their judgments of each candidate. The computer showed the shots of both “Candidate A” and “Candidate B” when participants made comparative judgments about them.

2.3 Procedure

Instructions were displayed on each participant’s computer screen, and an experimenter read them aloud. The instructions explained that each participant would watch 21 pairs of 10-second video clips of candidates for governor. Each clip in a pair would show one of the two major candidates: one Democrat, one Republican. After each clip, the participant would rate the candidate on several characteristics, and after every pair of clips, the participant would compare the two candidates. Participants were told that they would be asked which candidate in each pair was the Democrat. To encourage accurate guessing, one of the elections would be selected randomly, and the participant would earn an extra \$1 for guessing correctly in that election. Similarly, participants would be asked which candidate had won the actual election and would be paid an additional \$1 for guessing correctly in a randomly chosen election.

We asked participants whether they had grown up in the U.S. and in which ZIP code. We did not show any clips from an election in the state where a participant grew up. We also asked participants after each clip whether they recognized the candidate and, if so, who it is. We dropped from the analysis a participant’s ratings of candidates from any election in which the participant claimed to recognize one of the candidates (although we still paid participants for accurate guesses about victory and party identity in these cases). Because essentially all participants were Massachusetts residents at the time of the study, we also excluded from our analysis any Massachusetts elections.³

In the May 2005 sessions, participants knew that they would watch some of the clips with full sound, some with muddled sound, and some without sound. During the instructions, participants listened to two versions of a sample soundtrack, one with full sound and one with muddled sound. In the January and March 2006 sessions, participants knew that all of the clips would be silent.

After each clip, participants were asked to rate, on a 4-point scale, how much the candidate in the clip seemed “physically attractive,” “likeable,” “a good leader,” and “liberal or conservative.” After each pair of clips, participants answered “A” or “B” to each of the following questions:

³Participants whose home state was not Massachusetts did sometimes see clips from Massachusetts elections, but the data from these clips were excluded from our analysis.

- In which clip did you like the speaker more?
- One of these candidates is a Democrat, and one is a Republican. Which one do you think is the *Democrat*?
- Who would you vote for in an election in your home state?

If you do not live in the U.S., please answer this question as best you can for Massachusetts.

- Who do you think actually won this election for governor?

After all the clips were finished, we asked participants to rate (on a 4-point scale) how liberal/conservative they considered themselves, which political party they identified with more strongly, and how interested they are in politics. We also asked whether they had voted in the 2004 presidential election or, if ineligible, whether they would have. Finally, we asked a few demographic questions (college major, year in school, gender, mother's and father's education, and standardized test scores).

In sessions four and five, we asked a few debriefing questions at the very end of the questionnaire. We asked, on a scale from 1 to 10,

- When you watched video clips with full sound [video clips with muddled sound / silent video clips], how confident were you (on average) in your prediction about who actually won the election?
- With full sound [muddled sound / silent clips], how confident were you about which candidate was the Democrat?

We asked these two questions for each of the three sound conditions. We also asked participants about their strategies for making predictions for full sound and silent clips.

2.4 Measuring the Economic and Political Predictors of Election Outcomes

We collected data on the candidates and outcomes of the gubernatorial elections in our sample from the *CQ Voting and Elections Collection* (2005). We also obtained data on a number of political and economic predictors of election outcomes. Because of the modest size of our sample,

we tried to choose the political and economic predictors of election outcomes that seem to occur most frequently and robustly in the empirical literature on explaining and forecasting vote shares.

We obtained the following variables in order to construct possible economic predictors of election outcomes:

- *Per capita income.* Nearly all studies of the economic determinants of gubernatorial elections include some measure of recent growth in per capita income (see Peltzman, 1987; Chubb, 1988; Adams and Kenny, 1989; Levernier, 1992; Kone and Winters, 1993; Besley and Case, 1995; Partin, 1995; Niemi, Stanley and Vogel, 1995; Lowry, Alt, and Ferree, 1998; Wolfers, 2002). We obtained annual data on state per capita income from the Bureau of Economic Analysis (<http://www.bea.gov/bean/regional/data.htm>). Because a number of authors have argued that voters measure state economic performance relative to national trends, we have also computed national personal income as the average of state personal income, weighted by state populations as of 1995 (roughly the midpoint of our sample).⁴
- *Unemployment rate.* The unemployment rate is also frequently used as an index of state economic conditions in models of gubernatorial voting, although it appears less often in the literature than does income per capita (see Levernier, 1992; Kone and Winters, 1993; Besley and Case, 1995; Leyden and Borrelli, 1995; Wolfers, 2002). We obtained annual data on state unemployment rates from the Bureau of Labor Statistics (<http://www.bls.gov/data/>). As with income, we compute a national unemployment measure as the weighted average of the state unemployment measures.
- *Per capita revenues.* A number of authors have argued that voters respond to state fiscal policy (Peltzman, 1992; Kone and Winters, 1993; Niemi, Stanley and Vogel, 1995), and to differences between a state's fiscal policy and that of neighboring states (Besley and Case, 1995). Following Peltzman (1992), we use state per capita revenues as our primary measure of fiscal policy. We obtained information on state revenues per capita from the Census at <http://www.census.gov/govs/www/state.html>. In appendix A, we demonstrate that our

⁴We obtained data on state population in 1995 from the U.S. Census at <http://www.census.gov/population/projections/state/stpjpop.txt>. We use the average across U.S. states rather than reported national figures to ensure that the scale and definition of the variable is comparable between the state and national indices.

results are robust to using a tax-simulation-based measure as in Besley and Case (1995).

We also obtained data on the following political predictors of election outcomes:

- *Incumbency status.* Lee (2001) shows, using a regression-discontinuity design, that incumbency has a significant causal impact on vote shares in congressional elections. We identified the incumbent in each race (if any) using the *CQ Voting and Elections Collection* (2005). Our measure of incumbency is an index equal to 1 when the Democrat is an incumbent, 0 when neither candidate is an incumbent, and -1 when the Republican is an incumbent.⁵
- *Campaign spending.* Levitt (1994) argues that campaign spending has a statistically reliable impact on election outcomes in congressional contests. To measure campaign spending, we use data from Jensen and Beyle’s (2003) updated Gubernatorial Campaign Finance Data Project. This database provides information on the total campaign expenditures of each major party candidate. Our primary measure of campaign spending is the difference in log(expenditure) between the Democrat and Republican.⁶ In the six elections for which we lack spending information for one or both candidates, we impute this variable at the state mean difference in log spending over the 1988-2003 period.
- *Historical vote shares.* It is common to include measures of historical election outcomes in regression models of gubernatorial contests, as a proxy for party strength. We compute a measure of the average share of the two-party vote received by Democrats in the 1972-1987 gubernatorial elections in the state. We use this time period because it precedes all of the elections in our experimental sample.

3 Participants’ Success in Predicting Electoral Outcomes

Participants in our study performed extremely well in predicting the outcomes of the electoral contests that the video clips portrayed. Across our 58 elections, an average of 58 percent of

⁵Unreported regressions indicate that Democratic and Republican incumbency have similar effects on the two-party vote share, so that allowing for greater flexibility does not significantly increase the predictive power of the incumbency status variable.

⁶As with incumbency status, we do not find substantial asymmetries in the effects of campaign spending between Democratic and Republican candidates, so we do not lose much predictive power from constructing this spending index.

participants correctly guessed the winner of the election. With a standard error of around 2 percent, a t -test can definitively reject the null hypothesis that participants performed no better than chance (50 percent accuracy) in forecasting the election outcomes ($p = 0.002$).

Participants' ratings are also very highly correlated with actual vote shares across elections. In figure 1, we graph the actual two-party vote shares in our sample of 58 elections against the share of study participants who predicted that the Democrat would win the election. There is a visually striking positive relationship between these two measures, and the correlation coefficient is a highly statistically significant 0.46 ($p < 0.001$). Moreover, the relationship does not appear to be driven by outliers: the Spearman rank-correlation coefficient between participants' predictions and actual vote shares is large (0.42) and strongly statistically significant ($p = 0.001$).

A regression approach reveals similar patterns. Column (1) of table 1 shows that an increase of one percentage point in the share predicting a Democratic victory is associated with an increase of about one-quarter of a percentage point in the actual two-party vote share of the Democratic candidate. This relationship is highly statistically significant, and the predictions of our laboratory participants account for over one-fifth of the overall variation in two-party vote shares across the elections in this sample. We will provide more discussion of the relative power of alternative forecasting models in section 4, but to give a sense of magnitudes the R^2 of our laboratory-generated predictor is only slightly lower than we would obtain using as a predictor a measure of the incumbency status of the candidates.

Indeed, there is reason to expect that the R^2 in column (1) might understate the true explanatory power of visual forecasts, because of sampling error in participants' ratings. To check that this bias does indeed weaken our findings, in column (2) of table 1 we present results for the sample of elections for which we have over 30 raters, where econometric theory would suggest a fairly limited bias from measurement error. As expected, both the coefficient and R^2 of the model increase in this case, with the coefficient changing by about 10 percent. We have also estimated a maximum likelihood model (results not shown) that explicitly models the sampling error in our measure. In that model, we find a coefficient of about 0.27 on participants' ratings, with an R^2 of about 26 percent. Although these models indicate that our estimates of participants' predictive power are attenuated, throughout the body of the paper we will conservatively treat our sample-based measures as though they were not subject to sampling variability.

The remaining columns of table 1 discuss a variety of robustness checks. In column (3), we restrict to cases in which both candidates are white males (about two-thirds of our sample), in order to test whether participants' accuracy results merely from race or gender cues.⁷ We find that the coefficient and R^2 on this restricted sample are comparable to those in the overall sample. Similarly, in column (4), we include a control for whether the Democrat appears to be the taller candidate, as judged from footage on the original debate DVDs (e.g., handshakes) not shown to participants.⁸ (The clips we showed to participants show only the head and torso of one candidate at a time, so it is unlikely that participants could judge relative height from the clips.) We find that height exerts a positive, but small and statistically insignificant, effect on vote shares, and that including this variable makes little difference for our estimate of the predictive power of participants' ratings.

In column (5) of table 1, we use data from the 17 states with multiple elections in our sample to test how well participants do at predicting differences across elections *within* a state. Despite the reduction in precision that results from using a small share of the variation in the data, we still identify a large and statistically significant relationship between participants' ratings and the actual two-party vote share after including state fixed effects. The coefficient in this regression is, if anything, somewhat larger than the coefficient in the cross-sectional regression in column (1).⁹ Figure 2 illustrates the strong within-state relationship between participants' ratings and actual vote shares.

A final potential issue with interpreting our results as evidence of the forecasting power of participants' ratings is the possibility that, despite our efforts to exclude raters familiar with a candidate from our analysis, some informed raters remained in the sample. A first piece of evidence against this view is that, as we document further in section 5.1 below, participants in our sample (who claimed to be unfamiliar with the candidates) were unable to do better than random guessing

⁷We coded these cases conservatively, including only those debates in which it was obvious from the video clips themselves that both candidates were white males.

⁸The coding of heights was done from shots showing both candidates by a research assistant who did not know the outcomes of the sample elections or the share of participants predicting a Democratic victory.

⁹Related to the issue of cross-state variation in party strength, there is also the possibility that participants' responses are only effective in predicting extreme landslides. To check this issue, we have estimated a model that restricts attention to elections in which no major party candidate received more than 60 percent of the two-party vote (about two-thirds of the sample of elections). In this case, the coefficient drops somewhat, but the R^2 remains essentially the same as in the baseline model, increasing slightly from 0.22 to 0.23.

in identifying the party affiliations of the candidates.¹⁰ A second piece of evidence is that the recognizability of candidates in an election is not related to participants' accuracy. More specifically, participants who claimed not to recognize a candidate were no better at forecasting elections in which large numbers of *other* participants claimed to recognize one or more of the candidates. If the likelihood of recognizing a candidate is correlated across individuals within an election (which our data suggest it is), then this test suggests that even unconscious familiarity is unlikely to confound our estimates.

4 Comparisons with Economic and Political Predictors

In this section, we compare the accuracy of forecasts based on participants' predictions with political and economic factors frequently used in election forecasting. The forecasting value of participants' ratings survives controlling for these factors. Overall, we find that the performance of our measure is far better than economic factors, and comparable to some important political factors, in predicting vote shares in gubernatorial contests.

4.1 Economic Predictors of Election Outcomes

Table 2 shows our estimates of the forecasting power of alternative sets of economic variables. For each variable, we compute one-year growth rates, following a common practice in the literature on economic predictors of gubernatorial election outcomes. We then create an index equal to the growth rate of the variable if the incumbent governor is a Democrat, and equal to the negative of the growth rate if the incumbent governor is a Republican.¹¹ This specification amounts to assuming that the incumbent party is held responsible for the prevailing economic conditions at the time of the election, consistent with Fair (1978).

In addition to computing the R^2 for each specification shown, we have also computed an out-

¹⁰This is so despite the fact that participants who told us that they had recognized one or more candidates in a debate performed far better than chance in identifying political parties (results not shown). Thus, if our sample of non-recognizers were contaminated, we would expect to see better-than-random matching of party affiliations, which we do not. Moreover, participants were paid for correctly identifying the parties of the candidates, so they would have had a financial incentive to give the correct answer if they knew it.

¹¹We obtained the party of the current governor from <http://en.wikipedia.org/wiki/List_of_United_States_Governors>. If the incumbent governor was neither a Democrat nor a Republican, we coded the relevant index as having a value of zero.

of-sample measure of the fit of each model.¹² In particular, we compute the out-of-sample mean squared error by estimating the model repeatedly, leaving out a different observation each time, and computing the squared error of the predicted value for the omitted observation. We then compare the mean squared error of the model to that of a model including only a constant term. Finally, we compute an out-of-sample R^2 as the percentage reduction in mean squared error attributable to the inclusion of the explanatory variable. This statistic gives us an estimate of how well the model performs in explaining the variance of observations *not* used to fit the model. Unlike the traditional R^2 (but similar to the adjusted R^2), the out-of-sample R^2 can decrease as more variables are added to a model, if these variables do not achieve significant increases in goodness-of-fit.

For reference, the R^2 of a model using the share of experimental participants predicting a Democratic victory to predict the Democrat's two-party vote share is approximately 22 percent, and the out-of-sample R^2 is about 19 percent. This indicates that our experimental measure can reliably predict about one-fifth of the overall variation in two-party vote shares, even when we use the model to predict observations not included in the estimation.

In column (1) of table 2, we present estimates of a model that predicts election outcomes using the one-year growth in log(state personal income) prior to the election year. As expected, higher income growth is associated with greater electoral success, and the effect is both economically nontrivial and marginally statistically significant. However, this specification has an R^2 of less than six percent, with an out-of-sample R^2 of around two percent. This out-of-sample R^2 estimate is consistent with Wolfers' (2002) finding of a one to three percent adjusted R^2 for economic variables in explaining incumbent governors' electoral performance. On the whole, then, our estimates in column (1) suggest that income growth does predict election outcomes, but that its forecasting power is weaker than that of participants' ratings.

In the second panel of column (1), we show what happens to our estimate of the predictive power of participants' ratings, once we control for growth in state personal income. Not surprisingly, we find that inclusion of the economic variable leaves the magnitude and statistical significance of the coefficient on participants' ratings essentially unchanged. We also show the incremental out-of-sample R^2 of participants' ratings; that is, the change in out-of-sample R^2 from including

¹²See Goyal and Welch (forthcoming) for a recent discussion of the differences between in-sample and out-of-sample forecasting evaluations.

participants' ratings in the economic forecasting model. This calculation indicates an improvement of nearly 20 percentage points in the out-of-sample forecasting power of the model. These findings provide further evidence of the robustness of participants' ratings as an election forecaster, even after conditioning on economic factors such as income growth.

In column (2) of table 2, we augment the specification of column (1) by adding a measure of the one-year change in the unemployment rate. This variable enters negatively as expected, and its inclusion diminishes our estimate of the importance of income growth. However, the gain in R^2 is only two percentage points, resulting in an overall R^2 of about 8 percent. Moreover, because the additional variable does not result in a great improvement in predictive power, the out-of-sample R^2 measure penalizes the specification heavily, resulting in a tiny negative out-of-sample R^2 , that is essentially zero. In other words, adding the change in unemployment to the model tends to reduce its out-of-sample performance. As the second part of column (2) shows, including the unemployment rate growth measure does not meaningfully affect the magnitude or statistical significance of the coefficient on participants' ratings.

A number of authors (e.g., Adams and Kenny, 1989; Lowry, Alt, and Ferry, 1998) have hypothesized that voters judge states' economic performance relative to the performance of the national economy. In column (3) of table 2, we implement a model of this type, regressing vote shares on the one-year growth in national personal income as well as the difference between state and national income growth. Consistent with the "benchmarking" hypothesis, we do find a positive relationship between state performance net of national performance and vote shares, although the coefficient is small and statistically insignificant. Consistent with Wolfers' (2002) finding that voters are sensitive to economic factors beyond the control of governors, we find a positive and statistically significant effect of national income growth on the two-party vote share. However, the R^2 of the model is only seven percent, and the inclusion of the statistically insignificant measure of state growth relative to national growth results in a *negative* out-of-sample R^2 of about seven percent. Thus, although our point estimates in this model are consistent with theoretical predictions, the model's predictive performance is relatively low. Additionally, inclusion of these variables does not affect the economic or statistical significance of our measure of participants' ratings, and including this measure greatly improves the forecasting power of the model.

Besley and Case (1995) posited that voters judge states' economic policies relative to those of

their geographic neighbors. In column (4) of table 2, we implement this hypothesis as a predictive model, using state revenues per capita as a measure of fiscal policy (Peltzman, 1992). In addition to a measure of a state’s own policy, we include an analogous measure of the mean policy of other states in the same Census division. Consistent with the yardstick competition model, we find that states are penalized for extracting more revenues, but that, for a given level of the growth in state revenues, states are rewarded for being in a Census division with greater growth in revenues. In other words, voters seem to reward a political party for keeping revenues low while neighboring states’ revenues are rising. Although the signs and magnitudes of the coefficients are broadly consistent with the yardstick competition model, these two variables explain only about six percent of the overall variation in vote shares, and have an out-of-sample R^2 of less than one percent. Moreover, their inclusion does not diminish the importance of participants’ ratings, and if anything leads to a slightly larger coefficient on the share of participants predicting a Democrat victory. Thus, while we do find support for the yardstick competition theory, its power as a purely predictive model appears to be low relative to the personal factors we measure in our experiment.

The models in table 2 consistently confirm the qualitative predictions of previous researchers regarding effects of economic variables on election outcomes. However, these economic and policy variables in general explain a small portion of the variation in vote shares, and perform poorly relative to our experimental ratings in predicting election results out of sample. Of course, the specifications in table 2 do not exhaust the list of possible economic models of elections. In appendix A, we review a much larger list of possible models. None of the models we explore has an out-of-sample R^2 above 10 percent, and in no case does the inclusion of a set of economic predictors significantly reduce the estimated importance of personal factors in predicting vote shares.

4.2 Political Predictors of Election Outcomes

Table 3 presents a series of regression models that use political variables to predict the Democrat’s share of the two party vote. In column (1) of table 3, we attempt to predict vote shares using a historical mean of the Democrat’s share of the two-party vote. This variable has a small, statistically insignificant coefficient, an R^2 of essentially zero, and a negative out-of-sample R^2 .¹³ This finding is

¹³To check whether the weak performance of this variable is due to our use of a historical, rather than a recent lag, we have re-estimated this model using the Democrat’s share of the vote in the most recent prior election (results not shown). The finding that past vote shares do not robustly predict current vote shares is also true for this alternative

not surprising, since party ideology tends to be more fluid at the state level, meaning that political parties often have difficulty maintaining long-term holds on governor's offices, even in states that have reliable positions in national politics. As the second panel of the table shows, including the historical election variable does not diminish our estimate of the importance of personal appeal as an election forecaster.

In column (2) of table 3, we predict vote shares using an index of the incumbency status of the candidates. We estimate that being an incumbent results in roughly a 7 percentage point electoral advantage, which is quite similar to Lee's (2001) discontinuity-based estimate of the effect of incumbency in congressional elections. This variable has an out-of-sample R^2 of about 23 percent, which indicates that the incumbency index is slightly better than participants' ratings in predicting vote shares. However, as the second panel of the table shows, including a measure of incumbency status does not eliminate the statistical importance of our measure of personal appeal, although it does reduce the estimated coefficient somewhat.

In column (3), we predict vote shares using a measure of the difference in the log of campaign spending between the two candidates. This specification must be taken with special caution, since a number of authors have argued that cross-sectional estimates of the effect of campaign spending suffer from significant endogeneity bias (e.g., Levitt, 1994; Gerber, 1998). With that caveat in mind, we find that an increase of one point in the difference in log spending is associated with an increase of about six percentage points in favor of the Democratic candidate, which is comparable to Gerber's (1998) instrumental variables estimate for Senate candidates but far larger than Levitt's (1994) fixed-effects estimate for congressional candidates. This variable has an out-of-sample R^2 of about 33 percent, which is larger than the fit from laboratory ratings alone. In the second panel of the table, we report that including the difference in campaign spending reduces the estimated coefficient on participants' ratings, but this variable remains statistically significant.

In column (4) we include all three political variables simultaneously. This model has an out-of-sample R^2 of about 36 percent, an improvement of about 16 percentage points over the model with laboratory ratings alone, but only marginally better than a prediction based only on differences in campaign spending. Including all of these measures diminishes the coefficient on participants' predictions somewhat, but the laboratory measure is still marginally statistically significant. More-

specification.

over, although the incremental out-of-sample R^2 of the laboratory measure is only two percent in this case, when we restrict attention to elections with over 30 laboratory raters, the incremental out-of-sample R^2 rises to nearly 7 percent, suggesting that measurement error may be attenuating the forecasting power of the laboratory measure.

On the whole, then, the political predictors we examine perform better than the economic predictors, and are either comparable to or somewhat better than our laboratory measure in predicting election outcomes. The variable that most closely approximates the predictive power of our laboratory measure is an index of incumbency status, suggesting that participants' ratings are comparable to incumbency status as a predictor of gubernatorial election outcomes.

We also find that including a range of political variables diminishes, but does not eliminate, the predictive power of participants' ratings. From the point of view of evaluating the causal determinants of election outcomes, it is unclear how to interpret this finding. If incumbency and campaign spending can be taken as exogenous, then the evidence in table 3 would indicate that the personal characteristics of candidates as measured by our laboratory-based index have a smaller effect on election outcomes once the effects of incumbency and campaign spending are taken into account. On the other hand, if (as seems likely) both incumbency and campaign spending are endogenous to a candidate's characteristics, then these regressions might indicate that part of the effect of charisma or other personal characteristics on election outcomes might operate through candidates' past election success or superior campaign financing.

5 Explaining the Accuracy of Participants' Predictions

Having established that participants' election forecasts are highly predictive of actual vote shares, we turn in this section to an exploration of the factors that influence participants' ratings. We show that policy inferences do not play an important role in explaining the accuracy of participants' forecasts, and that if anything adding policy information to the video clips (by turning on the sound) seems to worsen forecast accuracy. Next, we show that participants' own preferences over gubernatorial candidates are only weakly predictive of electoral success. Finally, we discuss the possibility that participants' accuracy results from their ability to detect candidates' own confidence in their prospects for victory, and report that participants' ratings are predictive even of the

outcomes of elections that were deemed to be close at the time of the debate. Taken together, these findings are consistent with (but do not definitively establish) the hypothesis that participants were detecting aspects of candidates' appearance and non-verbal behavior that could play a causal role in the election outcome.

5.1 Policy Inferences

If participants are able to infer candidates' policy positions from the video clips they saw, this could potentially contribute to their ability to forecast electoral success. Some simple calculations suggest that policy information is not likely to be an important component of participants' prediction process. Across the 58 elections in our study, an average of 53 percent of participants (with a standard error of 2 percent) were correctly able to identify which candidate is the Democrat in the contest after seeing the silent video clips. This average is statistically indistinguishable from random guessing ($p = 0.176$).

We conducted an experiment to study how additional policy information affects participants' ability to forecast election outcomes. In our first (May 2005) round of laboratory exercises, we randomly assigned one-third of each participants' elections to be silent, one-third to include the sound from the original debate, and one-third to be "muddled" so that the pitch and tone of the speakers' voice was audible but the words were unintelligible.

As we expected, adding sound to the video clips greatly improved participants' accuracy in guessing the identity of the Democratic candidate. Part A of figure 3 shows that participants rating elections with sound correctly identified the Democratic candidate 58 percent of the time, which is highly statistically distinguishable from random guessing ($p = 0.008$). By contrast, participants rating elections in the silent and muddled conditions correctly identified the Democrat only 52 and 48 percent of the time, respectively, neither of which can be distinguished statistically from random guessing (silent: $p = 0.540$; muddled: $p = 0.668$). Additionally, although the mean share correctly identifying the Democrat in the silent and muddled conditions cannot be distinguished statistically ($p = 0.237$), the mean share in the silent condition is marginally statistically different from that in the full sound condition ($p = 0.055$), and the mean share in the muddled condition is highly statistically different from that in the full sound condition ($p = 0.002$).

The fact that only full sound—and not muddled sound—improves the accuracy of party identifi-

cation shows that the improvement in accuracy is not a result of information in the pitch or tone of the candidates' voices. Rather, it is the content of their speech that provides relevant information on their policy positions.

Part A of the figure also shows that participants were more confident in their guesses of the candidates' political affiliations in the full sound condition than in the muddled condition, and more confident in their guesses in the muddled condition than in the silent condition. (These contrasts are all highly statistically significant, with p -values below 0.001.) Although participants were wrong to express greater confidence in their predictions in the muddled condition than in the silent condition, they were correct in thinking they had performed better in identifying the Democratic candidate in the full sound condition than in the silent condition.

The results are very different when we turn to participants' guesses about the outcome of the election, where the addition of sound to the video clips tended to *worsen* predictive accuracy. Participants rating elections in the silent and muddled conditions correctly identified the winner of the contest 57 percent of the time, but those rating clips with sound guessed correctly only 53 percent of the time. Although the differences among these conditions are not statistically significant,¹⁴ they contrast strongly with participants' reported confidence in their guesses, which indicates much greater confidence in the full sound condition than in the silent and muddled conditions. (The contrasts among the self-reported confidence indices are all highly statistically significant with p -values below 0.001.) Additionally, the fact that performance in the muddled condition is so similar to that in the silent condition suggests that it is the content, rather than the tone or pitch, of the candidates' speech that leads to the difference between the full sound and silent conditions.¹⁵

The finding that additional policy information worsens judgment has methodological and substantive implications beyond our study. First, from the perspective election forecasting, it suggests the critical importance of finding naïve raters who are unfamiliar with the candidates in question, and of not informing the raters of the policy positions of the respective candidates. Second, it may help to explain why expert forecasters, who are highly informed about and attentive to policy

¹⁴This calculation of statistical significance refers to aggregate, election-level comparisons of the overall rates of successful predictions across different conditions. When we take advantage of our within-subject design, by estimating a regression model of the probability of a successful prediction as a function of rater and debate fixed effects as well as dummies for the experimental condition, we find a marginally statistically significant difference between the full sound and silent conditions ($p = 0.078$) and no difference between the silent and muddled conditions.

¹⁵Informal conversations with participants in our study suggest that they did in fact believe that the verbal content contained important information for determining the election winner.

matters, are often found to perform no better than chance in predicting elections (Tetlock, 1999).

5.2 Ratings of Candidate Characteristics

In addition to asking participants to judge how actual voters would respond to the candidates, we asked them several questions about their own personal feelings about the candidates. We requested ratings (on a 1-4 scale) of whether each candidate was physically attractive, likeable, and a good leader. We also asked each participant to tell us how she would vote in a contest between the two candidates depicted in the video clips.¹⁶ By comparing the predictive power of these ratings to the predictive power of participants' guesses about election outcomes, we can assess whether participants' guesses incorporate information beyond that contained in the factors they rated, and relatedly, whether their guesses merely reflect their own, personal response to the candidates.

Table 4 shows that participants' accuracy in forecasting electoral outcomes does not result from merely reporting their own personal feelings about the candidates. Column (1) of the table shows the two-way correlations between the Democrat's share of the two-party vote and various participant ratings. As the column shows, only the share of participants predicting the Democrat to win is reliably correlated with the election outcome. The share of participants who said that they would vote for the Democrat is essentially uncorrelated with the election outcome, and the differences in the ratings of the two candidates are only modestly (and mostly statistically insignificantly) correlated with the actual vote share. Only the rating of whether the candidate seems like a good leader is marginally statistically significantly correlated with the actual vote share, with a correlation of around 0.23.

Comparing columns (2) and (3) reveals that the ratings of candidate characteristics are much more highly correlated with participants' own reported voting preferences than with their guesses about the preferences of actual voters. This is consistent with the fact that these ratings are only weakly related to actual election outcomes. Indeed, the one significant exception is the rating of leadership quality, which is highly correlated both with the share who report that they would vote for the Democrat and with the share predicting the Democrat to win (and, as shown in column (1), moderately correlated with the vote share in the actual election).

¹⁶We also asked the participant which candidate she liked more. Across elections, the share of participants who report liking the Democrat more is very highly correlated with the share who say they would vote for the Democrat, and has similar statistical properties to the latter variable.

A closer study of the determinants of participants' stated voting intentions is revealing about the relative weakness of this variable in predicting actual vote shares. Among debates rated by participants who described themselves as Democrats, 81 percent said they would vote for the candidate they thought was the Democrat, and only 20 percent said they would vote for the candidate they thought was the Republican. Similar asymmetries are present for participants who described themselves as Republicans. By contrast, Democrats and Republicans did not show such own-party biases when asked to predict the *winner* of the election. These findings suggest that participants suppressed their own leanings in forecasting election outcomes, but, as we would expect, allowed these preferences to affect their stated voting intentions. Because participants' individual political beliefs are not those of the state's median voter, and because participants' guesses about candidates' party affiliations were no better than chance, suppression of their own-party biases led to superior performance in forecasting vote shares.¹⁷ This finding suggests that participants were responding to information beyond their own personal tastes when guessing the election outcomes. It also implies that the accuracy of naïve raters is likely to be greater when they are asked to judge likely outcomes directly, rather than to gauge their own personal reactions to the candidates.

5.3 Candidate Confidence

One possible mechanism for our findings is that candidates who are likely to win an election appear more confident. If our participants detected differences in the confidence of the two candidates and used these differences to gauge the likelihood of victory for each candidate, then this mechanism could potentially explain the predictive power of the participants' judgments. If this mechanism were operating, this would not diminish the forecasting value of participants' judgments, but it would call into question any causal interpretation of models such as those in table 1.

One way to test for this possibility would be to restrict attention to cases where the election was considered to be close as of the time of the debate, so that candidates should be about equally confident of winning. To do this, we searched the Lexis-Nexis and ProQuest news databases, and obtained news coverage from the period before the debate. We code an election as close if either the news article itself identifies it as close or "a dead heat," or if the polling numbers are within two

¹⁷Interestingly, we find little evidence for individual differences in predictive accuracy: the accuracy of respondents' guesses is not correlated with gender, SAT scores, interest in politics, or political preferences. Moreover, random effects models indicate little individual-specific variation in predictive accuracy.

standard errors of equality. When we were unable to find news coverage on an election, or when the coverage did not discuss the likely outcome of the race, we did not code it as close. Among the 26 close races in our sample, participants' ratings are positively but not statistically significantly related to the election outcome. However, when we restrict attention to the 22 cases in which we have more than 30 raters for the election, we find a statistically significant relationship between the Democrat's share of the two-party vote and the share of our participants predicting the Democrat to win ($p = 0.016$), and an R^2 that is comparable to those we report in table 1. A maximum-likelihood model that adjusts explicitly for measurement error confirms a statistically significant relationship between participants' ratings and the election outcomes. Although these findings do not conclusively rule out candidate confidence as the mechanism behind our findings, they do suggest that participants' ratings may have predictive power even in cases in which candidates are fairly unsure of their prospects for victory.

6 Conclusions

In this paper, we show that naïve participants can accurately predict election outcomes based on short selections of video. The predictive power of participants' ratings survives controls for candidate, race, gender, and height, as well as for state fixed effects. Moreover, participants' ratings outperform a range of models that attempt to predict election outcomes based on economic circumstances. Models based on political characteristics such as incumbency perform as well or better than participants' ratings, but including participants' ratings does tend to improve the predictive power even of these factors. These findings suggest that the intuitive judgments of naïve raters may provide valuable information for forecasting election results.

Because candidate characteristics may be endogenous to the political circumstances of the state, our findings are best thought of as providing evidence on visual information as a forecasting tool, rather than as a causal factor in determining election outcomes. However, if a causal interpretation were appropriate, this would raise the interesting question of why all candidates for high office are not immensely appealing along the dimensions we measure. We note, however, that the attributes we measure may bring significant returns in the private labor market (e.g., Biddle and Hamermesh, 1998) as well as in the political sphere. Moreover, although high political office may be a desirable

position, political parties often offer candidacy to high office as a reward for loyal service in lower, less desirable offices. Hence for a highly appealing individual, the expected return to a political career may not be that great relative to other occupations.

The view that personal appeal yields large dividends in electoral contests suggests testable hypotheses that we have not considered here. First, if appeal has a universal component that translates well across locations, more appealing candidates may sort into larger, more significant jurisdictions in order to maximize the gains they reap from their personal attributes (Rosen, 1981). Second, if policy positions carry intrinsic value to politicians, then highly appealing candidates may choose to “spend” some of their electoral advantage by taking more extreme positions. These hypotheses may themselves have important implications for the functioning of political markets.

References

- [1] James D. Adams and Lawrence W. Kenny. The retention of state governors. *Public Choice*, 62:1–13, 1989.
- [2] Alberto Alesina and Howard Rosenthal. *Partisan politics, divided government, and the economy*. Political Economic of Institutions and Decisions. Cambridge University Press, Cambridge, UK, 1995.
- [3] Alberto Alesina, Nouriel Roubini, and Gerald D. Cohen. *Political Cycles and the Macroeconomy*. MIT Press, Cambridge, Massachusetts, 1997.
- [4] Nalini Ambady, Debi LaPlante, Thai Nguyen, Robert Rosenthal, Nigel Chaumeton, and Wendy Levinson. Surgeons’ tone of voice: A clue to malpractice history. *Surgery*, 132:5–9, July 2002.
- [5] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.
- [6] Niclas Berggren, Henrik Jordahl, and Panu Poutvaara. The looks of a winner: Beauty, gender and electoral success. *University of Helsinki Mimeograph*, 2006.
- [7] Timothy Besley and Anne Case. Incumbent behavior: Vote-seeking, tax-setting, and yardstick competition. *American Economic Review*, 85(1):25–45, March 1995.
- [8] Jeff E. Biddle and Daniel S. Hamermesh. Beauty, productivity, and discrimination: Lawyers’ looks and lucre. *Journal of Labor Economics*, 16(1):172–201, January 1998.
- [9] Colin Camerer, George Loewenstein, and Martin Weber. The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, 97(5):1232–1254, October 1989.
- [10] John E. Chubb. Institutions, the economy, and the dynamics of state elections. *American Political Science Review*, 82(1):133–154, March 1988.
- [11] Nancy L. Etcoff, Paul Ekman, John J. Magee, and Mark G. Frank. Lie detection and language comprehension. *Nature*, 405:139, May 11 2000.
- [12] Ray C. Fair. The effect of economic events on votes for president. *Review of Economics and Statistics*, 60(2):159–173, April 1978.
- [13] Ray C. Fair. Econometrics and presidential elections. *Journal of Economic Perspectives*, 10(3):89–102, Summer 1996.
- [14] Daniel Feenberg and Elisabeth Coutts. An introduction to the TAXSIM model. *Journal of Policy Analysis and Management*, 12(1), Winter 1993.
- [15] Morris P. Fiorina. Whatever happened to the median voter? *Stanford University Mimeograph*, 1999.
- [16] Alan Gerber. Estimating the effect of campaign spending on senate election outcomes using instrumental variables. *American Political Science Review*, 92(2):401–411, June 1998.

- [17] Edward L. Glaeser, Giacomo A. M. Ponzetto, and Jesse M. Shapiro. Strategic extremism: Why Republicans and Democrats divide on religious values. *Quarterly Journal of Economics*, 120(4):1283–1330, November 2005.
- [18] Amit Goyal and Ivo Welch. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, September 2006.
- [19] Daniel S. Hamermesh. Changing looks and changing "discrimination": The beauty of economists. *University of Texas mimeograph*, May 2005.
- [20] Daniel S. Hamermesh and Jeff E. Biddle. Beauty and the labor market. *American Economic Review*, 84(5):1174–1194, December 1994.
- [21] Jennifer M. Jensen and Thad Beyle. Of footnotes, missing data, and lessons for 50-state data collection: The Gubernatorial Campaign Finance Data Project, 1977-2001. *State Politics and Policy Quarterly*, 3(2):203–214, Summer 2003.
- [22] Benjamin F. Jones and Benjamin A. Olken. Do leaders matter? *Quarterly Journal of Economics*, 120(3):835–864, August 2005.
- [23] Susan L. Kone and Richard F. Winters. Taxes and voting: Electoral redistribution in the American states. *Journal of Politics*, 55(1):22–40, February 1993.
- [24] David S. Lee. The electoral advantage to incumbency and voters' valuation of politicians' experience: A regression discontinuity analysis of elections to the U.S. House. 2001. NBER Working Paper 8441.
- [25] David S. Lee, Enrico Moretti, and Matthew J. Butler. Do voters affect or elect policies? Evidence from the U.S. House. *Quarterly Journal of Economics*, 119(3):807–859, August 2004.
- [26] William Levernier. The effect of relative economic performance on the outcome of gubernatorial elections. *Public Choice*, 74:181–190, 1992.
- [27] Steven D. Levitt. Using repeat challengers to estimate the effect of campaign spending on election outcomes in the U.S. House. *Journal of Political Economy*, 102(4):777–798, August 1994.
- [28] Kevin M. Leyden and Stephen A. Borrelli. The effect of state economic conditions on gubernatorial elections: Does unified government make a difference? *Political Research Quarterly*, 48(2):275–290, June 1995.
- [29] Robert C. Lowry, James E. Alt, and Karen E. Ferree. Fiscal policy outcomes and electoral accountability in American states. *American Political Science Review*, 92(4):759–774, December 1998.
- [30] Markus M. Mobius and Tanya S. Rosenblat. Why beauty matters. *American Economic Review*, 96(1):222–235, March 2006.
- [31] Richard G. Niemi, Harold W. Stanley, and Ronald J. Vogel. State economies and state taxes: Do voters hold governors accountable? *American Journal of Political Science*, 39(4):936–957, November 1995.
- [32] Randall W. Partin. Economic conditions and gubernatorial contests: Is the state executive held accountable? *American Politics Quarterly*, 23(1):81–95, January 1995.

- [33] Sam Peltzman. Economic conditions and gubernatorial elections. *American Economic Review*, 77(2):293–297, May 1987.
- [34] Sam Peltzman. Voters as fiscal conservatives. *Quarterly Journal of Economics*, 107(2):327–361, May 1992.
- [35] Congressional Quarterly. *CQ Voting and Elections Collection*. CQ Electronic Library, <http://library.cqpress.com>, 2005.
- [36] Peter L. Rogers, Klaus R. Scherer, and Robert Rosenthal. Content filtering human speech: A simple electronic system. *Behavioral Research Methods and Instrumentation*, 3:16–18, 1971.
- [37] Sherwin Rosen. The economics of superstars. *American Economic Review*, 71(5):845–858, December 1981.
- [38] Shawn W. Rosenberg, Lisa Bohan, Patrick McCafferty, and Kevin Harris. The image and the vote: The effect of candidate presentation on voter preference. *American Journal of Political Science*, 30(1):108–127, February 1986.
- [39] James N. Schubert, Carmen Strungaru, Margaret Curren, and Wulf Schiefenhovel. Physische Erscheinung und die Einschätzung von politischen Kandidatinnen und Kandidaten. In Klaus Kamps and Meredith Watts, editors, *Biopolitics: Politikwissenschaft jenseits des Kulturismus*. Nomos Verlagsgesellschaft, Baden-Baden, 1998.
- [40] Philip E. Tetlock. Theory-driven reasoning about plausible pasts and probable futures in world politics: Are we prisoners of our preconceptions? *American Journal of Political Science*, 43(2):335–366, April 1999.
- [41] Alexander Todorov, Anesu N. Mandisodza, Amir Goren, and Crystal C. Hall. Inference of competence from faces predict electoral outcomes. *Science*, 308:1623–1626, June 10 2005.
- [42] Justin Wolfers. Are voters rational? evidence from gubernatorial elections. *Stanford GSB Research Paper Series*, 1730, March 2002.

Table 1 *The predictive power of participants' forecasts*

Dependent variable: Democrat share of two-party vote					
	(1)	(2)	(3)	(4)	(5)
Share predicting a Democrat victory	0.2424 (0.0618)	0.2793 (0.0597)	0.2721 (0.0990)	0.2383 (0.0714)	0.2794 (0.1138)
Democrat is taller				0.0215 (0.0305)	
Sample	All	More than 30 raters	Both candidates white males	Relative heights clear from video	2+ elections in state
State fixed effects?	No	No	No	No	Yes
R^2	0.2158	0.3075	0.1695	0.2600	0.5194
N	58	52	39	40	37

Notes: Results are from OLS regressions, with standard errors in parentheses. “Share predicting a Democrat victory” refers to the share of experimental participants who said they thought the Democratic candidate would win the gubernatorial election against the Republican candidate. “More than 30 raters” refers to elections that were viewed by over 30 study participants. “Relative heights clear from video” refers to a judgment from a selection of debate footage showing both candidates side by side (even though clips seen by participants showed each candidate alone). All calculations exclude respondents who claimed to recognize one or both of the candidates.

Table 2 *Economic predictors of election outcomes*

Dependent variable: Democrat's share of two-party vote				
<i>Index of one-year growth in:</i>	(1)	(2)	(3)	(4)
log(state personal income)	0.5386 (0.2948)	0.4470 (0.3067)		
State unemployment rate		-0.0186 (0.0174)		
log(state personal income)- log(national personal income)			0.0970 (0.5985)	
log(national personal income)			0.6221 (0.3115)	
log(state per capita revenues)				-0.2883 (0.2570)
log(per capita revenues in Census division)				0.4731 (0.2450)
R^2	0.0562	0.0754	0.0684	0.0647
Out-of-sample R^2	0.0172	-0.0000	-0.0665	0.0030
N	58	58	58	58
<i>After controlling for the above:</i>				
Share predicting a Democrat victory	0.2392 (0.0603)	0.2339 (0.0619)	0.2358 (0.0614)	0.2513 (0.0602)
Incremental out-of-sample R^2	0.1973	0.1813	0.2283	0.2250

Notes: Results are from OLS regressions, with standard errors in parentheses. “Share predicting a Democrat victory” refers to the share of experimental participants who said they thought the democratic candidate would win the gubernatorial election against the Republican candidate. “Index of one-year growth in log(state personal income)” is equal to the one-year growth (relative to the year prior to the election) of log of state personal income if the incumbent governor at the time of the election was a Democrat, equal to the negative of the growth of log(state personal income) if the incumbent governor was a Republican, and equal to zero if the incumbent governor was neither a Democrat nor a Republican. Other indices are defined analogously. “Out-of-sample R^2 ” is the out-of-sample mean squared prediction error of the model (estimated by leaving out each observation in sequence) divided by the out-of-sample mean squared prediction error of a constant-only model. “Incremental out-of-sample R^2 ” is the difference in out-of-sample R^2 between the specification including participants’ ratings and the specification excluding that variable.

Table 3 *Political predictors of election outcomes*

Dependent variable: Democrat’s share of two-party vote

	(1)	(2)	(3)	(4)
Average Democrat share of two-party vote, 1972-1987	0.0178 (0.2004)			-0.1957 (0.1582)
Difference in incumbency status		0.0737 (0.0166)		0.0409 (0.0176)
Difference in log(campaign spending)			0.0642 (0.0115)	0.0510 (0.0131)
R^2	0.0001	0.2594	0.3580	0.4265
Out-of-sample R^2	-0.0374	0.2279	0.3345	0.3565
N	58	58	58	58
<i>After controlling for the above:</i>				
Share predicting a Democrat victory	0.2433 (0.0625)	0.1609 (0.0626)	0.1398 (0.0586)	0.1122 (0.0593)
Incremental out-of-sample R^2	0.2020	0.0662	0.0478	0.0215

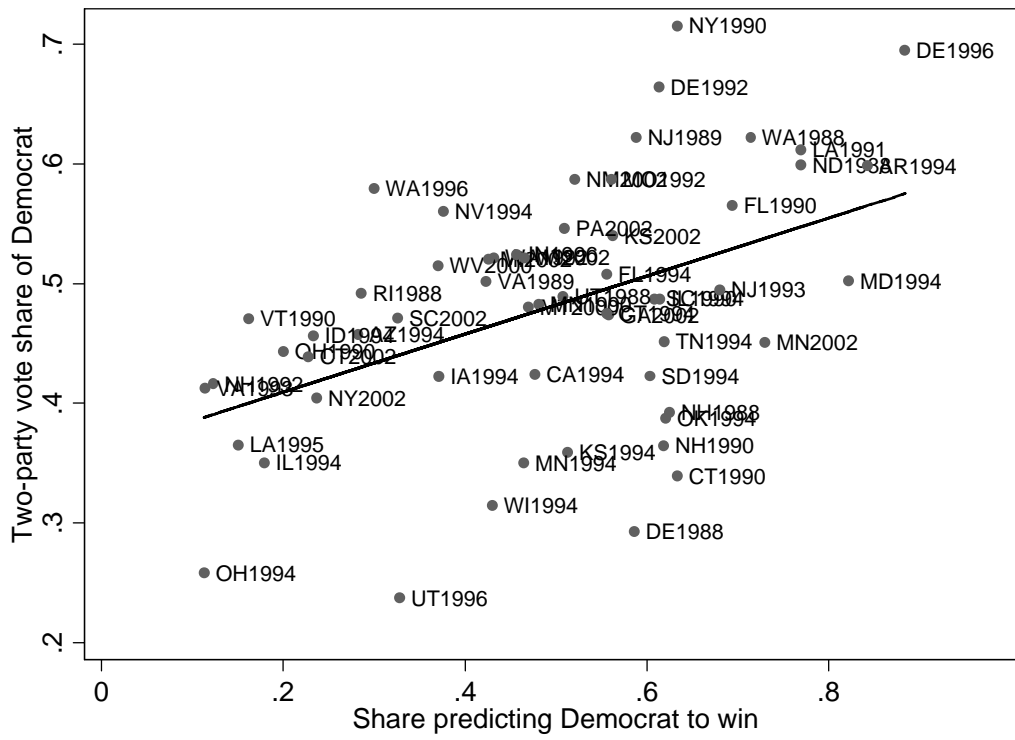
Notes: Results are from OLS regressions, with standard errors in parentheses. “Share predicting a Democrat victory” refers to the share of experimental participants who said they thought the democratic candidate would win the gubernatorial election against the Republican candidate. “Average Democrat share of two-party vote, 1972-1987” is the mean share of the two-party vote received by the Democratic candidate in gubernatorial elections in the state from years 1972 through 1987. “Difference in incumbency status” is equal to 1 if the Democratic candidate is an incumbent governor, -1 if the Republican candidate is an incumbent, and 0 if neither the Republican nor the Democratic candidate is an incumbent. “Difference in log(campaign spending)” is equal to the difference in the log of campaign spending between the Democrat and Republican, and is imputed at the state mean when missing. “Out-of-sample R^2 ” is the out-of-sample mean squared prediction error of the model (estimated by leaving out each observation in sequence) divided by the out-of-sample mean squared error of a constant-only model. “Incremental out-of-sample R^2 ” is the difference in out-of-sample R^2 between the specification including participants’ ratings and the specification excluding that variable.

Table 4 *Predictions, preferences, and electoral outcomes*Correlation matrix (*p-values in parentheses*)

	(1)	(2)	(3)
	Democrat's share of two-party vote	Share predicting Democrat to win	Share who would vote for Democrat
Share predicting Democrat to win	0.4646 (0.0002)	—	—
Share who would vote for Democrat	0.0723 (0.5895)	0.3141 (0.0164)	—
Average difference in ratings of candidate as:			
Physically attractive	0.1492 (0.2636)	0.2797 (0.0335)	0.7939 (<0.0001)
Likeable	0.1096 (0.4128)	0.2982 (0.0230)	0.8817 (<0.0001)
Good leader	0.2270 (0.0866)	0.6319 (<0.0001)	0.7329 (<0.0001)

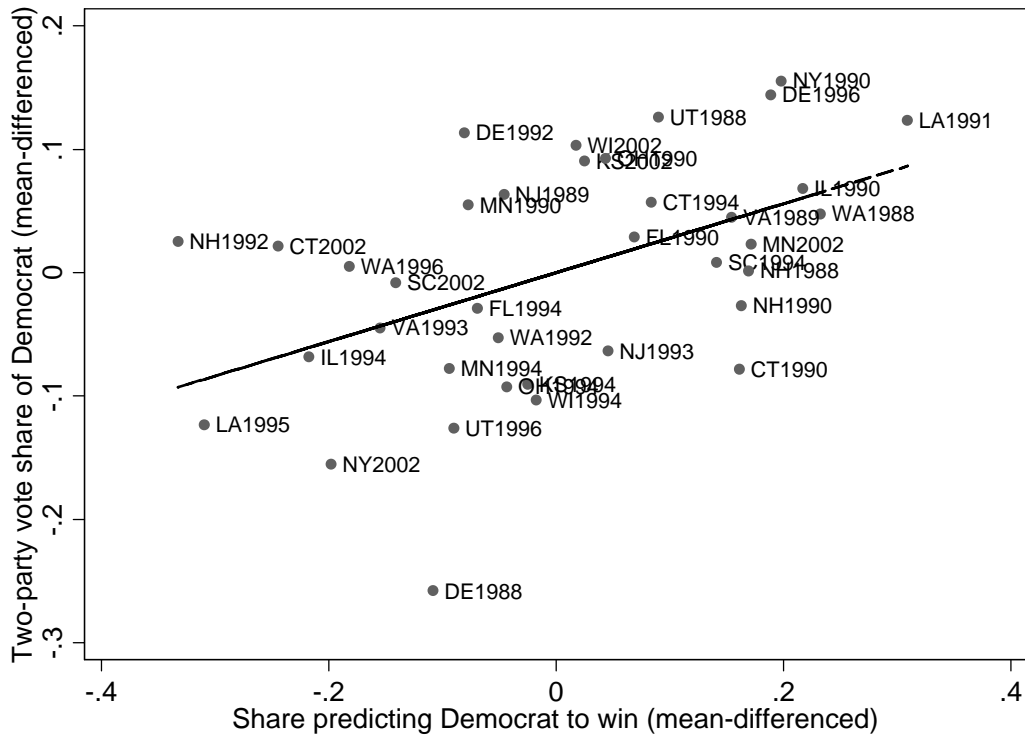
Notes: Table shows two-way correlation coefficients, with p-values in parentheses. All calculations exclude respondents who claimed to recognize one or both of the candidates.

Figure 1 *Predicted and actual two-party vote shares*



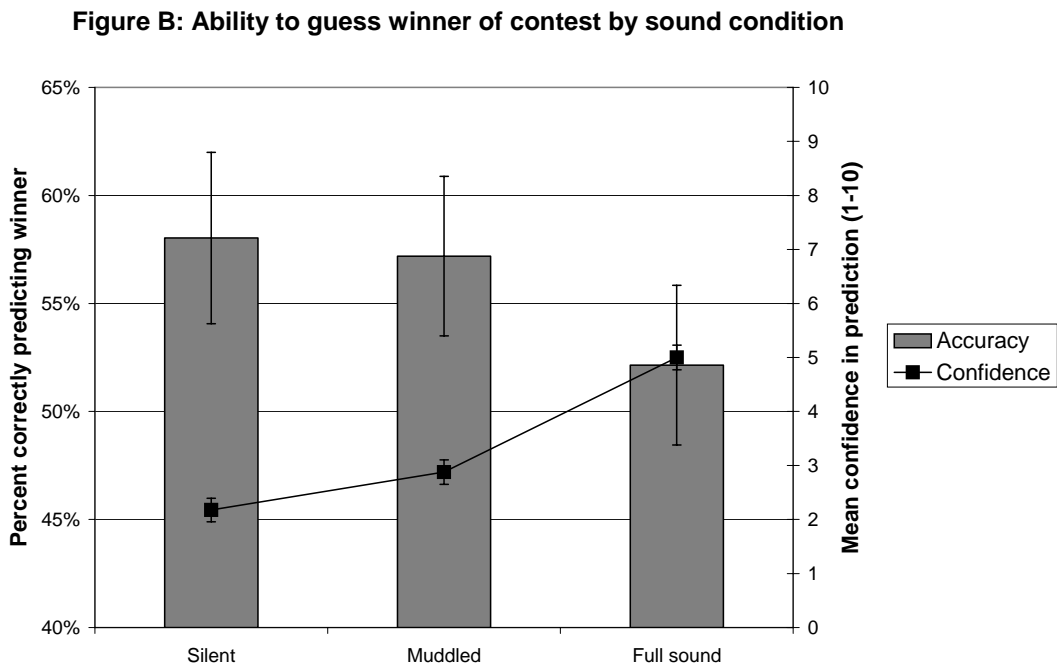
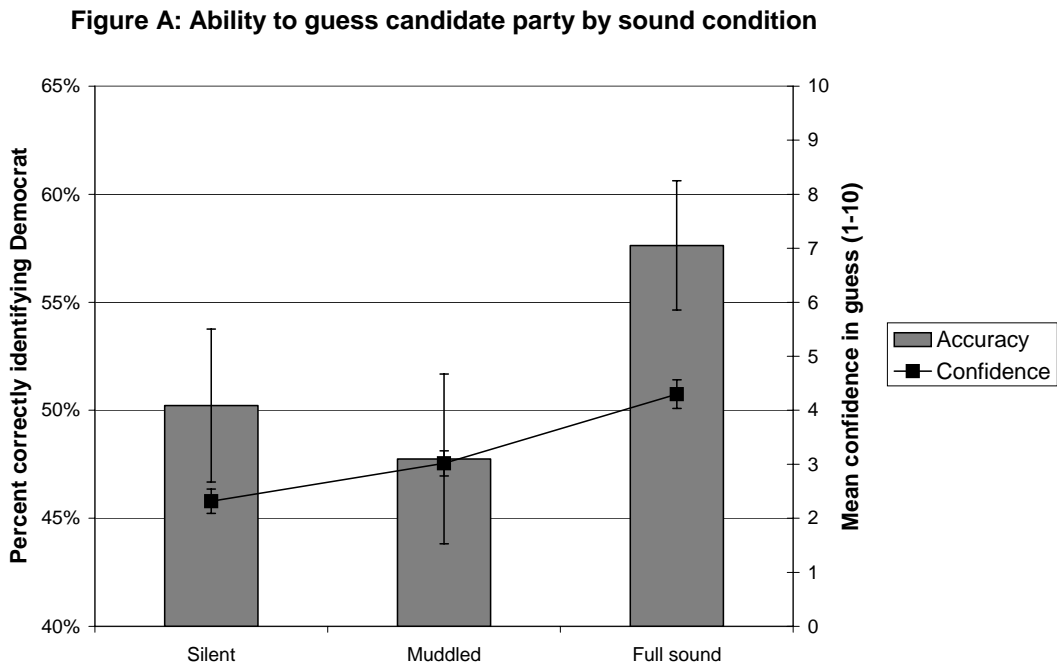
Notes: Figure shows share of two-party vote received by Democratic candidate on y-axis, and share of experimental participants (in silent condition) who predicted the Democratic candidate to win the election. Predictions from participants who claimed to recognize one or both of the candidates are excluded from the analysis. Number of elections is 58.

Figure 2 *Within-state evidence on forecast accuracy*



Notes: Y-axis shows share of two-party vote received by Democratic candidate, less the average of that variable over the sample elections in the same state. X-axis shows share of experimental participants (in silent condition) who predicted the Democratic candidate to win the election, less the average of that variable over the sample elections in the same state. Predictions from participants who claimed to recognize one or both of the candidates are excluded from the analysis. Number of elections is 37.

Figure 3 *The effect of policy information on forecast accuracy*



Notes: Error bars are ± 1 standard error. Data are from the first (May 2005) round of the study. Number of elections (for accuracy measures) is 33. Number of participants (for confidence measures) is 57.

A Appendix: Alternative Economic Predictors of Gubernatorial Elections

In this appendix, we examine several alternative economic predictors of gubernatorial elections, as a supplement to table 2. The discussion below refers to the appendix table. Each specification in the appendix table regresses the Democrat’s share of the two-party vote on a different set of economic predictors of election outcomes for our sample of 58 elections. For each specification, we report the R^2 , the out-of-sample R^2 , and the incremental out-of-sample R^2 from adding participants’ ratings to the model. We also report an “adjusted” coefficient on the share of participants predicting a Democratic victory, after controlling for the economic factors listed. In all cases, this coefficient is similar in magnitude and statistical precision to the coefficients we report in table 1. In addition, in all cases the out-of-sample R^2 of the economic model is below 10 percent (and the incremental out-of-sample R^2 from participants’ ratings is at least 14 percent), indicating that these alternative sets of economic predictors have significantly less predictive power than our laboratory-based predictor.

In column (1) of table 2 we examine the predictive power of one-year growth rates of state personal income. However, voters may be comparing current economic performance to the performance as of the previous election, in which case four-year lags could be more appropriate. In specification (1) of the appendix table we examine the predictive power of the four-year growth rate in log income. Consistent with Fair (1978), we find that this variable is a weaker predictor than the one-year growth rate, and has no out-of-sample predictive power. In specification (2) we augment specification (1) by adding a measure of the four-year growth rate in unemployment, and find no improvement in out-of-sample fit.

Specification (3) implements a model in which voters completely ignore state trends and focus only on national income growth in deciding how to vote. The one-year growth in national income predicts about three percent of the variation in vote shares out of sample. Specification (4) adds the national unemployment rate growth to specification (3), resulting in an out-of-sample R^2 of about 9 percent, the highest of our various economic forecasting models.

In table 2 we estimate a model in which voters compare state revenue growth to growth in revenues of neighboring states. An alternative possibility is that they compare revenue growth to national levels, which we check in specification (5). In this model, we include the one-year growth rate in log(state revenues per capita), along with the growth in the log of the population-weighted average revenue of all other states. This specification has no out-of-sample predictive power.

Although Peltzman (1992) uses revenues to measure state fiscal policy, Besley and Case (1995) suggest using income tax levels as measured by the NBER’s TAXSIM program.¹⁸ In specification (6), we parallel the specification in table 2, but use a TAXSIM-based measure of state revenues. In particular, we compute for each state and year the state income tax liability for a married, single-earner household with two dependents that earns \$35,000 per year. While we do find some evidence that higher taxes provoke a voter response, this model has weak out-of-sample forecasting power.

A number of authors (Chubb, 1988; Levernier, 1992; Kone and Winters, 1993; Alesina and Rosenthal, 1995) have suggested that local races might be affected by presidential “coattails,” in the sense that voters may attribute the successes and failures of the president to others in the same political party. In specification (7), we predict gubernatorial elections using the Democrat’s share of the two-party vote in the most recent presidential election, but find that this specification has only weak forecasting power.

In specification (8), we implement an alternative model of presidential coattails, in which voters attribute variation in national economic conditions to the incumbent president’s party. The model

¹⁸See Feenberg and Coutts (1993) and <<http://www.nber.org/~taxsim/>>.

predicts the Democrat's share of the gubernatorial vote using a measure of national income growth, a measure of whether the president is a Democrat, and the interaction of the two, which may be seen as a simple representation of a model in which national trends are attributed to the gubernatorial candidate of the same party as the president. This specification has moderate out-of-sample forecasting power, successfully predicting about 5 percent of the overall variation in two-party vote shares.

In specification (9) we allow for voters to treat income growth and income decline differently. In particular, we allow for different coefficients on income growth depending on whether growth was positive or negative over the previous year, and we also include a measure of whether growth was positive in the previous year. In out-of-sample tests, this specification predicts about 4 percent of the variation in the Democrat's share of the two-party vote.

Appendix Table: Alternative Economic Predictors of Gubernatorial Elections

Specification	Unadjusted R^2 (<i>out-of-sample</i> R^2)	Adjusted coeff. on lab measure (standard error)	Incremental out-of-sample R^2 of lab measure
(1) Four-year growth in log(state personal income)	0.0229 (-0.0197)	0.2432 (0.0613)	0.2011
(2) (1) + four-year growth in unemployment rate	0.0401 (-0.0512)	0.2515 (0.0609)	0.2281
(3) One-year growth in log(national personal income)	0.0680 (0.0294)	0.2317 (0.0607)	0.1811
(4) (3) + one-year growth in unemployment rate	0.1539 (0.0918)	0.2119 (0.0600)	0.1485
(5) One-year growth in log(state revenue per capita) + log(national average state revenue per capita)	0.0354 (-0.0290)	0.2661 (0.0612)	0.2487
(6) One-year growth in TAXSIM state taxes + growth in average TAXSIM taxes in Census division	0.0677 (-0.0233)	0.2265 (0.0625)	0.2113
(7) Democrat's share of two-party vote in most recent presidential election	0.0279 (-0.0185)	0.2343 (0.0631)	0.1809
(8) One-year growth in log(national personal income) + Democrat is president + Democrat is president \times growth in log(national income)	0.1443 (0.0485)	0.2286 (0.0593)	0.1867
(9) One-year growth is positive + growth in log(state income) + Positive growth \times growth in log(income)	0.1242 (0.0401)	0.2223 (0.0614)	0.1620

Notes: See appendix A for details.