The Rational Voter, Thinly Sliced:  
Personal Appeal as an Election Forecaster

Daniel J. Benjamin  
Harvard University  

Jesse M. Shapiro*  
University of Chicago and NBER  

April 24, 2006

Abstract

We construct a measure of gubernatorial candidates’ personal appeal by showing 10-second, silent video clips of unfamiliar gubernatorial debates to a group of experimental participants and asking 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. Adding policy information to the video clips by turning on the sound tends, if anything, to worsen participants’ accuracy. In a horse race of alternative forecasting models, personal factors as measured in our study significantly outperform economic variables in predicting vote shares, and are comparable in predictive power to a measure of incumbency status.

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 Nalini Ambady, Chris Chabris, Luis Garicano, Matt Gentzkow, Ed Glaeser, David Laibson, Steve Levitt, Kevin Murphy, Emily Oster, Jane Risen, Emmanuel Saez, Andrei Shleifer, and seminar participants at the University of Chicago for helpful comments. We are very grateful to Suje Chang, Ethan Lieber, Dina Mishra, Marina Neissner, 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: benjamin@fas.harvard.edu, 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 determinants 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. 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 study the importance of candidates’ personal appeal, relative to economic and political factors, in explaining electoral outcomes. By definition, “personal appeal” refers to a politician’s appeal to voters as a person, apart from policy positions (or party affiliation). Although it may be difficult to describe the factors that determine personal appeal, it can be measured by how people react to a politician when those people lack information about policy stands. Indeed, reactions to even small amounts of visual information may be highly informative about personal appeal, as a large literature in social psychology establishes that judgments about other people from “thin slices”—exposures to expressive behavior as brief as a few seconds—are highly predictive of reactions to much longer exposures (see Ambady and Rosenthal, 1992 for a review).

We showed experimental participants 10-second silent video clips of the major party candidates in 58 gubernatorial elections from 1988-2002. We assessed the candidates’ relative appeal by asking the participants who they predicted would win. Ratings are highly reliable across different debates from the same elections. Moreover, participants’ predictions are highly related to electoral outcomes, and can account for over 20 percent of the variation in two-party vote shares in these elections. Participants’ ratings are strongly related to electoral outcomes even after we control for state fixed effects, indicating that participants’ ratings are not merely proxying for stable differences across states in the relative appeal of different political parties.

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. This finding is consistent with earlier results suggesting that verbal information can interfere with more instinctive visual judgments (e.g., Etcoff et al, 2000). 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).

We also find that participants’ own tastes do not seem to explain the accuracy of their forecasts: participants’ ratings of likability, physical attractiveness, and participants’ own preferences among the candidates are only weakly related to election outcomes. Rather, participants seem to be able to infer how other individuals—specifically, voters in the relevant jurisdiction—would react to the two candidates. As a final robustness check, we test extensively for biases coming from participants’ greater familiarity with the victorious candidate, and find no evidence for such a confound.

After analyzing the results of our laboratory study, we turn to a horse race comparison between alternative methods of election forecasting. We find that the personal factor we measure performs far better than economic indices in predicting election outcomes, and is comparable in explanatory power 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 five-variable model in predicting the vote shares in our sample of 58 elections.


We make several contributions relative to this existing literature. 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. Our approach also sheds light
on the importance of participants’ own preferences in determining their guesses about election outcomes. 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. Finally, our regression framework permits us to test whether measures of personal appeal remain important after controlling for economic and political predictors of electoral success, which could themselves have an influence on the charisma of the chosen candidate, and to compare the relative predictive power of these factors.

In addition to the economics of voter behavior, this paper contributes to a growing literature on the role of beauty in labor markets (Hamermesh and Biddle, 1994; Biddle and Hamermesh, 1998; Mobius and Rosenblat, 2006). Although we cannot make definitive causal claims based on our estimates, our findings are suggestive of a substantial return to personal appeal in political competition.

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

2 Laboratory Procedures for Eliciting Election Predictions

We showed participants 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. We analyzed how accurately participants could consciously predict the actual election outcomes on the basis of these “thin slices,” as well as how well their snap assessments of the candidates’ personal characteristics predicted the outcomes.

There were 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. 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 experi-
ment 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.\textsuperscript{1}

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 participants 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 recognized one of the candidates (although we still paid participants for accurate responses in these cases). Because essentially all participants were Massachusetts residents at the time of the study, we also exclude from our analysis any Massachusetts elections, although some participants did see clips from these elections.

In the May 2005 sessions, participants knew that they would watch some of the clips with full

\textsuperscript{1}Statistical tests show no difference in participants’ ability to forecast election outcomes across the three rounds of sessions.
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:

- 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 questions 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 answer 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 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.
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 57 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 ($p = 0.003$).

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 economically and statistically significant 0.47 ($p < 0.001$). 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 (0.44) 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 5, 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.

It is possible that the correlation of personal appeal with election outcomes is just an artifact of persistent differences between states. For example, in states where the Democrat regularly wins, the Republicans might tend to run unappealing candidates. Because there are 17 states with multiple elections in our sample, we can address this concern by asking how well participants do at predicting differences across elections within a state. Figure 2 shows that there is a strong relationship between participants’ ratings and actual vote shares, even after we subtract state-level

---

2 We collected data on the candidates and outcomes of the gubernatorial elections in our sample from the *CQ Voting and Elections Collection* (2005).
averages of these variables. Column (2) of table 1 poses this question in a regression form, using state fixed effects to difference out permanent state-specific variation in the Democrat’s share of the two-party vote. 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. The coefficient in this regression is, if anything, somewhat larger than the coefficient in the cross-sectional regression in column (1). Given the relative imprecision of the estimate in column (2), we can only be tentative in generalizing this conclusion out of sample, but it does suggest that participants’ ratings do not merely pick up on permanent differences across states.

To show that participants’ responses are not merely effective in predicting extreme landslides, column (3) of table 1 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 substantially, but remains highly statistically significant. Moreover, the $R^2$ remains essentially the same, dropping slightly from 0.22 to 0.21.

A final interpretational point is that, because the number of participants rating each election is necessarily finite, our key independent variable is measured with error, leading to an attenuation of the relationship we estimate. To check that this bias does indeed weaken our findings, in column (4) 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. Throughout the body of the paper we will conservatively treat our sample-based measures as though they were not subject to sampling variability.\footnote{Using asymptotic theory to estimate the degree of attenuation bias in our estimates, we find that correcting the bias due to a finite number of participants makes only a slight difference in our estimates, changing the estimated coefficient on the share predicting a Democratic victory by less than one percent. This suggests that some of the difference between the elections rated by many participants and the overall sample comes from factors other than measurement error.}

4 Determinants 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. We then present a range of evidence demonstrating that unmeasured candidate familiarity is unlikely to play a role in our findings. Next, we show that participants’ own preferences over gubernatorial candidates are only weakly predictive of electoral success. Finally, we argue that the notion of personal appeal measured by participants’ ratings is highly stable across different debates within an election and, to a lesser extent, different short video clips within the same debate.

4.1 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 the Democratic candidate in the contest. This average is statistically indistinguishable from random guessing ($t = 1.36, p = 0.181$).

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 ($t = 2.83, p = 0.008$). 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: $t = 0.62, p = 0.540$; muddled: $t = -0.43, p = 0.668$). Additionally, although the mean share correctly identifying the Democrat in the silent and muddled conditions cannot be distinguished statistically ($t = 1.21, p = 0.237$), the mean share in the silent condition
is marginally statistically different from that in the full sound condition \((t = 1.99, \ p = 0.055)\), and
the mean share in the muddled condition is highly statistically different from that in the full sound condition \((t = 3.32, \ p = 0.002)\).

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 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, 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. Lastly, 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.

From the standpoint of standard economic theory, the fact that adding information worsens participants’ election forecasts seems paradoxical. However, it is consistent with a range of evidence from psychology. When the relevant source of information is non-verbal, but individuals believe
that verbal content is more informative, the addition of verbal information can worsen judgments. For example, aphasics—who have difficulty interpreting language and must therefore rely more heavily on nonverbal cues—perform better than typical experimental subjects in detecting lying in videotaped monologues (Ettedgui et al., 2000).

The finding that additional policy information worsens judgment has methodological and substantive implications beyond our study. First, from the perspective of measuring personal appeal, it suggests the critical importance of finding naive 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 matters, are often found to perform no better than chance in predicting elections (Tetlock, 1999).

4.2 Candidate Familiarity

We designed our study to rule out the concern that, because election winners are more recognizable than election losers, participants might be able to correctly guess electoral outcomes based on mere familiarity. We never showed a participant a debate from an election in his or her home state, and we exclude elections from the participants’ current state of residence (Massachusetts). We also asked participants to note cases in which they recognized a candidate, and we have excluded those cases from our analysis.

Several pieces of evidence suggest that these steps were successful in removing the influence of candidate familiarity on our results. Most obviously, as we noted in the previous subsection, participants who claimed not to recognize either candidate were unable to do better than random guessing in identifying the party affiliations of the candidates. This 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. This evidence seems difficult to resolve with the view that participants often knew the identities of one or both candidates but chose not to report this to us.

A remaining possibility is that participants were unconsciously familiar with a candidate, but did not report this to us because they themselves were unaware of it. To test for this concern,
we take advantage of the fact that the familiarity of a candidate is likely to be correlated across participants. In table 2, we report regressions on separate samples of elections with below- and above-median candidate familiarity. We measure familiarity by the share of all raters who reported recognizing one or both candidates in the contest. Since our regressions include only participants who did not report recognizing a candidate, these regressions can be interpreted as measuring how the accuracy of a participant’s ratings depend on how often other participants claimed to recognize a candidate in the election. If unmeasured familiarity were a significant concern, we would expect to see greater accuracy in elections in which a larger share of participants reported recognizing a candidate.

Table 2 shows that, if anything, participants’ predictions were somewhat less accurate in elections with highly recognizable candidates. Both the coefficient relating participants’ rating to the two-party vote share, and the $R^2$ of the regression, are lower in the sample of elections with above-median rates of recognition. The relationship between ratings and vote shares is statistically significant in both samples, and the coefficients in the two regressions are not statistically different from one another. Finally, the average share of participants correctly identifying the winner is statistically insignificantly lower in the high-recognition sample than in the low-recognition sample. In light of these findings, it is difficult to believe that unmeasured candidate familiarity significantly confounds our estimates.

4.3 Personal Preferences

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. Most importantly, we asked each participant to tell us how she would vote in a contest between the two candidates depicted in the video clips. We also requested ratings (on a 1-4 scale) of whether each candidate was physically attractive, likeable, and a good leader.

Table 3 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

---

4 We 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.
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 0.24.

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 voting outcomes. Indeed, the one significant exception is the rating of leadership quality, which is highly correlated with both the share predicting the Democrat to win and the share who report that they would vote for the Democrat. The fact that the leadership rating is highly correlated with predicted vote outcomes is consistent with its (relatively) superior performance in predicting actual vote shares in column (1).

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, 82 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.
### 4.4 Candidate-, Debate-, and Clip-specific Factors

By comparing participants’ ratings of different debates from the same election, and of different video clips from the same debate, we can assess the extent to which participants’ judgments of personal appeal are reliable across contexts. Across the 10 elections for which we have two debates, the cross-debate correlation in the share of participants who predict the Democrat to win the election is 0.80, which is economically large and strongly statistically significant ($p = 0.0052$).

Because we created three short video clips for each candidate in each debate (for a total of 9 possible clip pairs per debate), we can conduct the same type of analysis to determine the reliability of participants’ ratings across different video clips. The pairwise correlation between the share predicting the Democrat to win across different video clips from the same debate is a moderate 0.35, and is highly statistically significant ($p < 0.0001$). Because only an average of five participants rated each clip set, this correlation is significantly attenuated by measurement error, and therefore masks a stronger relationship among the population ratings of different clips.

For a more structured decomposition the relative importance of different sources of variance in participants’ ratings, we have estimated a multilevel random-effects model, allowing for election- and debate-level effects in determining the share predicting a Democratic victory at the level of the individual video clip. Of the total variance across all clips of about 0.11, the model assigns about 0.02 to election-level factors and an additional 0.02 to debate-level factors. Using information on the number of participants rating each clip set, we calculate that the measurement error in the ratings accounts for a variance of about 0.03. These estimates therefore suggest that about half of the total variance in participants’ ratings comes from election- and debate-specific factors, with the remainder coming from clip-specific factors.\(^5\)

\(^5\)To estimate the variance due to measurement error, we calculate

$$\frac{\hat{p} (1 - \hat{p})}{N}$$

for each clip set, where $N$ is the number of participants rating the clip set and $\hat{p}$ is the share of participants predicting the Democrat to win. We then compute the average of this quantity across all clip sets in our sample.
5 Predicting Gubernatorial Elections

In this section, we compare the accuracy of forecasts based on participants’ predictions with political and economic factors frequently used in election forecasting. We find that the performance of our measure of personal appeal is far better than economic factors, and comparable to some important political factors, in predicting vote shares in gubernatorial contests.

5.1 Measuring 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. The political factors we measured were incumbency status and campaign spending.

- **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 will be 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. (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.)

- **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 will be the difference in log(expenditure) between the Democrat and Republican. (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

---

6 Experiments with including lagged vote shares, and historical average vote shares, did not reveal any significant improvement in explanatory power from our set of political variables.
spending index.) 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.

Several authors have argued that economic conditions and policies impact elections (see, e.g., Fair, 1978; Peltzman, 1987; Besley and Case, 1995). We will focus on state personal income and unemployment, which play an important role of many studies of economic predictors of elections:

- **Unemployment rate.** We obtained annual data on state unemployment rates from the Bureau of Labor Statistics [http://www.bls.gov/data/]. For each election, we created an index equal to the election year unemployment rate if the incumbent governor is a Democrat, and equal to the negative of the unemployment rate if the incumbent governor is a Republican. 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). Alternative specifications of the unemployment index, including changes in unemployment and measures using unemployment rates relative to state and year means, did not improve the accuracy of our forecasting models.

- **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]. As with the unemployment rate, we constructed an index equal to the log of per capita income when a Democrat is the incumbent governor, and the negative of the log of per capita income when a Republican is the incumbent governor. This approach assumes that the incumbent party is rewarded for favorable economic conditions. We did not find any significant improvement in model fit when we used per capita income relative to state and year averages.

We have also collected data on policy variables such as tax rates, revenues, and expenditures, but including these variables does not tend to improve our ability to forecast vote shares, once we have included the unemployment and income variables.

---

7 We 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 unemployment index as having a value of zero.
5.2 Testing the Predictive Power of Economic, Political, and Personal Factors

Table 4 shows our estimates of the forecasting power of alternative sets of explanatory variables. 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. In particular, we compute the out-of-sample root mean squared error ($RMSE$) 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 ($MSE$) of the model to the $MSE$ of a model including only a constant term. Finally, we compute an out-of-sample $R^2$ as the percentage reduction in $MSE$ 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.

Column (1) of the table repeats the first specification of table 1, which shows that participants’ predictions of victory are strongly related to vote shares, and can explain over 20 percent of the overall variation in the Democrat’s share of the two-party vote. The out-of-sample $R^2$ is also on the order of 20 percent, indicating that a model using participants’ ratings to predict votes out of sample achieves an improvement of 20 percent relative to a model with only a constant term.

We turn to economic predictors of the vote share in column (2). As expected, higher income is associated with greater electoral success, and high unemployment is associated with poorer electoral outcomes. These two variables are jointly marginally statistically significant, but only achieve about a three percent improvement in out-of-sample goodness of fit relative to a constant-only regression. Thus, although economic variables do seem to have predictive power, it is relatively weak compared with the personal factors measured by our laboratory exercise. Our estimates are very consistent with Wolfers’ (2002) finding of a one to three percent adjusted $R^2$ for economic variables in explaining incumbent governors’ electoral performance.

In column (3), we predict vote shares using an index of the incumbency status of the candidates. Being an incumbent results in roughly an 8 percentage point electoral advantage, relative to a state in which neither candidate is an incumbent, 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 24 percent, which indicates that the incumbency index is slightly better than participants’ ratings in predicting vote shares. Interestingly, unreported specifications find
that including our measure of personal appeal reduces the estimated incumbency effect to about 6 percentage points, consistent with Lee’s (2001) argument that standard estimates of incumbency advantage may be biased due to the endogenous selection of incumbents into office.

In column (4), we add a measure of the difference in the log of campaign spending between the two candidates. This specification must be taken with 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 five 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. Adding this variable to the model produces a significant increase in explanatory power, resulting in an out-of-sample $R^2$ of about 37 percent. This level of model fit is significantly larger than the fit from laboratory ratings alone, although this contrast is somewhat difficult to interpret, because the ability to raise funds is likely to be a function of a candidate’s personal appeal, making it difficult to cleanly separate the role of campaign spending from the explanatory power of our laboratory participants’ predictions.

In columns (5) and (6), we include all three sets of variables simultaneously. As column (5) shows, including all factors other than campaign spending results in an out-of-sample $R^2$ of about 34 percent, an improvement of about 14 percentage points over the model with laboratory ratings alone. Interestingly, although the coefficient on participants’ ratings is somewhat lower in this specification, it is still robustly statistically significant despite the inclusion of the other explanatory variables. Column (6) shows that adding campaign spending to the model improves the out-of-sample $R^2$ by about 5 percentage points, which is small but nontrivial. In this last specification, the coefficient on participants’ ratings is significant only at the 10 percent level, although we note that the estimated effects of the other variables are also somewhat lower in the full specification.

Some additional information about the explanatory power of alternative specifications can be gleaned from figure 4, which graphs actual and predicted vote shares based on the models in columns (1), (2), and (4) of table 4. Clearly visible are the powerful relationship between election outcomes and personal factors, the weaker but still apparent relationship between vote shares and economic variables, and the strong relationship with political predictors. In this last case, however, the graph
makes clear that much of the predictive power of political variables comes from landslide elections. Indeed, when we restrict attention to elections in which neither candidate received more than 60 percent of the vote, the fit of the political model decreases substantially, and is much lower than the fit of our laboratory-based predictor.

6 Conclusions

Although a large body of existing research demonstrates that candidates’ policy positions powerfully influence individual voting behavior (e.g., Campbell, Converse, Miller, and Stokes, 1960), we find that ratings based solely on short selections of silent video are highly predictive of vote shares in gubernatorial contests. Hotelling’s (1929) prediction that both candidates in two-party contests will tailor their policy positions to those of the median voter offers a possible resolution of these seemingly contradictory findings. In Hotelling’s model, the winning candidate is chosen, almost by definition, for reasons other than policy positions. Since politicians may be able to choose their policy stands to gain votes much more easily than they can choose their appearance and non-verbal behavior, these latter factors may end up becoming a major determinant of aggregate election outcomes, even if individual-level choices are largely driven by policy preferences.

A causal interpretation of our findings also raises the question of why all candidates for high office are not immensely appealing along the dimensions we measure. We note, however, that while we find evidence suggestive of substantial returns to personal appeal in political markets, these attributes may also bring significant returns in the private labor market (e.g., Biddle and Hamermesh, 1998). 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


<table>
<thead>
<tr>
<th>Sample</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Share predicting a Democrat victory</td>
<td>0.2452</td>
<td>0.2840</td>
<td>0.1301</td>
<td>0.2795</td>
</tr>
<tr>
<td>(0.0613)</td>
<td>(0.1144)</td>
<td>(0.0410)</td>
<td>(0.0593)</td>
<td></td>
</tr>
<tr>
<td>Sample</td>
<td>All</td>
<td>More than one</td>
<td>Competitive</td>
<td>More than 30</td>
</tr>
<tr>
<td>election in state</td>
<td></td>
<td></td>
<td>elections</td>
<td>raters</td>
</tr>
<tr>
<td>State fixed effects?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.2225</td>
<td>0.5973</td>
<td>0.2092</td>
<td>0.3075</td>
</tr>
<tr>
<td>$N$</td>
<td>58</td>
<td>37</td>
<td>40</td>
<td>52</td>
</tr>
</tbody>
</table>

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. “Competitive elections” refers to elections in which neither candidate received more than 60 percent of the vote. “More than 30 raters” refers to elections that were viewed by over 30 study participants. All calculations exclude respondents who claimed to recognize one or both of the candidates.
Table 2  Candidate recognizability and participants’ forecast accuracy

Dependent variable: Democrat share of two-party vote

<table>
<thead>
<tr>
<th>Sample</th>
<th>Share recognizing a candidate:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Below-median</td>
<td>Above-median</td>
<td>Difference</td>
<td></td>
</tr>
<tr>
<td>Share predicting a</td>
<td>0.2647</td>
<td>0.1985</td>
<td>-0.0662</td>
<td></td>
</tr>
<tr>
<td>Democrat victory</td>
<td>(0.0795)</td>
<td>(0.0956)</td>
<td>(0.1240)</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.2908</td>
<td>0.1376</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>29</td>
<td>29</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>Average share correctly guessing candidates’ political parties</td>
<td>0.5851</td>
<td>0.5605</td>
<td>-0.0247</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0346)</td>
<td>(0.0338)</td>
<td>(0.0484)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: The first two columns show results of OLS regressions of Democrat share of two-party vote on the share of participants predicting a Democrat victory. The first column shows results for elections in which a below-median share of participants rating the election claimed to recognize one or both of the candidates. The second column shows results for elections in which an above-median share of participants rating the election claimed to recognize one or both of the candidates. However, all calculations exclude respondents who claimed to recognize one or both of the candidates. The difference between the regression coefficients in the two samples is calculated using a fully interacted model that nests the specifications in the first two columns.
Table 3  Predictions, preferences, and electoral outcomes

Correlation matrix (p-values in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat’s share of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>two-party vote</td>
<td>0.4717</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Share predicting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Democrat to win</td>
<td>(0.0002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share who would</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vote for Democrat</td>
<td>0.0891</td>
<td>0.3078</td>
<td>—</td>
</tr>
<tr>
<td>Average difference in</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ratings of candidate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>as:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physically attractive</td>
<td>0.1626</td>
<td>0.2851</td>
<td>0.7939</td>
</tr>
<tr>
<td></td>
<td>(0.2227)</td>
<td>(0.0301)</td>
<td>(&lt;0.0001)</td>
</tr>
<tr>
<td>Likeable</td>
<td>0.1224</td>
<td>0.2965</td>
<td>0.8838</td>
</tr>
<tr>
<td></td>
<td>(0.3602)</td>
<td>(0.0238)</td>
<td>(&lt;0.0001)</td>
</tr>
<tr>
<td>Good leader</td>
<td>0.2392</td>
<td>0.6455</td>
<td>0.7328</td>
</tr>
<tr>
<td></td>
<td>(0.0706)</td>
<td>(&lt;0.0001)</td>
<td>(&lt;0.0001)</td>
</tr>
</tbody>
</table>

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.
### Table 4 Political, economic, and personal factors as election predictors

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

<table>
<thead>
<tr>
<th>Predictor</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Share predicting a Democrat victory</td>
<td>0.2452</td>
<td></td>
<td></td>
<td>0.1450</td>
<td>0.1049</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0613)</td>
<td></td>
<td></td>
<td>(0.0588)</td>
<td>(0.0586)</td>
<td></td>
</tr>
<tr>
<td>Index of log(state personal income)</td>
<td>0.0132</td>
<td></td>
<td>0.0101</td>
<td>0.0047</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0063)</td>
<td></td>
<td>(0.0051)</td>
<td>(0.0054)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Index of unemployment rate</td>
<td>-0.0206</td>
<td></td>
<td>-0.0227</td>
<td>-0.0130</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0115)</td>
<td></td>
<td>(0.0092)</td>
<td>(0.0097)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference in incumbency status</td>
<td></td>
<td>0.0764</td>
<td>0.0411</td>
<td>0.0823</td>
<td>0.0623</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0168)</td>
<td>(0.0179)</td>
<td>(0.0218)</td>
<td>(0.0224)</td>
<td></td>
</tr>
<tr>
<td>Difference in log(campaign spending)</td>
<td></td>
<td>0.0481</td>
<td></td>
<td>0.0348</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0131)</td>
<td></td>
<td>(0.0143)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|          | 0.2225   | 0.0925   | 0.2706  | 0.4140   | 0.4408  | 0.4985   |
|          | 58       | 58       | 58      | 58       | 58      | 58       |
| Out-of-sample $RMSE$                 | 0.0924   | 0.1019   | 0.0901  | 0.0818   | 0.0839  | 0.0810   |
| Out-of-sample $R^2$                  | 0.2001   | 0.0263   | 0.2389  | 0.3720   | 0.3394  | 0.3855   |

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 log(state personal income)” is equal to the log of state personal income if the incumbent governor at the time of the election was a Democrat, equal to the negative 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. “Index of unemployment” is defined analogously. “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.
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

Figure A: Ability to guess candidate party by sound condition

![Graph showing the effect of sound condition on the percent correctly identifying the Democrat party. The x-axis represents sound conditions: Silent, Muddled, Full sound. The y-axis represents percent correctly identifying Democrat. Error bars indicate ±1 standard error.](image)

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

![Graph showing the effect of sound condition on the percent correctly predicting the winner of a contest. The x-axis represents sound conditions: Silent, Muddled, Full sound. The y-axis represents percent correctly predicting the winner. Error bars indicate ±1 standard error.](image)

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.
**Figure 4** Visual comparisons of alternative forecasting models