Thin-Slice Forecasts of Gubernatorial Elections

Daniel J. Benjamin*

Cornell University and Institute for Social Research

Jesse M. Shapiro**

University of Chicago and NBER

First Draft: June 28, 2005

This Draft: December 9, 2007

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' forecasts significantly outperform economic variables in predicting vote shares, and are comparable in predictive power to a measure of incumbency status. Participants' forecasts seem to rest on judgments of candidates' personal attributes (such as likeability), rather than inferences about candidates' policy positions. Though conclusive causal inference is not possible in our context, our findings may be seen as suggestive evidence of a causal effect of candidate appeal on election outcomes.

JEL classification: D72, J45, Z19

Keywords: thin slices, charisma, elections

Acknowledgments

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, Steve Coate, Stefano DellaVigna, Ray Fair, Luis Garicano, Matt Gentzkow, Ed Glaeser, David Laibson, Steve Levitt, Ulrike Malmendier, Hal Movius, Kevin M. Murphy, Ted O'Donoghue, Emily Oster, Jane Risen, Emmanuel Saez, Bruce Sacerdote, Andrei Shleifer, Matt Weinzerl, Rick Wilson, the editor, two anonymous referees, and seminar participants at the University of Chicago, Harvard University, the Stanford Institute for Theoretical Economics, the University of Michigan, Michigan State University, University of Texas-Austin, and Texas A&M 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 and P01-AG26571 to the Institute for Social Research; the Institute for Humane Studies; and the National Science Foundation for financial support. We are very grateful to Sarah Bommarito, Sujie Chang, Ghim Chuan, Chris Convery, Jonathan Hall, Rebecca Hausner, Ethan Lieber, Annette Leung, Hans Lo, Dina Mishra, Marina Niessner, Shawn Nelson, Mark Petzold, Krishna Rao, Tiye Sherrod, David Sokoler, Bernardo Vas, and Narendra Vempati 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: db468@cornell.edu, telephone: 607-255-6314. Mailing address: Uris Hall, Economics Department, Cornell University, Ithaca NY 14853.

** Corresponding author. E-mail: jmshapir@uchicago.edu, telephone: 773-834-2688, fax: 773-834-8172. Mailing address: University of Chicago, 5807 S. Woodlawn Avenue, Chicago, IL 60637.

1 Introduction

From 1988 to 2003, the standard deviation in the two-party vote share in U.S. gubernatorial elections was 12 percentage points, and the interquartile range was from 40 percent to 53 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, forthcoming). 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; Snowberg, Wolfers, and Zitzewitz, 2007).

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 what drives participants' forecasts. 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 tended, if anything, to worsen. In contrast, we show that variation in participants' ratings of candidate likeability, physical attractiveness, and leadership can account for about one-third of the accuracy of participants' forecasts. Finally, we argue that forecasting skill is fairly homogeneous within our sample of participants.

Finally, we discuss possible causal interpretations of our findings. 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, just as they influence outcomes in other labor markets (Hamermesh and Biddle, 1994; Biddle and Hamermesh, 1998; Mobius and Rosenblat, 2006). While we cannot conclusively demonstrate a causal effect of candidate appeal, we argue using a range of robustness checks and alternative specifications that our data do not clearly support alternative causal mechanisms.

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 (2006) studies the role of attractiveness in American Economic Association elections using students' ratings of still photographs. King and Leigh (2006) and Klein and Rosar (2005) find that ratings of physical attractiveness predict election outcomes in Australia and Germany, respectively. 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. 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, unlike existing work, we assess the incremental predictive power of personal appeal, after accounting for economic and political predictors of electoral success, and we 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. 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 (an exception is Schubert et al, 1998, who also used debate footage).

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 forecasts of highly informed experts often perform no better than chance in predicting political events like elections

(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 demonstrate 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 briefly discusses possible causal interpretations of our findings. Section 7 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 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

[1] The 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.

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 (see section 5.1).

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.[2]

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.[3] 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.[4]

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:

- 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?

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

[3] Feedback from informal interviews with participants after early sessions led us to make small changes to the experimental procedure in later sessions, such as modifying the proportion of the silent, muddled, and full-sound clips each participant viewed from 3/21 – 3/21 – 15/21 to 7/21 – 7/21 – 7/21, respectively (see Benjamin and Shapiro, 2007, for details).

[4] Statistical tests show no difference in participants' ability to forecast election outcomes across the three rounds of sessions.

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 limited 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.* We obtained annual data on state per capita income from the Bureau of Economic Analysis (<http://www.bea.gov/bea/regional/data.htm>). 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).[5]

- *Unemployment rate.* 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.* 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 results are robust to using a tax-simulation-based measure.

- *House prices.* For quarterly data on house prices at the state level, we used the Office of Federal Housing Enterprise Oversight's (OFHEO) House Price Index.[6]

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

- *Incumbent candidate and party.* We identified the incumbent candidate (if any) and party in each race using the *CQ Voting and Elections Collection* (2005).[7]

[5] We 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.

[6] Downloaded from their website <http://www.ofheo.gov/hpi_download.aspx>, as of November, 2007.

[7] We cross-checked information on the incumbent party using

- *Campaign spending.* 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. 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 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 ($p = 0.002$). (Except where noted, all analysis uses data from the silent condition.)

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

<http://en.wikipedia.org/wiki/List_of_United_States_Governors> (in January 2006) and the National Governors Association website <http://www.nga.org> (in December 2007).

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.[8] 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.

The R^2 in column (1) might understate the true explanatory power of participants' 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

[8] For simplicity of interpretation, we focus throughout on linear regression of vote share on the share of participants predicting a Democratic victory. This specification is appropriate if voters follow a random-utility model with uniformly distributed choice errors, in which the share predicting a Democratic victory measures a candidate characteristic that influences the relative utility from voting for the Democratic candidate. Results are virtually identical in terms of effect size and R^2 when we instead estimate a model in which the choice error is normally distributed.

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 present 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.[9] 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.[10] (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

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

[10] The 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.

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).[11]

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 in identifying the party affiliations of the candidates.[12] 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.

[11] Related 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.

[12] This is so despite the fact that 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.

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, equal to the negative of the growth rate if the incumbent governor is a Republican, and equal to zero if the incumbent governor is neither a Republican nor a Democrat. 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-of-sample measure of the fit of each model.[13] 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

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

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 for the incumbent party, 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 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 Ferree, 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 also 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) argue 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.

Wolfers (2002) proposes the inclusion of house prices in models of gubernatorial elections, because economic theory suggests that changes in house prices capitalize many important aspects of a governor's policies. In column (5) of table 2, we use the log of the state house price index as a measure of house prices.[14] We find that although the point estimate is consistent with the theory, changes in state house prices fare no better than other economic variables in predicting

[14] For consistency with our other economic predictors, we use the change in the log index beween the fourth quarter of the election year and the fourth quarter of the previous year. Results are comparable if we use a four-year lag.

election outcomes, with a negative out-of-sample R^2. Moreover, including house prices does not substantively affect the predictive power of our participants' forecasts.

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.[15] 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. Our measure of incumbency is an index equal to 1 when the Democrat is an incumbent,

[15] To check whether the weak performance of this variable is due to our use of an historical lag, 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 specification.

0 when neither candidate is an incumbent, and -1 when the Republican is an incumbent.[16] We estimate that being an incumbent results in roughly a 7 percentage point electoral advantage, which is quite similar to Lee's (forthcoming) 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.[17] 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

[16] Unreported 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.

[17] As 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.

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. Moreover, 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.

5 What Do Participants' Predictions Measure?

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. Participants' ratings seem to be driven by personal attributes of candidates (such as likeability), rather than inferences about their policy positions. Moreover, these attributes seem to be universally detectable (at least within our sample population), in the sense that different raters performed similarly in forecasting election outcomes.

5.1 Policy Inferences

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 2 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 identification 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, if anything, to *worsen* predictive accuracy (see part B of figure 2). 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 in election-level tests,[18] 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.[19]

[18] When we estimate 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 (with standard errors clustered by rater), we find a marginally statistically significant difference between the full sound and silent conditions ($p = 0.057$) and no difference between the silent and muddled conditions. Moreover, when we pool results from the silent and muddled conditions, we find that these are jointly statistically significantly different from the full sound condition ($p = 0.028$).

[19] Informal 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.

5.2 Participant Heterogeneity

One possible interpretation of our findings is that a small fraction of the population is highly skilled at forecasting electoral success based on visual cues. In fact, we find little evidence for individual differences in predictive accuracy: the accuracy of respondents' guesses is not statistically significantly correlated with gender, SAT scores, interest in politics, or political preferences. Moreover, random effects models indicate little individual-specific variation in predictive accuracy.[20] In other words, the factors measured in participants' ratings appear to be no more visible to some raters than to others.

5.3 Candidate Attributes

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. All three of these ratings are individually statistically significantly correlated with the share of participants predicting the Democrat to win.

For a more quantitative evaluation of the role of these factors, we regress the share predicting the Democrat to win on the vector of candidate characteristics (results not shown), and extract the predicted values from the regression. A regression of actual vote shares on the resulting index has an R^2 of approximately 0.07, suggesting that on the order of one-third of participants' forecast accuracy can be attributed to their impressions of the attractiveness, likeability, and leadership qualities of the candidates. Thus, while these factors leave the majority of participants' predictive power unexplained, they can nevertheless account for a nontrivial fraction of participants' forecast

[20] We also find no evidence that predictive power varies significantly with the interaction of candidate and respondent characteristics.

accuracy.[21]

6 Discussion and Interpretation

A direct interpretation of the predictive power of participants' ratings is that the ratings measure some characteristic of a candidate, which might broadly be called "appeal" or "charisma," that has a direct, causal influence on electoral success. To reach such a conclusion, however, it is important to rule out competing explanations for participants' predictive power. Below we briefly discuss evidence on three potential biases in a causal interpretation of our estimates. While we find no evidence to suggest the presence of significant confounds to a causal interpretation, we cannot conclusively rule out such concerns. We analyze these issues at greater length in Benjamin and Shapiro (2007).

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. A simple test for this confound is to directly measure whether the candidates appear confident and ask whether controlling for perceived confidence affects the estimated predictive power of participants' ratings. To this end, we asked a set of 10 research assistants unfamiliar with our hypotheses (and with the election results) to rate the apparent confidence of the candidates in our sample.[22] The difference in average confi-

[21] We also find that participants' own preferences, as measured by responses to questions about which candidate the participant would vote for, are only weakly predictive of vote shares. Analysis suggests this may be because participants' own political preferences (which are idiosyncratic and hence not predictive of state-level election outcomes) may have played a larger role in driving their stated voting preferences than their predictions about the election winner.

[22] The research assistants' ratings are reasonably correlated with one another, with an effective reliability (R) of 0.77 (Rosenthal, 1987), suggesting that the ratings did identify a common element

dence ratings received by the Democrat and Republican candidates in each race in our sample is not statistically significantly correlated with either the Democrat's share of the two-party vote or the share of participants predicting the Democrat to win. As a consequence, including this confidence measure as a control in a regression of vote shares on participants' ratings has no meaningful effect on the magnitude or statistical significance of the estimated effect of candidate appeal on electoral success.

Another potential confound is candidate selection: If political parties are more likely to choose appealing or charismatic candidates when their chances of victory are greater (or if appealing candidates are more likely to run in such periods), then the regressions in table 1 could overstate the strength of the causal relationship between candidates' visible characteristics and vote shares.[23] If the circumstances that cause the appealing candidate's party to win are serially correlated, then we would expect to find a positive relationship between participants' ratings and *lagged* vote shares, but instead we find a small and statistically insignificant relationship.[24] We can also test directly whether participants' ratings are correlated with plausibly exogenous, measurable predictors of

among the clips.

[23] We thank a referee for pointing out that selection bias may cause us to *underestimate* the relationship between personal appeal and election outcome. Conditional on winning the primary race, a less appealing candidate may be stronger on non-visual attributes that contribute to electoral success.

[24] A seemingly unrelated regression shows that the relationship between participants' ratings and lagged vote shares is marginally statistically significantly different from the relationship between participants' ratings and current vote shares ($p = 0.063$). In some races, one or more of the candidates from the current race may have participated in the previous race. In a subsample of races with no incumbent participating, we continue to find no relationship between lagged vote shares and participants' ratings. Moreover, results are similar when we use a longer historical average vote share for the state (from 1972 to 1987) rather than the vote share for the previous

electoral outcomes: historical average vote shares, state personal income, national personal income, state unemployment rate, state per capita revenues, and per capita revenues in Census division. A regression of the thin-slice forecast on these variables reveals no jointly or individually significant relationship. Of course, an absence of correlation with observables does not prove an absence of correlation with unobservables.

A final potential confound is that visual attributes of candidates are correlated with other characteristics—such as intelligence or managerial skill—that are themselves valuable for governors to possess, and it is these other skills that cause electoral success. Indeed, our earlier evidence (subsection 5.3 above) shows that participants' ratings are correlated with perceptions of "good leadership," which could imply that participants are (or believe they are) able to predict job performance based on the silent clips. As a contrasting piece of evidence, however, we note that undergraduate research assistants' ratings of candidates' *competence* do not exert a significant independent effect on vote shares, once we control for participants' predictions of the election outcome.[25] These contrasting findings leave open the question of whether the relationship between participants' ratings and vote shares is driven by inference about underlying candidate quality.

election.

[25] We asked 10 research assistants to do these ratings – as well as the ratings of confidence discussed above – because we did not include a question about competence (or confidence) in our original design. Note, however, that the effective reliability (R) is only 0.49 (significantly lower than in the case of the confidence ratings). Note also that, while this result stands in partial contrast to Todorov et al. (2005), our specification is not directly comparable to his because we control for participants' forecasts. When we regress vote shares on competence ratings without controlling for forecasts, we find (consistent with Todorov et al.'s results) that competence is a statistically significant predictor of vote shares.

7 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.

Our findings do not conclusively show that candidate appeal causally affects 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. For example, 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). If policy positions carry intrinsic value to politicians, then highly appealing candidates may choose to adopt different policy positions from less appealing ones. 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] Daniel J. Benjamin and Jesse M. Shapiro. Thin-slice forecasts of gubernatorial outcomes. *NBER Working Paper No. 12660*, November 2007.

[7] Niclas Berggren, Henrik Jordahl, and Panu Poutvaara. The looks of a winner: Beauty, gender and electoral success. *University of Helsinki Mimeograph*, 2006.

[8] Timothy Besley and Anne Case. Incumbent behavior: Vote-seeking, tax-setting, and yardstick competition. *American Economic Review*, 85(1):25–45, March 1995.

[9] 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.

[10] 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.

[11] John E. Chubb. Institutions, the economy, and the dynamics of state elections. *American Political Science Review*, 82(1):133–154, March 1988.

[12] Nancy L. Etcoff, Paul Ekman, John J. Magee, and Mark G. Frank. Lie detection and language comprehension. *Nature*, 405:139, May 11 2000.

[13] Ray C. Fair. The effect of economic events on votes for president. *Review of Economics and Statistics*, 60(2):159–173, April 1978.

[14] Ray C. Fair. Econometrics and presidential elections. *Journal of Economic Perspectives*, 10(3):89–102, Summer 1996.

[15] Daniel Feenberg and Elisabeth Coutts. An introduction to the TAXSIM model. *Journal of Policy Analysis and Management*, 12(1), Winter 1993.

[16] Morris P. Fiorina. Whatever happened to the median voter? *Stanford University Mimeograph*, 1999.

[17] 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.

[18] 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.

[19] Amit Goyal and Ivo Welch. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*. Forthcoming.

[20] Daniel S. Hamermesh. Changing looks and changing "discrimination": The beauty of economists. *Economics Letters*, 93:405–412, 2006.

[21] Daniel S. Hamermesh and Jeff E. Biddle. Beauty and the labor market. *American Economic Review*, 84(5):1174–1194, December 1994.

[22] 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.

[23] Benjamin F. Jones and Benjamin A. Olken. Do leaders matter? *Quarterly Journal of Economics*, 120(3):835–864, August 2005.

[24] Amy King and Andrew Leigh. Beautiful politicians. *Australian National University Mimeograph*, 2006.

[25] Markus Klein and Ulrich Rosar. Physical attractiveness and electoral success. an empirical investigation on candidates in constituencies at the German federal election 2002. *Politische Vierteljahresschrift*, 46(2):263–287, June 2005.

[26] 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.

[27] David S. Lee. Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, forthcoming.

[28] 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.

[29] William Levernier. The effect of relative economic performance on the outcome of gubernatorial elections. *Public Choice*, 74:181–190, 1992.

[30] 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.

[31] 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.

[32] 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.

[33] Markus M. Mobius and Tanya S. Rosenblat. Why beauty matters. *American Economic Review*, 96(1):222–235, March 2006.

[34] 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.

[35] Randall W. Partin. Economic conditions and gubernatorial contests: Is the state executive held accountable? *American Politics Quarterly*, 23(1):81–95, January 1995.

[36] Sam Peltzman. Economic conditions and gubernatorial elections. *American Economic Review*, 77(2):293–297, May 1987.

[37] Sam Peltzman. Voters as fiscal conservatives. *Quarterly Journal of Economics*, 107(2):327–361, May 1992.

[38] Congressional Quarterly. *CQ Voting and Elections Collection.* CQ Electronic Library, http://library.cqpress.com, 2005.

[39] 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.

[40] Sherwin Rosen. The economics of superstars. *American Economic Review*, 71(5):845–858, December 1981.

[41] 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.

[42] Robert Rosenthal. *Judgment Studies: Design, Analysis, and Meta-Analysis*. Cambridge University Press, New York, 1987.

[43] 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 jensets des Kulturismus*. Nomos Verlagsgesellschaft, Baden-Baden, 1998.

[44] Erik Snowberg, Justin Wolfers, and Eric Zitzewitz. Partisan impacts on the economy: Evidence from prediction markets and close elections. *Quarterly Journal of Economics*, 122(2):807–829, May 2007.

[45] 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.

[46] 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.

[47] Justin Wolfers. Are voters rational? evidence from gubernatorial elections. *Stanford GSB Research Paper Series*, 1730, March 2002.

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.[26] 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 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

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

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 (*out-of-sample* 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 × 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 × growth in log(income)	0.1242 (*0.0401*)	0.2223 (0.0614)	0.1620

Notes: See appendix A for details.

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	0.3042	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 (in silent condition) 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

Index of one-year growth in:	(1)	(2)	(3)	(4)	(5)
log(state personal income)	0.5386	0.4470			
	(0.2948)	(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)	
log(state house price index)					0.2991
					(0.2714)
R^2	0.0562	0.0754	0.0684	0.0647	0.0212
Out-of-sample R^2	0.0172	-0.0000	-0.0665	0.0030	-0.0207
N	58	58	58	58	58
After controlling for the above:					
Share predicting a Democrat victory	0.2392 (0.0603)	0.2339 (0.0619)	0.2358 (0.0614)	0.2513 (0.0602)	0.2381 (0.0620)
Incremental out-of-sample R^2	0.1973	0.1813	0.2283	0.2250	0.1834

Notes: Results are from OLS regressions, with standard errors in parentheses. "Share predicting a Democrat victory" refers to the share of experimental participants (in silent condition) 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.

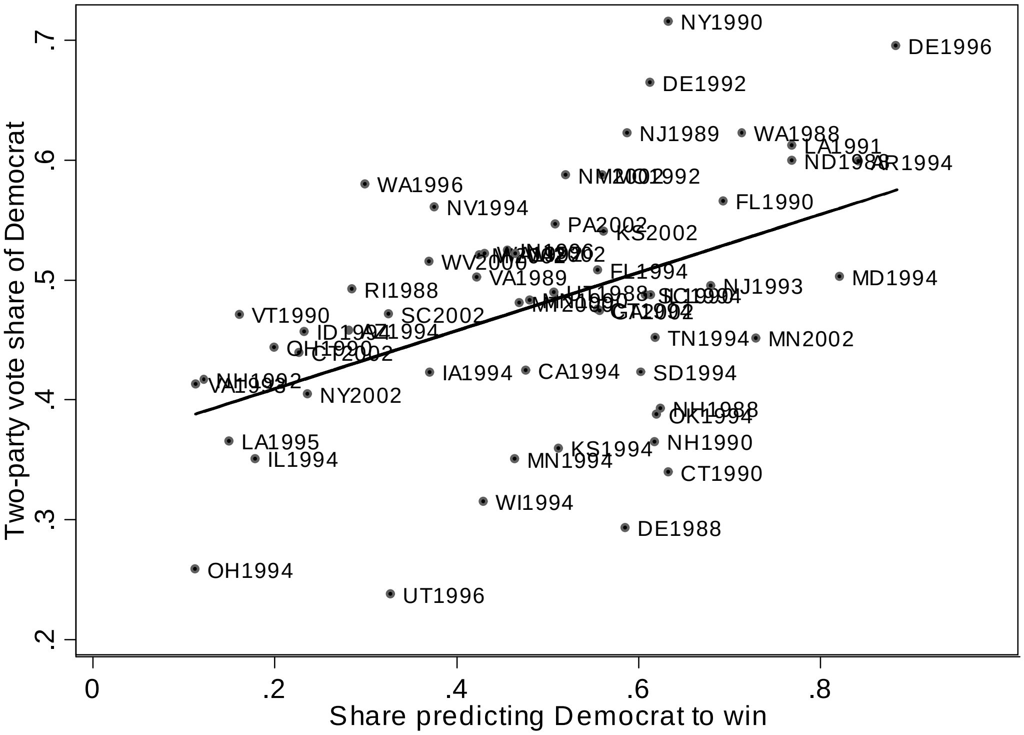
Table 3 *Political predictors of election outcomes*

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
After controlling for the above:				
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 (in silent condition) 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.

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 *The effect of policy information on forecast accuracy*

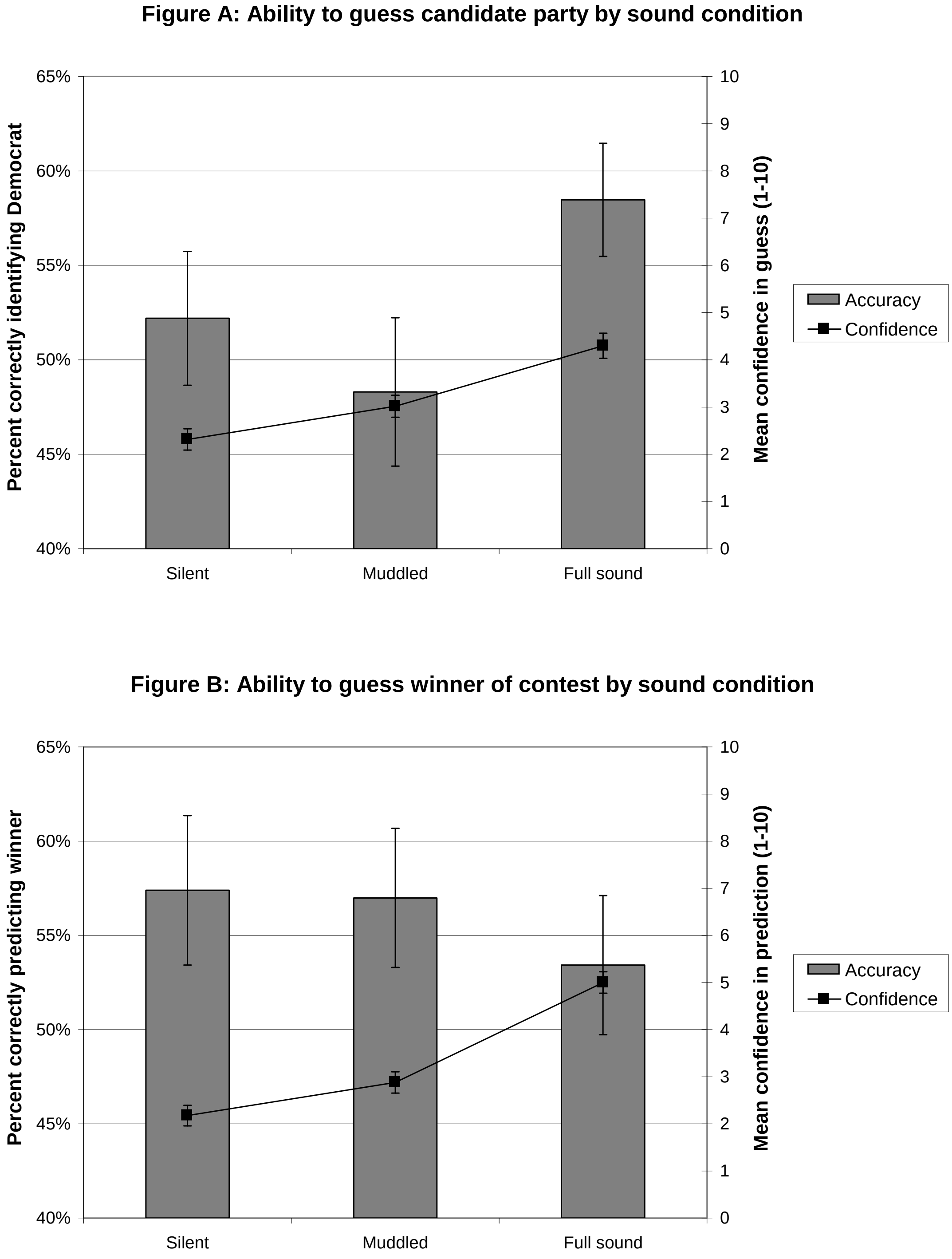


Figure A: Ability to guess candidate party by sound condition

Figure B: Ability to guess winner of contest by sound condition

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.