

The Mythical Swing Voter

Andrew Gelman¹, Sharad Goel², Douglas Rivers³ and David Rothschild^{4*}

¹*Columbia University, USA; gelman@stat.columbia.edu*

²*Stanford University, USA; scgoel@stanford.edu*

³*Stanford University, USA; rivers@stanford.edu*

⁴*Microsoft Research, USA; davidmr@microsoft.com*

ABSTRACT

Most surveys conducted during the 2012 U.S. presidential campaign showed large swings in support for the Democratic and Republican candidates, especially before and after the first presidential debate. Using a combination of traditional cross-sectional surveys, a unique panel survey (in terms of scale, frequency, and source), and a high response rate panel, we find that daily sample composition varied more in response to campaign events than did vote intentions. Multilevel regression and post-stratification (MRP) is used to correct for this selection bias. Demographic post-stratification, similar to that used in most academic and media polls, is inadequate, but the addition of attitudinal variables (party identification, ideological self-placement, and past vote) appears to make selection ignorable in our data. We conclude that vote swings in 2012 were mostly sample artifacts and that real swings were quite small. While this account is at odds with most contemporaneous analyses,

*We thank Jake Hofman, Neil Malhotra, and Duncan Watts for their comments, and the National Science Foundation for partial support for this research. We also thank the audiences at MPSA, AAPOR, Toulouse Network for Information

Supplementary Material available from:

http://dx.doi.org/10.1561/100.00015031_supp

MS submitted on 26 February 2015; final version received 30 November 2015

ISSN 1554-0626; DOI 10.1561/100.00015031

© 2016 A. Gelman, S. Goel, D. Rivers and D. Rothschild

it better corresponds with our understanding of partisan polarization in modern American politics.

Keywords: Elections; swing voters; multilevel regression and post-stratification

1 Introduction

In a competitive political environment, a relatively small number of voters can shift control of Congress and the Presidency from one party to the other, or to divided government. Polls do indeed show substantial variation in voting intentions over the course of campaigns. This suggests that swing voters are key to understanding the changing fortunes of Democrats and Republicans in recent national elections. This is certainly the view of political professionals and media observers. Campaigns spend enormous sums — over \$2.6 billion in the 2012 presidential election cycle — trying to target “persuadable voters.” Poll aggregators track day-to-day swings in the proportion of voters supporting each candidate. Political scientists have debated whether swings in the polls are a response to campaign events or are reversions to predictable positions as voters become more informed about the candidates (Gelman and King, 1993; Hillygus and Jackman, 2003; Kaplan *et al.*, 2012). Both researchers and campaign participants seem to agree that polls accurately measure vote intentions and that these are malleable. While there is disagreement about the causes of swings, no one appears to have questioned their existence.

But there is a puzzle: candidates appeal to swing voters in debates, campaigns target advertising toward swing voters, journalists discuss swing voters, and the polls do indeed swing — but it is hard to find voters who have actually switched sides. Partly this is because most polls are based on independent cross-sections of respondents, and so vote switching cannot be directly observed.¹ But there are also theoretical reasons to be skeptical about the degree of volatility found in

Technology, Stanford, Microsoft Research, University of Pennsylvania, Duke, and Santa Clara for their feedback during talks on this work.

¹Individual-level changes must be inferred from aggregate shifts in candidate preference between polls, and this inference depends upon the assumption that the

election polls. If, as is widely agreed, there is a high degree of partisan polarization in the American electorate, it seems implausible that many voters will switch support from one party to the other because of minor campaign events (Baldassarri and Gelman, 2008; Fiorina and Abrams, 2008; Levendusky, 2009).²

In this paper we focus on apparent vote shifts surrounding the debates between Barack Obama and Mitt Romney during the 2012 U.S. presidential election campaign. We argue that the apparent swings in vote intention represent mostly changes in sample composition — not changes in opinion — and that these “phantom swings” arise from sample selection bias in survey participation. To make this case, we draw on three sources of evidence: (1) traditional cross-sectional surveys; (2) a novel large-scale panel survey; and (3) the RAND American Life Panel. Previous studies have tended to assume that campaign events cause changes in vote intentions, while ignoring the possibility that they may cause changes in survey participation. We show that in 2012, campaign events were more strongly correlated with changes in survey participation than with changes in vote intention. As a consequence, inferences about the impact of campaign events from changes in polling averages involve invalid sample comparisons, similar to uncontrolled differences among treatment groups.

We further show how one can correct for this sample bias. If survey variables such as vote intention are independent of sample selection conditional upon a set of covariates, various methods can be used to obtain consistent estimates of population parameters. Using the method of multilevel regression and post-stratification (MRP), we show that conditioning upon standard demographics (age, race, gender, education) is inadequate to remove the selection bias present in our data. However, the introduction of controls for party ID, ideology, and past vote among the covariates appears to substantially eliminate selection effects. While

sample selection mechanism does not change at the same time. Of course, vote switching is directly observable in panel data, but there are few election panels with multiple interviews of the same respondents, and even fewer panels are large enough to provide reliable estimates of even moderate-sized vote swings.

²To be clear, we are discussing *net* change in support for candidates. Panel surveys show much larger amounts of *gross* change in vote intention between waves which are offset by changes in the opposite direction. See, for example, Table 7.1 of Sides and Vavreck (2014).

the use of party ID weighting is controversial in cross-sectional studies (Allsop and Weisberg, 1988; Kaminska and Barnes, 2008), most of these problems can be avoided in a panel design.³ In panels, post-stratification on baseline attitudes avoids endogeneity problems associated with cross-sectional party ID weighting, even if these attitudes are not stable over the campaign.

2 Evidence from Empirical Studies

2.1 *Study 1: Sample Selection Bias in Cross-sectional Polls*

In mid-September, Obama led Romney by about 4% in the *Huffington Post* polling average and seemed to be coasting to an easy reelection victory. However, as shown in Figure 1a, following the first presidential debate on October 3, the polls reversed and Romney led in nine of the twelve polls conducted in the following week (of the remaining three, one was a tie and Obama led in the other two). On average, Romney led Obama by slightly over 1% in the polls with field periods starting between October 4 and 8. It was not until after the third debate (on October 22) that Obama regained a small lead in the polling averages, which he maintained until election day. At the time, it was commonly agreed that Obama had performed poorly in the first presidential debate but had recovered in later debates. This account is consistent with the existence of a pool of swing voters who switched back and forth between the candidates.

However, other data from the same surveys cast doubt on the claim that the first presidential debate caused a swing of this magnitude. Consider, for example, the Pew Research surveys. In the September 12–16 Pew survey, Obama led Romney 51–42 among registered voters, but the two candidates were tied 46–46 in the October 4–7 survey. The 5% swing to Romney sounds impressive until it is compared to how the same respondents recalled voting in 2008. In the September 12–16 sample, 47% recalled voting for Obama in 2008, but this dropped to 42% in the October 4–7 sample. Recalled vote for McCain also rose

³See Reilly *et al.* (2001) for a potential work-around in cross-sectional studies.

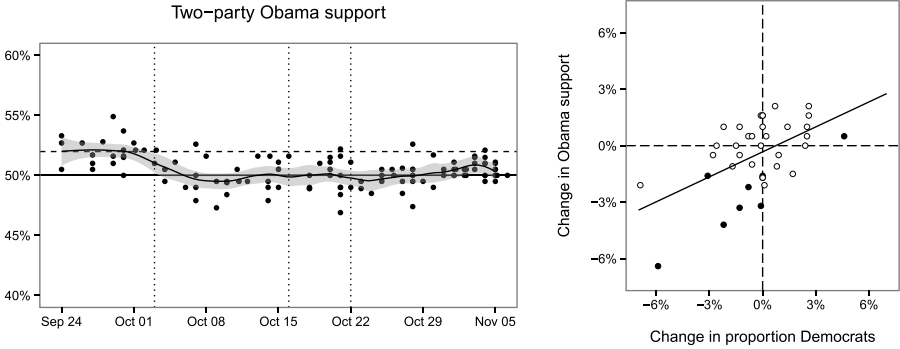


Figure 1: (a) Estimated support for Obama (among those who support either Obama or Romney), as reported in major media polls, where each point corresponds to a single poll. The dashed horizontal line indicates the final vote share, and the dotted vertical lines indicate the three presidential debates. (b) The change in two-party support for Obama versus the change in the fraction of respondents who identify as Democrats. Each point indicates the reported change in consecutive polls conducted before and after the October 3 debate by the same polling organization; the solid points correspond to polls that were fielded within 20 days both before and after the debate. The solid line is the regression line, and the positive correlation indicates that observed swings in the polls correspond to swings in the proportion of Democrats or Republicans who respond to the polls. The figure illustrates how the sharp drop in measured support for Obama around the first debate (Panel A) is strikingly correlated with a drop in the fraction of Democrats responding to major media polls (Panel B).

by 5% in this pair of surveys (from 32% to 37%). The swing toward Romney in the two polls was identical to the increase in recalled voting for McCain.

Similarly, Figure 1b shows that throughout the election cycle and across polling organizations, Obama support is positively correlated with the proportion of survey respondents who say they are Democrats. Each point in the plot represents a pair of consecutive surveys conducted by the same polling organization before and after the first presidential debate. The scatterplot compares the change in two-party Obama support to the change in the proportion of respondents who self-identify as Democrats in the pre- and post-debate surveys. The estimated support for Obama is positively correlated with the proportion of Democratic party identifiers in each sample. For the subset of polls that

were in the field within 20 days before and after the debate (indicated by the solid points), the effect is even more pronounced.

There are at least two potential explanations for these patterns in the data. One possibility is that the debate changed people's voting intentions, their memory of how they had voted in the previous election, and their party identification (Himmelweit *et al.*, 1978). Or, alternatively, the samples before and after the first debate were different (i.e., the pre-debate surveys contained more Democrats and 2008 Obama voters, while the ones afterward contained more Republicans and 2008 McCain voters).

It is impossible to distinguish between these explanations using cross-sectional data. Respondents in the September and October Pew samples do not overlap, so we cannot tell whether more of the September respondents would have supported Romney if they had been reinterviewed in October. The October interviews are with a different sample and, while more say they intend to vote for Romney than those in the September sample, we do not know whether these respondents were less supportive of Romney in September, since they were not interviewed in September.

2.2 Study 2: The Xbox Panel Survey

2.2.1 Survey Design and Methodology

We address the shortcomings of cross-sectional surveys discussed above by fielding a large-scale online panel survey. During the 2012 U.S. presidential campaign, we conducted 750,148 interviews with 345,858 unique respondents on the Xbox gaming platform during the 45 days preceding the election. Xbox Live subscribers were asked to provide baseline information about themselves in a registration survey, including demographics, party identification, and ideological self-placement. Each day, a new survey was offered and respondents could choose whether they wished to complete it. The analysis reported here is based upon the 83,283 users who responded at least once prior to the first presidential debate on October 3. In total, these respondents completed 336,805 interviews, or an average of about four interviews per respondent. Over 20,000 panelists completed at least five interviews and over 5,000 answered surveys on 15 or more days. The average number of respondents in our analysis sample each day was about 7,500. The Xbox panel

provides abundant data on actual shifts in vote intention by a particular set of voters during the 2012 presidential campaign, and the size of the Xbox panel supports estimation of MRP models which adjust for different types of selection bias.

Our analysis has two steps. We first show that with demographic adjustments, the Xbox data reproduce swings found in media polls during the 2012 campaign. That is, if one adjusts for the variables typically used for weighting phone or Internet samples, daily Xbox surveys exhibit the same sort of patterns found in conventional polls. Second, because the Xbox data come from a panel with baseline measurements of party ID and other attitudes, it is feasible to correct for variations in survey participation due to partisanship, ideology, and past vote. The correlation of within-panel response rates with party ID, for example, varies over the course of the campaign. Using MRP with an expanded set of covariates enables us to distinguish between actual vote swings and compositional changes in daily samples. With these adjustments, most of the apparent swings in vote intention disappear.

The Xbox panel is not representative of the electorate, with Xbox respondents predominantly young and male. As shown in Figure 2,

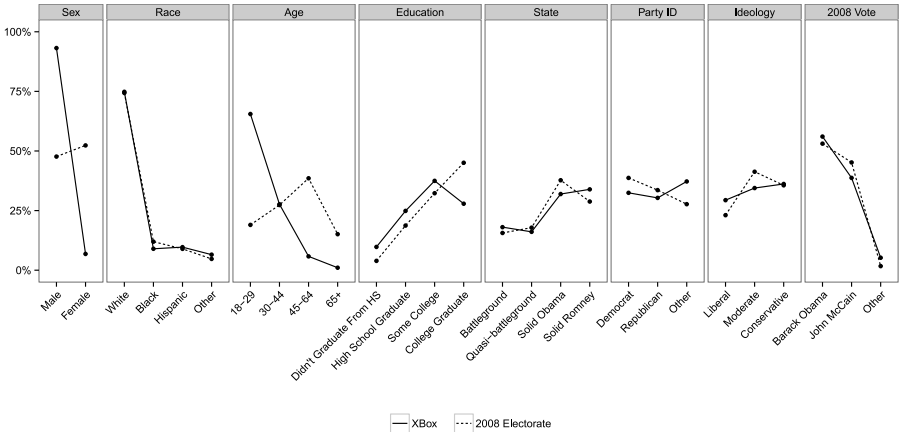


Figure 2: Demographic and partisan composition of the Xbox panel and the 2008 electorate. There are large differences in the age distribution and gender composition of the Xbox panel and the 2012 exit poll. Without adjustment, Xbox data consistently overstate support for Romney. However, the large size of the Xbox panel permits satisfactory adjustment even for large skews.

66% of Xbox panelists are between 18 and 29 years old, compared to only 18% of respondents in the 2008 exit poll,⁴ while men make up 93% of Xbox panelists but only 47% of voters in the exit poll. With a typical-sized sample of 1,000 or so, it would be difficult to correct skews this large, but the scale of the Xbox panel compensates for its many sins. For example, despite the small proportion of women among Xbox panelists, there are over 5,000 women in our sample, which is an order of magnitude more than the number of women in an RDD sample of 1,000.

The method of MRP is described in Gelman and Little (1997). Briefly, post-stratification is a standard framework for correcting for known differences between sample and target populations (Little, 1993). The idea is to partition the population into cells (defined by the cross-classification of various attributes of respondents), use the sample to estimate the mean of a survey variable within each cell, and finally to aggregate the cell-level estimates by weighting each cell by its proportion in the population. In conventional post-stratification, cell means are estimated using the sample mean within each cell. This estimate is unbiased if selection is ignorable (i.e., if sample selection is independent of survey variables conditional upon the variables defining the post-stratification). In other words, the key assumption is that within each cell individuals who partake in the survey have vote choices suitably similar to those who choose not to take the survey to make them feasible substitutes for the non-responders. This ignorability assumption is more plausible if more variables are conditioned upon. However, adding more variables to the post-stratification increases the number of cells at an exponential rate. If any cell is empty in the sample (which is guaranteed to occur if the number of cells exceeds the sample size), then the conventional post-stratification estimator is not defined. Even if every cell is nonempty, there can still be problems because estimates of cell means are noisy in small cells. Collapsing cells reduces variability, but can leave substantial amounts of selection bias. MRP addresses this problem by using hierarchical Bayesian regression to obtain stable

⁴As discussed later, we chose to use the 2008 exit poll data for post-stratification so that the analysis relies only upon information available before the 2012 election. Relying upon 2008 data demonstrates the feasibility of this approach for forecasting. Similar results are obtained by post-stratifying on 2012 exit poll demographics and attitudes.

estimates of cell means (Gelman and Hill, 2006). This technique has been successfully used in the study of public opinion and voting (Ghitza and Gelman, 2013; Lax and Phillips, 2009).

We initially apply MRP by partitioning the population into 6,258 cells based upon demographics and state of residence (2 gender \times 4 race \times 4 age \times 4 education \times 50 states plus the District of Columbia).⁵ One cell, for example, corresponds to 30- to 44-year-old white male college graduates living in California. Using each day's sample, we then fit separate multilevel logistic regression models that predict respondents' stated vote intention on that day as a function of their demographic attributes. Key to our analysis is that cells means (i.e., average vote intention) on any given day are accurately estimated by the regression models. We evaluate this assumption in the Appendix and find the model indeed generates accurate group-level estimates despite being based on a non-representative sample of respondents. We additionally assume that the distribution of voter demographics for each state would be the same as that found in the 2008 exit poll. See the Appendix for additional details on modeling and methods.

2.2.2 Xbox Panel Results

Figure 3a shows the estimated daily proportion of voters intending to vote for Obama (excluding minor party voters and non-voters).⁶ After adjustment for demographics by MRP, the daily Xbox estimates of voting intention are quite similar to daily polling averages from media polls shown in Figure 1. In particular, the most striking feature of this time series is the precipitous decline in Obama's support following the first presidential debate on October 3 (indicated by the left-most dotted vertical line). This swing was widely interpreted as a real and important shift in vote intentions. For example, Nate Silver wrote in the *New York Times* on October 6, "Mr. Romney has not only improved his own standing but also taken voters away from Mr. Obama's column," and

⁵The survey system used for the Xbox project was limited to four response options per question, except for state of residence, which used a text box for input.

⁶We smooth the estimates over a four-day moving window, matching the typical duration for which standard telephone polls were in the field in the 2012 election cycle.

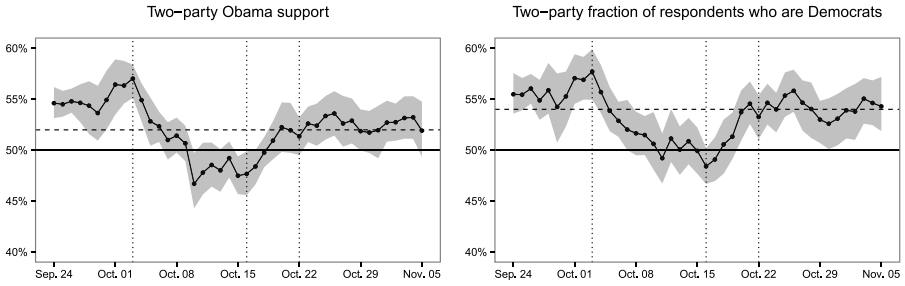


Figure 3: (a) Among respondents who support either Barack Obama or Mitt Romney, estimated support for Obama (with 95% confidence bands), adjusted for demographics. The dashed horizontal line indicates the final vote share, and the dotted vertical lines indicate the three presidential debates. This demographically adjusted series is a close match to what was obtained by national media polls during this period. (b) Among respondents who report affiliation with one of the two major parties, the estimated proportion who identify as Democrats (with 95% confidence bands), adjusted for demographics. The dashed horizontal lines indicate the final party identification share, and the dotted vertical lines indicate the three presidential debates. The pattern in the two figures is strikingly similar, suggesting that most of the apparent changes in public opinion are actually artifacts of differential nonresponse.

Karl Rove declared in the *Wall Street Journal* the following day, “Mr. Romney’s bounce is significant.”

But was the swing in Romney support in the polls real? Figure 3b shows the daily proportion of respondents, after adjusting for demographics, who say they are Democrats or Republicans (omitting independents). For the two weeks following the first debate, Democrats were simply much less likely than Republicans to participate in the survey, even after adjustment for demographic differences in the daily samples. For example, among 30- to 44-year-old white male college graduates living in California, more of the respondents were self-identified Republicans after the debate than in the days leading up to it. Demographic adjustment alone is inadequate to correct selection bias due to partisanship.

An important methodological concern is the potential endogeneity of attitudinal variables, such as party ID, in voting decisions. If some respondents change their party identification and vote intention simultaneously, then using current party ID to weight a cross-sectional survey to a past party ID benchmark is both inaccurate and arbitrary. This

problem has deterred most media polls from using party ID for weighting. The approach used here, however, avoids the endogeneity problem because we are adjusting *past* party ID to a *past* party ID benchmark. That is, current vote intention is post-stratified on a pre-determined variable (baseline party ID) that does not change over the course of the panel.

The other objection to post-stratification on partisanship is that, unlike demographics (where we have Census data), we lack reliable benchmarks for its baseline distribution. This is less of a problem than it might seem. First, the approximate distribution of party ID can be obtained from other surveys. In our analysis, we used the 2008 exit poll for the joint distribution of all variables. Second, the swing estimates are not particularly sensitive to which baseline is used, since swings are similar within the different party ID groups. The party ID benchmark has a larger impact on the estimated candidate lead, but even this does not vary a lot within the range of plausible party ID distributions. In the Appendix, we compare estimates based upon covariate distributions from the 2008 and 2012 exit polls, and find the two lead to similar results.

In Figure 4, we compare MRP adjustments using only demographics (shown in light gray) and both demographic and attitudinal variables (a black line with dark gray confidence bounds). The additional attitudinal variables used for post-stratification were party identification (Democratic, Republican, Independent, and other), ideology (liberal, moderate, and conservative), and 2008 presidential vote (Obama, McCain, “other”, and “did not vote”). Again, we applied MRP to adjust the daily samples for selection bias, but now the adjustment allows for selection correlated with both attitudinal and demographic variables.

In Figure 4, the swings shown in Figure 3 largely disappear. The addition of attitudinal variables in the MRP model corrects for differential response rates by party ID and other attitudinal variables at different points in the campaign. Compared to the demographic-only post-stratification (shown in gray), post-stratification on both demographics and party ID greatly reduces (but does not entirely eliminate) the swings in vote intention after the first presidential debate. Adjusting only for demographics yields a six-point drop in support for Obama in the four days following the first presidential debate; adjusting for both demographics and partisanship reduces the drop in support for

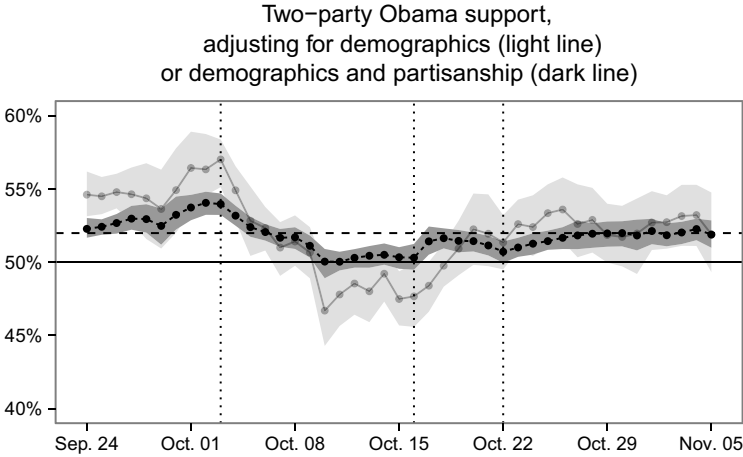


Figure 4: Obama share of the two-party vote preference (with 95% confidence bands) estimated from the Xbox panel under two different post-stratification models: the dark line shows results after adjusting for both demographics and partisanship, and the light line adjusts only for demographics (identical to Figure 3a). The surveys adjusted for partisanship show less than half the variation of the surveys adjusted for demographics alone, suggesting that most of the apparent changes in support during this period were artifacts of partisan nonresponse.

Obama to between 2% and 3% percent. More generally, adjusting for partisanship reduces swings by more than 50% compared to adjusting for demographics alone. In the demographics-only post-stratification, Romney takes a small lead following the first debate (similar to that observed in contemporaneous media polls). In contrast, the demographics and party ID adjustment leave Obama with a lead throughout the campaign. Correctly estimated, most of the apparent swings were sample artifacts, not actual change.

Next, in Figure 5, we consider estimated swings around the debate within demographic and partisan groups. Not surprisingly, the small net change that does occur is concentrated among independents, moderates, and those who did not vote in 2008. Of the relatively few supporters gained by Romney, the majority were previously undecided.

To this point, we have focused on net changes in voting intention for Obama over Romney and found, after correcting for partisan nonresponse, a nearly stable lead for Obama throughout the 2012 election campaign. This result is, in principle, consistent with two competing

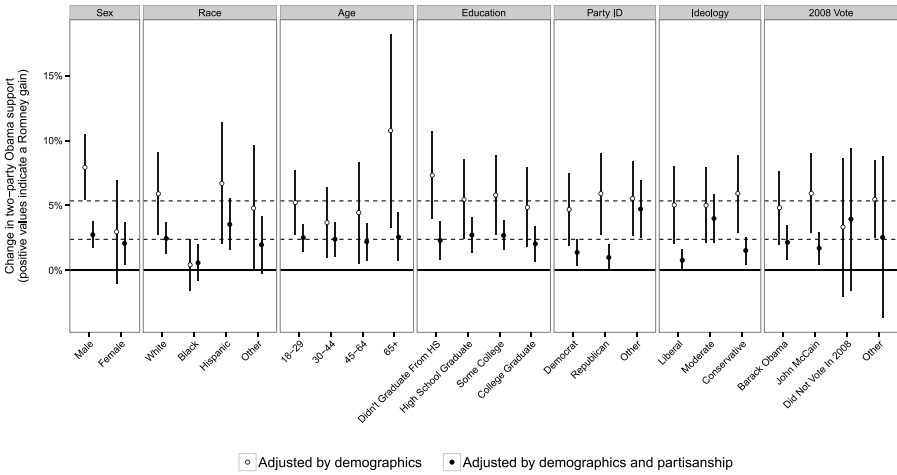


Figure 5: Estimated swings in two-party Obama support between the day before and four days after the first presidential debate under two different post-stratification models, separated by subpopulation. The vertical lines represent the overall average movement under the each model. The horizontal lines correspond to 95% confidence intervals.

hypotheses. One possibility is that relatively large numbers of supporters for both candidates may have switched their vote intention, resulting in little net movement; the other is that only a relatively small number of individuals may have changed their allegiance. Furthermore, gross changes could be from intending to vote for one major-party candidate and changing to intending to vote for the other, or from switching one’s support from a major-party candidate to “other”. We conclude our analysis by examining individual-level changes of opinion around the first presidential debate.

Figure 6 shows, as one may have expected, that only a small percentage of individuals (3%) switched their support. Notably, the largest fraction of switches results from individuals who supported Obama prior to the first debate and then switched their support to “other.” Here “other” incorporates both undecideds and third party voters, who were negligible in 2012 and who may have ended up abstaining or even supporting Obama. On the other hand, only 0.5% of panelists switched from Obama to Romney in the weeks around the first debate, with 0.2% switching from Romney to Obama. Contrary to most popular accounts

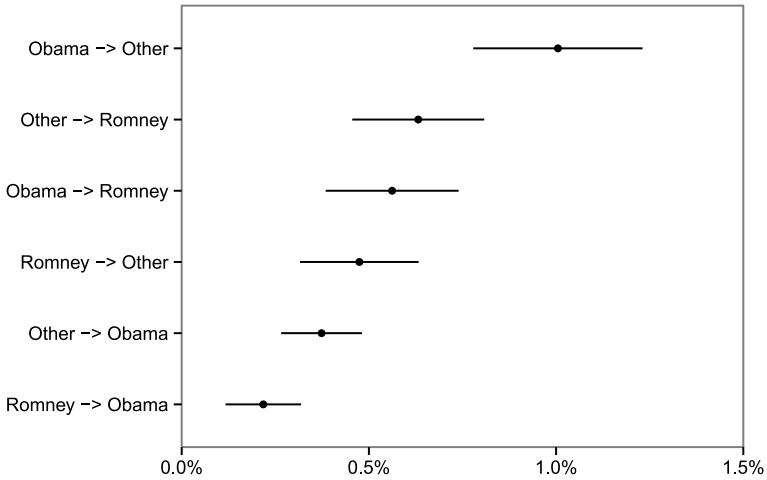


Figure 6: Estimated proportion of the electorate that switched their support from one candidate to another during the one week immediately before and after the first presidential debate, with 95% confidence intervals. We find that only 0.5% of individuals switched their support from Obama to Romney.

of the campaign, the Xbox panel shows little evidence of Romney picking up support from Obama voters after the first debate.

2.3 Study 3: The RAND American Life Panel

The daily participation rate of Xbox panelists is similar to the response rate for RDD media polls. In contrast, some Internet panels have much higher rates of participation. For example, the RAND American Life Panel (ALP) pays respondents two dollars for each completed interview and achieves an impressive 80% response rate. This means that unlike the Xbox panel or most phone surveys, sample composition is much more stable between waves and interwave selection effects are minimal. To test our claims further, we examine results from the RAND Continuous 2012 Presidential Election Poll. Starting in July 2012, RAND polled a fixed panel of 3,666 people each week, asking each participant the likelihood he or she would vote for each presidential candidate (for example, a respondent could specify 60% likelihood to vote for Obama, 35% likelihood for Romney, and 5% likelihood for someone else). Participants were additionally asked how likely they were to vote

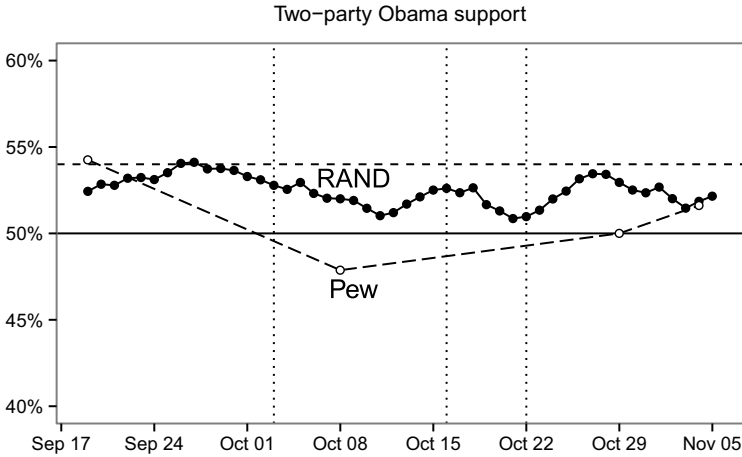


Figure 7: Support for Obama (among respondents who expressed support for one of the two major-party candidates) as reported by the RAND (solid line) and Pew Research (dashed line) surveys. The dashed horizontal line indicates the final vote share, and the dotted vertical lines indicate the three presidential debates. As with most traditional surveys (see Figure 1), the Pew poll indicates a substantial drop in support for Obama after the first debate. However, the high response rate RAND panel, which should not be susceptible to partisan nonresponse, shows much smaller swings.

in the election, and their assessed probability of Obama winning the election. Each day, one-seventh of the panel (approximately 500 people) were prompted to answer these three questions and had seven days to respond — though in practice, most responded immediately (Gutsche *et al.*, 2014).

Comparing changes in the RAND panel with those in cross-sectional surveys provides a rough estimate of how much of the measured drop in Obama’s support after the first debate is due to selection bias. Although the RAND panel may be biased in estimating the *level* of Obama’s support, the high within-panel response rate (about 80%) means that between-wave selection bias is relatively small. In contrast, Pew reports that its typical response rate is under 10%, so that the potential for differential selection processes between surveys is large. Figure 7 compares estimated two-party vote share for Obama in the RAND and Pew surveys during the 2012 campaign. The solid line reproduces the results of the RAND survey as reported by Gutsche

et al. (2014), where each point represents a seven-day rolling average.⁷ The RAND estimate shows a low of 51% in Obama support occurring in the days after the first debate. This estimate is nearly identical to the Xbox estimate of 50%. In contrast, Pew shows Obama support dropping from 55% to 48%.

3 Discussion

By considering three qualitatively different sources of evidence — traditional cross-sectional surveys, a large-scale opt-in panel, and a high response rate panel — we find that much of the apparent swings in vote intention can be explained by sample selection bias. In panel surveys, even ones with low response rates, real population changes can be inferred by post-stratifying on attitudinal variables measured at the start of the panel. In cross-sectional survey designs, it can be difficult to correct for selection bias without assuming that attitudinal variables do not fluctuate over time. Though the proportion of Democrats and Republicans in presidential election exit polls is quite stable, there is also evidence that party ID does fluctuate somewhat between elections. This makes cross-sectional party ID corrections controversial, but the failure to adjust sample composition for anything other than demographics should be equally controversial. Methods exist for such adjustment, making use of the assumption that the post-stratifying variable (in this case, party identification) evolves slowly (Reilly *et al.*, 2001). But even the naive approach of post-stratifying on current partisanship works reasonably well (see the Appendix for details).

We have not treated the problem of turnout. Likelihood of voting may vary over the campaign and the proclivity to take a survey could be an indicator of likelihood to vote. Consequently, it is possible that cross-sectional poll estimates could be good predictors of actual vote, even if they are misleading about changes in preference. This argument is speculative. In fact, as seen in Figure 3b, the relative

⁷Whereas Gutsche *et al.* (2014) separately plot support for Obama and Romney, we combine these two into a single line indicating two-party Obama support; we otherwise make no adjustments to their reported numbers. The estimated number of votes for each candidate is based on one's stated likelihood of voting, and one's stated likelihood of voting for each candidate conditional on voting.

dearth of Democratic sample respondents was short-lived. By the third debate, there were as many Democrats participating in Xbox surveys as there had been before the first. Furthermore, this runs counter to the remarkable stability of the partisan composition of the electorate: in every presidential election from 1984 to 2012, Democrats have comprised between 37% and 39% of voters, and men have comprised between 46% and 48% of voters.

The temptation to over-interpret bumps in election polls can be difficult to resist, so our findings provide a cautionary tale. The existence of a pivotal set of voters attentively listening to the presidential debates and switching sides is a much more satisfying narrative, both to pollsters and survey researchers, than a small, but persistent, set of sample selection biases. Correcting for these biases gives us a picture of public opinion and voting that corresponds better with our understanding of the intense partisan polarization in modern American politics.

A Methods and Materials

Xbox survey.

The only way to answer the polling questions was via the Xbox Live gaming platform. There was no invitation or permanent link to the poll, and so respondents had to locate it daily on the Xbox Live's home page and click into it. The first time a respondent opted-into the poll, they were directed to answer the nine demographics questions listed below. On all subsequent times, respondents were immediately directed to answer between three and five daily survey questions, one of which was always the vote intention question.

Intention Question: If the election were held today, who would you vote for?

Barack Obama\Mitt Romney\Other\Not Sure

Demographics Questions:

1. Who did you vote for in the 2008 Presidential election?
Barack Obama\John McCain\Other candidate\Did not vote in 2008

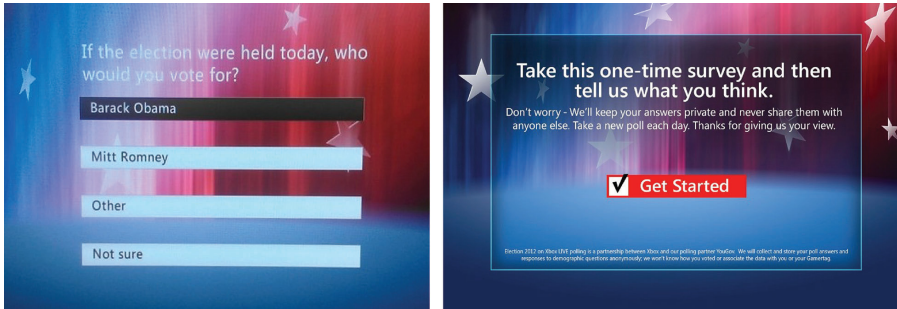


Figure A.1: The left panel shows the vote intention question, and the right panel shows what respondents were presented with during their first visit to the poll.

2. Thinking about politics these days, how would you describe your own political viewpoint?
Liberal\Moderate\Conservative\Not sure
3. Generally speaking, do you think of yourself as a . . . ?
Democrat\Republican\Independent\Other
4. Are you currently registered to vote?
Yes\No\Not sure
5. Are you male or female?
Male\Female
6. What is the highest level of education that you have completed?
Did not graduate from high school\High school graduate\Some college or 2-year college degree\4-year college degree or Postgraduate degree
7. What state do you live in?
Dropdown menu with states — listed alphabetically; including District of Columbia and “None of the above”
8. In what year were you born?
1947 or earlier\1948–1967\1968–1982\1983–1994
9. What is your race or ethnic group?
White\Black\Hispanic\Other

Demographic post-stratification.

We used multilevel regression and post-stratification (MRP) to produce daily estimates of candidate support. For each date d between September 24, 2012 and November 5, 2012, define the set of responses R_d to be those submitted on date d or on any of the three prior days. Daily estimates — which were smoothed over a four-day moving window — are generated by repeating the following MRP procedure separately on each subset of responses R_d . In the first step (multilevel regression), we fit two multilevel logistic regression models to predict panelists' vote intentions (Obama, Romney, or "other") as a function of their age, sex, race, education, and state. Each of these predictors is categorical: age (18–29, 30–44, 45–64, or 65 and older), sex (male or female), race (white, black, Hispanic or other), education (no high school diploma, high school graduate, some college, or college graduate), and residence (one of the 50 U.S. states or the District of Columbia).

We fit two binary logistic regressions sequentially. The first model predicts whether a respondent intends to vote for one of the major-party candidates (Obama or Romney), and the second model predicts whether they support Obama or Romney, conditional upon intending to vote for one of these two. Specifically, the first model is given by

$$\begin{aligned} & \Pr(Y_i \in \{\text{Obama, Romney}\}) \\ &= \text{logit}^{-1} \left(\alpha_0 + a_{j[i]}^{\text{age}} + a_{j[i]}^{\text{sex}} + a_{j[i]}^{\text{race}} + a_{j[i]}^{\text{edu}} + a_{j[i]}^{\text{state}} \right) \end{aligned} \quad (1)$$

where Y_i is the i th response (Obama, Romney, or other) in R_d , α_0 is the overall intercept, and $a_{j[i]}^{\text{age}}$, $a_{j[i]}^{\text{sex}}$, $a_{j[i]}^{\text{race}}$, $a_{j[i]}^{\text{edu}}$, and $a_{j[i]}^{\text{state}}$ are random effects for the i -th respondent. Here we follow the notation of Gelman and Hill (2006) to indicate, for example, that $a_{j[i]}^{\text{age}} \in \{a_{18-29}^{\text{age}}, a_{30-44}^{\text{age}}, a_{45-64}^{\text{age}}, a_{65+}^{\text{age}}\}$ depending on the age of the i -th respondent, with $a_{j[i]}^{\text{age}} \sim N(0, \sigma_{\text{age}}^2)$, where σ_{age}^2 is a parameter to be estimated from the data. In this manner, the multilevel model partially pools data across the four age categories — as opposed to fitting each of the four coefficients separately — boosting statistical power. The benefit of this multilevel approach is most apparent for categories with large numbers of levels (for example, geographic location), but for consistency and simplicity we use a fully hierarchical model.

The second of the nested models predicts whether one supports Obama given one supports a major-party candidate, and is fit on the

subset $M_d \subseteq R_d$ for which respondents declared support for one of the major-party candidates. For this subset, we again predict the i -th response as a function of age, sex, race, education, and geographic location. Namely, we fit the model

$$\begin{aligned} \Pr(Y_i = \text{Obama} | Y_i \in \{\text{Obama}, \text{Romney}\}) \\ = \text{logit}^{-1} \left(\beta_0 + b_{j[i]}^{\text{age}} + b_{j[i]}^{\text{sex}} + b_{j[i]}^{\text{race}} + b_{j[i]}^{\text{edu}} + b_{j[i]}^{\text{state}} \right). \end{aligned} \quad (2)$$

Once these two models are fit, we can estimate the likelihood any respondent will report support for Obama, Romney, or “other” as a function of his or her demographic attributes. For example, to estimate a respondent’s likelihood of supporting Obama, we simply multiply the estimates obtained under each of the two models.

By the above, for each of the 6,528 combinations of age, sex, race, education, and geographic location, we can estimate the likelihood that a hypothetical individual with those demographic attributes will support each candidate. In the second step of MRP (post-stratification), we weight these 6,528 estimates by the assumed fraction of such individuals in the electorate. For simplicity, transparency, and repeatability in future elections, in our primary analysis we assume the 2012 electorate mirrors the 2008 electorate, as estimated by exit polls. In particular, we use the full, individual-level data from the exit polls (not the summary cross-tabulations) to estimate the proportion of the electorate in each demographic cell. Our decision to hold fixed the demographic composition of likely voters obviates the need for a likely voter screen, allows us to separate support from enthusiasm or probability of voting, and generates estimates that are largely in line with those produced by leading polling organizations.

The final step in computing the demographic post-stratification estimates is to account for the *house effect*: the disproportionate number of Obama supporters even after adjusting for demographics. For example, older voters who participate in the Xbox survey are more likely to support Obama than their demographic counterparts in the general electorate. To compute this overall bias of our sample, we first fit models (1) and (2) on the entire 45 days of Xbox polling data, and then post-stratify to the 2008 electorate as before. This yields (demographically-adjusted) estimates for the overall proportion of supporters for Obama, Romney, and “other”. We next compute the analogous estimates via models (3)

and (4) that additionally include respondents' partisanship, as measured by 2008 vote, ideology, and party identification. (These latter models are described in more detail in the partisan post-stratification section below.) As expected, the overall proportion of Obama supporters is smaller under the partisanship models than under the purely demographic models, and the difference of one percentage point between the two estimates is the house effect for Obama. Thus, our final, daily, demographically post-stratified estimates of Obama support are obtained by subtracting the Obama house effect from the MRP estimates. A similar house correction is used to estimate support for Romney and "other".

Partisan post-stratification.

To correct simultaneously for both demographic and partisan skew, we mimic the MRP procedure described above, but we now include partisanship attributes in the predictive models. Specifically, we include a panelist's 2008 vote (Obama, McCain, or "other"), party identification (Democrat, Republican, or "other"), and ideology (liberal, moderate, or conservative). As noted in the main text, all three of these covariates are collected first time when a panelist participates in a survey, which is necessarily before the first presidential debate. The multilevel logistic regression models we use are identical in structure to those in models (1) and (2) but now include the added predictors. Namely, we have

$$\begin{aligned} \Pr(Y_i \in \{\text{Obama, Romney}\}) & \\ &= \text{logit}^{-1} \left(\alpha_0 + a_{j[i]}^{\text{age}} + a_{j[i]}^{\text{sex}} + a_{j[i]}^{\text{race}} + a_{j[i]}^{\text{edu}} + a_{j[i]}^{\text{state}} \right. \\ &\quad \left. + a_{j[i]}^{2008 \text{ vote}} + a_{j[i]}^{\text{party ID}} + a_{j[i]}^{\text{ideology}} \right) \end{aligned} \quad (3)$$

and

$$\begin{aligned} \Pr(Y_i = \text{Obama} | Y_i \in \{\text{Obama, Romney}\}) & \\ &= \text{logit}^{-1} \left(\beta_0 + b_{j[i]}^{\text{age}} + b_{j[i]}^{\text{sex}} + b_{j[i]}^{\text{race}} + b_{j[i]}^{\text{edu}} + b_{j[i]}^{\text{state}} \right. \\ &\quad \left. + b_{j[i]}^{2008 \text{ vote}} + b_{j[i]}^{\text{party ID}} + b_{j[i]}^{\text{ideology}} \right). \end{aligned} \quad (4)$$

As before, we post-stratify to the 2008 electorate, where in this case there are a total of 176,256 cells, corresponding to all possible combinations of age, sex, race, education, geographic location, 2008 vote,

party identification, and ideology. Since here we explicitly incorporate partisanship, we do not adjust for the house effect as we did with the purely demographic adjustment.

Change in support by group.

Figure 5 shows swings in support around the first presidential debate broken down by various subgroups (for example, support among political moderates), under both partisan and demographic estimation models. To generate these estimates, we start with the same fitted multilevel models as above, but instead of post-stratifying to the entire 2008 electorate, we post-stratify to the 2008 electorate within the subgroup of interest. Thus, in the case of political moderates, younger voters have less weight than in the national estimates since they make up a relatively smaller fraction of the target subgroup of interest.

Partisan nonresponse.

To compute the demographically-adjusted daily partisan composition of the Xbox sample (shown in Figure 3), we mimic the demographic MRP approach described above. In this case, however, instead of vote intention, our models predict party identification. Specifically, we use nested models of the following form:

$$\begin{aligned} \Pr(Y_i \in \{\text{Democrat, Republican}\}) \\ = \text{logit}^{-1} \left(\alpha_0 + a_{j[i]}^{\text{age}} + a_{j[i]}^{\text{sex}} + a_{j[i]}^{\text{race}} + a_{j[i]}^{\text{edu}} + a_{j[i]}^{\text{state}} \right) \end{aligned} \quad (5)$$

and

$$\begin{aligned} \Pr(Y_i = \text{Democrat} | Y_i \in \{\text{Democrat, Republican}\}) \\ = \text{logit}^{-1} \left(\beta_0 + b_{j[i]}^{\text{age}} + b_{j[i]}^{\text{sex}} + b_{j[i]}^{\text{race}} + b_{j[i]}^{\text{edu}} + b_{j[i]}^{\text{state}} \right). \end{aligned} \quad (6)$$

As before, smoothed, daily estimates are computed by separately fitting Equations (5) and (6) on the set of responses R_d collected in a moving four-day window. Final partisan composition is based on post-stratifying to the 2008 exit polls.

Individual-level opinion change.

To estimate rates of opinion change (shown in Figure 6), we take advantage of the ad hoc panel design of our survey, where 12,425

individuals responded both during the seven days before and during the seven days after the first debate. Specifically, for each of these panelists, we denote their last pre-debate response by y_i^{pre} and their first post-debate response by y_i^{post} . As before, we need to account for the demographic and partisan skew of our panel to make accurate estimates, for which we again use MRP. In this case we use four nested models. Mimicking Equations (3) and (4), the first two models, given by Equations (7) and (8), estimate panelists' pre-debate vote intention by decomposing their opinions into support for a major-party candidate, and then support for Obama conditional on supporting a major-party candidate. The third model, in Equation (9), estimates the probability that an individual switches his or her support (that $y_i^{\text{pre}} \neq y_i^{\text{post}}$). It has the same demographic and partisanship predictors as both (3) and (7), but additionally includes a coefficient for the panelist's pre-debate response (shown in bold). The fourth and final of the nested models, in Equation (10), estimates the likelihood that, conditional on switching, a panelist switches to the more Republican of the alternatives (an Obama supporter switching to Romney, or a Romney supporter switching to "other"). This model is likewise based on demographics, partisanship, and pre-debate response.

$$\begin{aligned} & \Pr(y_i^{\text{pre}} \in \{\text{Obama, Romney}\}) \\ &= \text{logit}^{-1}\left(\alpha_0 + a_{j[i]}^{\text{age}} + a_{j[i]}^{\text{sex}} + a_{j[i]}^{\text{race}} + a_{j[i]}^{\text{edu}} + a_{j[i]}^{\text{state}} \right. \\ & \quad \left. + a_{j[i]}^{2008 \text{ vote}} + a_{j[i]}^{\text{party ID}} + a_{j[i]}^{\text{ideology}}\right), \end{aligned} \tag{7}$$

$$\begin{aligned} & \Pr(y_i^{\text{pre}} = \text{Obama} | y_i \in \{\text{Obama, Romney}\}) \\ &= \text{logit}^{-1}\left(\beta_0 + b_{j[i]}^{\text{age}} + b_{j[i]}^{\text{sex}} + b_{j[i]}^{\text{race}} + b_{j[i]}^{\text{edu}} + b_{j[i]}^{\text{state}} \right. \\ & \quad \left. + b_{j[i]}^{2008 \text{ vote}} + b_{j[i]}^{\text{party ID}} + b_{j[i]}^{\text{ideology}}\right), \end{aligned} \tag{8}$$

$$\begin{aligned} & \Pr(y_i^{\text{pre}} \neq y_i^{\text{post}}) \\ &= \text{logit}^{-1}\left(\beta_0 + b_{j[i]}^{\text{age}} + b_{j[i]}^{\text{sex}} + b_{j[i]}^{\text{race}} + b_{j[i]}^{\text{edu}} + b_{j[i]}^{\text{state}} \right. \\ & \quad \left. + b_{j[i]}^{2008 \text{ vote}} + b_{j[i]}^{\text{party ID}} + b_{j[i]}^{\text{ideology}} + \mathbf{b}_{j[i]}^{\text{pre}}\right), \end{aligned} \tag{9}$$

and

$$\begin{aligned} \Pr \left(\mathbf{y}_i^{\text{post}} = \text{more Republican alternative} \mid \mathbf{y}_i^{\text{pre}} \neq \mathbf{y}_i^{\text{post}} \right) \\ = \text{logit}^{-1} \left(\beta_0 + b_{j[i]}^{\text{age}} + b_{j[i]}^{\text{sex}} + b_{j[i]}^{\text{race}} + b_{j[i]}^{\text{edu}} + b_{j[i]}^{\text{state}} \right. \\ \left. + b_{j[i]}^{2008 \text{ vote}} + b_{j[i]}^{\text{party ID}} + b_{j[i]}^{\text{ideology}} + \mathbf{b}_{j[i]}^{\text{pre}} \right). \end{aligned} \quad (10)$$

After fitting these four nested models, we post-stratify to the 2008 electorate as before.

Model calibration.

Our analysis is premised on the idea that despite the non-representative nature of the Xbox sample, our modeling approach is still able to generate accurate estimates of population-level vote intention. In part, this assumption is validated by the close agreement between actual and model-predicted two-party vote share (52% for Obama). We further evaluate model performance by examining actual and predicted election outcomes for various demographic subgroups, where actual outcomes are based on exit polling data and predicted outcomes are generated based on Xbox data available the day before the election. Figure A.2 shows that the model estimates are indeed in line with outcomes across all major demographic categories.

Cross-sectional analysis.

In our primary analysis, we relied on a panel of respondents who reported their ideology, party ID, and 2008 vote prior to the first presidential debate. By post-stratifying on these partisanship measures (along with other fixed, demographic characteristics), we are able to estimate vote intention for a static group of individuals that mirrors the 2008 electorate. In particular, this panel design avoids problems of endogeneity associated with weighting by partisanship in cross-sectional surveys, where stated party affiliations may change over time in concert with vote intention. For example, if at any given time, one's reported party ID perfectly reflects one's vote intention, post-stratifying by current party ID would misleadingly result in flat estimates of candidate support.

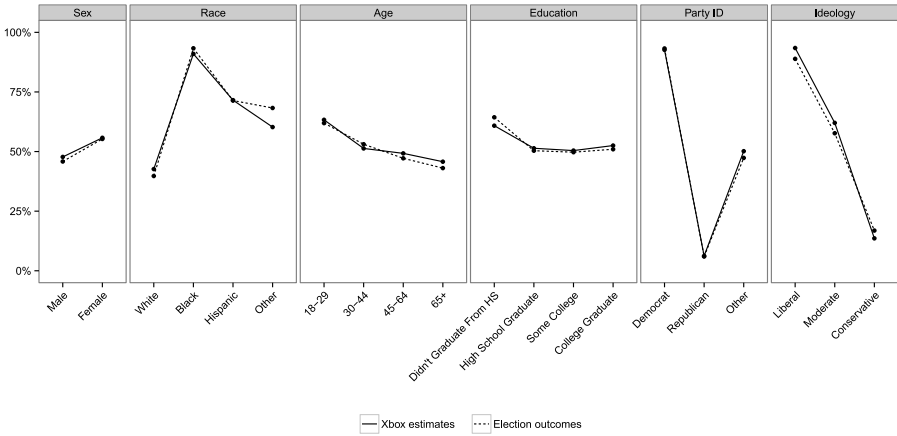


Figure A.2: Comparison of election outcomes (estimated from exit poll data) to Xbox predictions computed the day prior to the election. Despite being based on a highly non-representative sample, the Xbox predictions are largely in line with the election outcomes.

Nevertheless, despite the theoretical advantages of a panel analysis, cross-sectional data are often easier to collect. To check whether our approach can be applied to cross-sectional surveys, we limit our Xbox sample to the 327,432 first-time interviews — in which respondents simultaneously provide both partisanship and vote intention information — and then correct for partisan nonresponse via MRP as before. That is, we discard all follow-up interviews, where only vote intention was collected. Figure A.3 shows that post-stratifying on the cross-sectionally reported partisanship measures yields similar results to those obtained via the panel analysis. Thus, to a large extent, partisan nonresponse can be detected and adjusted for even via a naive statistical approach that does not account for possible movements in reported partisanship.

Raking.

MRP is a robust approach for identifying and correcting for partisan nonresponse. Our qualitative findings, however, can also be seen with

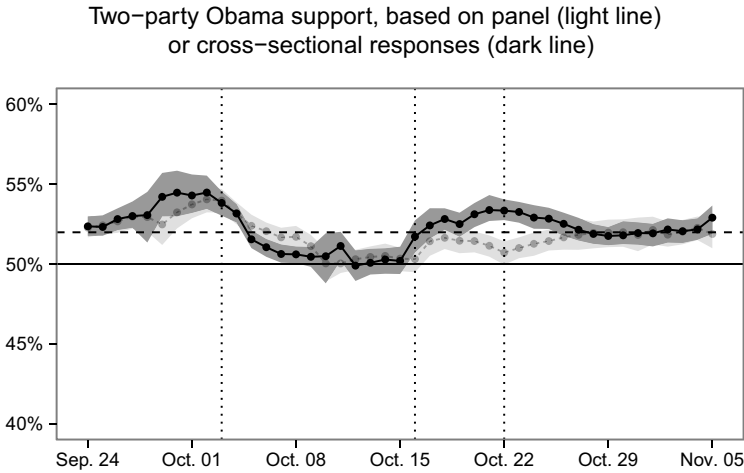


Figure A.3: Obama share of the two-party vote preference (with 95% confidence bands) estimated from the Xbox via a panel (light line) and a cross-section (dark line) of users. Estimates from the panel are post-stratified on partisanship measures (ideology, party ID, and 2008 vote) collected prior to the first debate, whereas the cross-sectional estimates are based on partisanship measures collected at the same time as voter intention. The similarity between the two curves suggests that in practice, cross-sectional data can be used to adjust for partisan nonresponse.

conventional survey adjustments, such as raking (DeBell and Krosnick, 2009). Specifically, the dashed line in Figure A.4 shows the results of raking the Xbox data with demographic variables (age, sex, race, and education); and the solid line shows the results of raking with both demographic and partisanship (party ID and 2008 vote) variables. As with our MRP estimates, these raking estimates are computed separately for each day based on data collected during the previous four days. Both raking models appear to consistently overestimate support for Obama by several percentage points throughout the campaign. However, the qualitative trends are largely consistent with the MRP estimates. In particular, whereas the demographic-only model shows a precipitous fall in support for Obama following the first debate, the demographic-plus-partisanship model shows a much more modest decline.

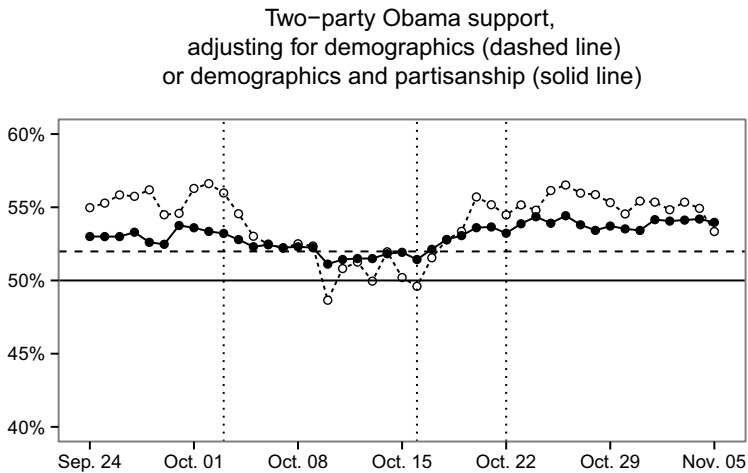


Figure A.4: Obama share of the two-party vote preference, as estimated from the Xbox data via two different raking models: one based on demographics (dashed line) and one based on demographics and partisanship (solid line). Though both models consistently overestimate support for Obama, they mimic the overall trends of the MRP estimates.

References

- Allsop, D. and H. F. Weisberg (1988), “Measuring Change in Party Identification in an Election Campaign”, *American Journal of Political Science*, 996–1017.
- Baldassarri, D. and A. Gelman (2008), “Partisans Without Constraint: Political Polarization and Trends in American Public Opinion”, *American Journal of Sociology*, 114(2), 408–46.
- DeBell, M. and J. A. Krosnick (2009), “Computing Weights for American National Election Study Survey Data”, ANES Technical Report series, no. nes012427. Ann Arbor, MI, and Palo Alto, CA: American National Election Studies. Available at: <http://www.electionstudies.org>.
- Fiorina, M. P. and S. J. Abrams (2008), “Political Polarization in the American Public”, *Annual Review of Political Science*, 11, 563–88.
- Gelman, A. and J. Hill (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

- Gelman, A. and G. King (1993), “Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?”, *British Journal of Political Science*, 23(04), 409–51.
- Gelman, A. and T. C. Little (1997), “Poststratification into Many Categories Using Hierarchical Logistic Regression”, *Survey Methodology*.
- Ghitza, Y. and A. Gelman (2013), “Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups”, *American Journal of Political Science*, 57(3), 762–76.
- Gutsche, T., A. Kapteyn, E. Meijer, and B. Weerman (2014), “The RAND Continuous 2012 Presidential Election Poll”, *Public Opinion Quarterly*.
- Hillygus, D. S. and S. Jackman (2003), “Voter Decision Making in Election 2000: Campaign Effects, Partisan Activation, and the Clinton Legacy”, *American Journal of Political Science*, 47(4), 583–96.
- Himmelweit, H. T., M. J. Biberian, and J. Stockdale (1978), “Memory for Past Vote: Implications of A Study of Bias in Recall”, *British Journal of Political Science*, 8(03), 365–75.
- Kaminska, O. and C. Barnes (2008), “Party Identification Weighting: Experiments to Improve Survey Quality”, in *Elections and Exit Polling*, NJ, Wiley Hoboken, 51–61.
- Kaplan, N., D. K. Park, and A. Gelman (2012), “Polls and Elections Understanding Persuasion and Activation in Presidential Campaigns: The Random Walk and Mean Reversion Models”, *Presidential Studies Quarterly*, 42(4), 843–66.
- Lax, J. R. and J. H. Phillips (2009), “How Should We Estimate Public Opinion in the States?”, *American Journal of Political Science*, 53(1), 107–21.
- Levendusky, M. (2009), *The Partisan Sort: How Liberals became Democrats and Conservatives became Republicans*, University of Chicago Press.
- Little, R. J. A. (1993), “Post-stratification: A Modeler’s Perspective”, *Journal of the American Statistical Association*, 88(423), 1001–12.
- Reilly, C., A. Gelman, and J. Katz (2001), “Poststratification Without Population Level Information on the Poststratifying Variable With Application to Political Polling”, *Journal of the American Statistical Association*, 96(453).
- Sides, J. and L. Vavreck (2014), *The gamble: Choice and Chance in the 2012 Presidential Election*, Princeton University Press.