High-Frequency Polling with Non-Representative Data

Andrew Gelman
Columbia University
gelman@stat.columbia.edu

Sharad Goel
Stanford University
scgoel@stanford.edu

David Rothschild
Microsoft Research
davidmr@microsoft.com

Wei Wang
Columbia University
ww2243@columbia.edu

Probability-based sampling methods, such as random-digit dialing, are a staple of modern polling and have been successfully used to gauge public opinion for nearly 80 years. Though historically effective, such traditional methods are often slow and expensive, and with declining response rates, even their accuracy has come into question. At the same time, non-representative polls, such as opt-in online surveys, have become increasingly fast and cheap. We show that with proper statistical adjustment, non-representative polling can be used to accurately and continuously track public sentiment, offering a new approach to public opinion research.

Introduction

Modern polling is based on the simple and theoretically appealing idea of probability sampling: if each member of the target population has a known, non-zero chance of being surveyed, then a small random sample of the population can be used to accurately estimate the distribution of attitudes in the entire population. This methodological approach has guided polling from the early days of in-home interviewing, through random-digit dialing (RDD) of landlines, to more recent mixed-mode polling of landlines and cellphones. Of course, it has never been possible to reach everyone in the population (e.g., those without phones or permanent addresses), or to guarantee 100% response rates. In practice, it is thus common to use probability-*based* sampling, in which a variety of post-sampling corrections, such as raking (Battaglia, Hoaglin, & Frankel, 2013), are applied to correct for coverage and non-response errors. Nevertheless, the idea that one should start with approximately representative samples has permeated the practice of polling for nearly a century.

In recent years, however, it has become increasingly difficult to construct representative samples, and traditional polling has veered further from its theoretical underpinnings. In part, the difficulty of collecting representative samples stems from the steep rise in mobile phone usage and the accompanying decline in landline penetration. Between 2004 and 2014, the percentage of Americans with a mobile phone but no landline went from less than 5% to 44% (Pew Research Center, 2014b). Whereas landlines are systematically catalogued and organized geographically, mobile phones are often unlisted and not used in their nominal geographic designation. Further, it is hard to tell who has only a mobile phone and who has both a mobile phone and a landline, which in turn leads to members of certain subgroups to be double counted. Compounding these issues of coverage, response rates have plummeted to as low as 5%, since people increasingly screen calls and even those who do answer are reluctant to participate in surveys (Pew Research Center, 2014a). These low response rates raise concerns that survey participants are not representative of the population at large. Moreover, given the difficulty in finding willing participants, low response rates also increase the time and cost to conduct surveys.

While the cost—in both time and money—of constructing representative samples has increased, it has simultaneously become substantially easier to quickly collect large, non-representative samples of public opinion thorough online, opt-in surveys. Such nonprobability-based sampling has generally been dismissed by the polling community, and not without reason. It is not obvious how to extract meaningful signal from a collection of participants whose opinions are often far from representative of the population-at-large. Indeed, the birth of modern polling can be traced to an infamous polling mishap during the 1936 U.S. presidential election campaign, in which the popular magazine Literary Digest predicted, based on a non-representative survey of its readers, a landslide victory for Republican candidate Alf Landon over the incumbent Franklin Roosevelt. Roosevelt, of course, won the election decisively, carrying every state except for Maine and Vermont.

Here we show that with proper statistical adjustment, even highly non-representative polls can yield accurate estimates of population-level attitudes. Our approach is to use multilevel-regression and poststratification (MRP). With MRP, we partition the data into thousands of demographic cells, estimate attitudes at the cell level with a multilevel regression model, and then aggregate cell-level estimates in accordance with the target population's demographic composition. We note that while MRP is well-known in the statistics community, it is primarily viewed as a means for reducing variance, not for correcting bias as we do here. We demonstrate this approach by estimating support for presidential candidates from an opt-in sample gathered on the Xbox gaming system. The survey was available on Xbox for the last 45 days of the election, and respondents were allowed to register their opinions up to once per day. This sample clearly has significant coverage error since respondents needed to have an Xbox to participate. Further, relative to the voting population, respondents were highly skewed in both gender and age. Thus, unsurprisingly, the raw responses do not accurately reflect public opinion. However, after applying MRP, we show this non-representative sample yields estimates inline with those from the best available traditional polls as well as the final election outcome.

Non-representative polling offers the promise of fast, cheap, and accurate assessments of public opinion. Traditional polls often require several days to reach statistically meaningful sample sizes, and because of the logistics and expense, major polling organizations are limited in the number of polls they can conduct. In contrast, high-frequency, non-representative polling can be used to continually track changes in sentiment, giving researchers, marketers, and campaign advisors a powerful new tool for studying scientific questions and making informed decisions. In this chapter, we summarize and extend results originally described in Wang, Rothschild, Goel, & Gelman (2015) and Gelman, Goel, Rivers, & Rothschild (2016).

The non-representative Xbox data

For the last 45 days before the 2012 election, we operated an opt-in polling application on the Xbox gaming platform; questions were changed daily and users could respond to questions up to once per day. There was limited coverage for this survey in that the only way to answer the polling questions was via the Xbox Live gaming platform. And the survey was truly opt-in: there was no invitation or permanent link to the poll, so respondents had to locate it daily on the Xbox Live home page. Each daily poll had three to five questions, but we always included one question on voter intention: "If the election were held today, who would you vote for?" Before taking their first poll, and only before their first poll, respondents were asked to provide a variety of demographic information about themselves, including their gender, race, age, education, state, party ID, political ideology, and 2008 U.S. presidential vote. Images and full details of the demographic questions are included in the Appendix. Nearly 350,000 people completed one poll, and over 30,000 completed five or more polls.

Participants in the Xbox survey are, unsurprisingly, not representative of the voting population. Figure 5.1 compares the voting population (estimated by the 2012 national exit poll) with the Xbox respondents. The key differences are in gender, where men compose 93% of the Xbox sample compared to 47% of the voting population, and age, where 18–29 year-olds constitute 65% of the Xbox sample compared to just 19% of the voting population. Political scientists have long observed that both gender and age are strongly correlated with voting preferences (Kaufmann & Petrocik, 1999), and so we would not expect the raw, unadjusted results of the voter intention question to accurately reflect sentiment in the general population. Figure 5.2 shows unadjusted Xbox estimates (solid line with solid circles) of Obama's two-party support (i.e., support for Obama divided by total support for Obama and Romney, excluding third-party candidates); the dashed line at 52% indicates the final two-party outcome of the election. The average support among Xbox respondents swings wildly and implausibly from day to day. As a point of comparison, the dotted line with white-filled circles shows the Pollster.com rolling average, the industry-standard estimate of support based on hundreds of traditional polls.

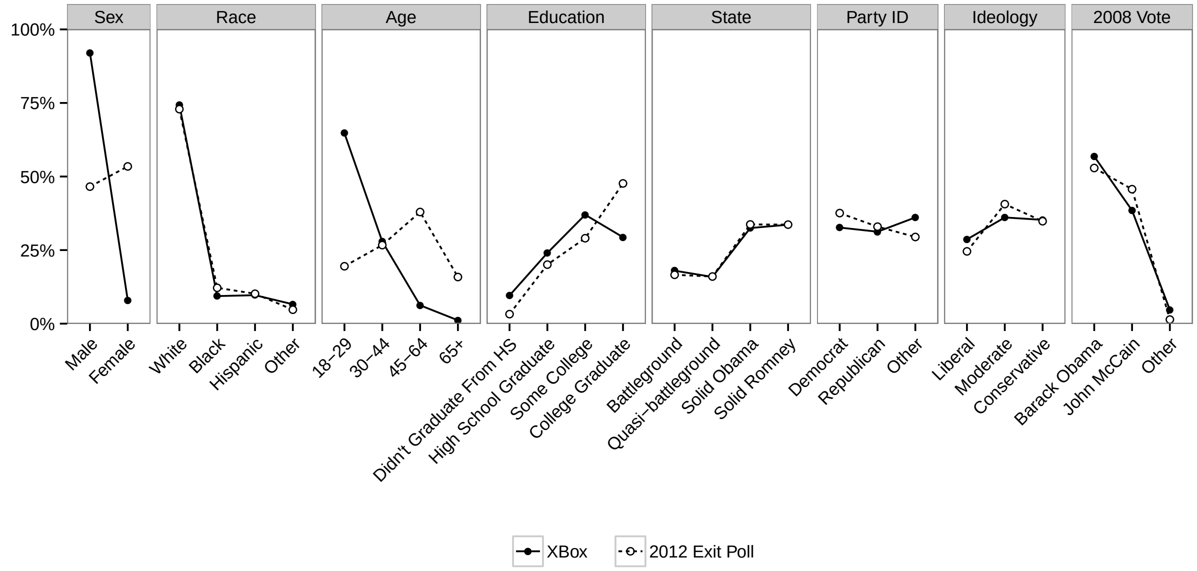


Figure 5.1: *A comparison of the demographic composition of participants in the Xbox survey and the 2012 electorate.*

Note: The 2012 electorate is measured by adjusted exit polls. The most prominent differences are in gender and age. For more detail on how we broke down geographical groups, please see the Appendix.

Statistically adjusted estimates

As Figure 5.2 shows, unadjusted estimates from non-representative samples do not reflect sentiment in the general population. The most common statistical approach to correcting such sample bias is raking (Battaglia et al., 2013), where weights are assigned to each respondent so that the marginal weighted distribution of respondent demographics (e.g., age, sex, and race) match those in the target population. The dashed line with “x” in Figure 5.3 shows the results of raking the Xbox data to match demographics (age, gender, race, and education) and partisanship (party ID and 2008 vote). Marginal distributions are matched to those from the 2008 electorate, as estimated from national exit polls. Raked estimates of support are computed separately for each day based on data collected during the previous four days. The adjusted estimates are certainly not perfect, and relative to the Pollster.com average, appear to consistently overestimate support for Obama. Nevertheless, it is perhaps surprising that even the simplest statistical correction can be used to extract meaningful signal from such a non-representative sample.

Multilevel regression and poststratification

Raking is a popular method for survey adjustment, but can suffer from high variance due to large respondent weights on under-sampled segments of the population. Here we consider

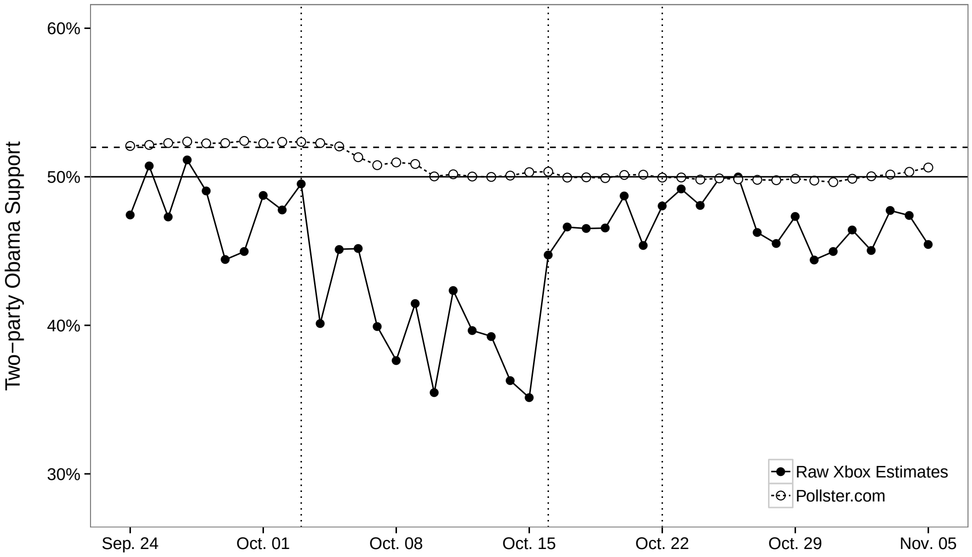


Figure 5.2: *Daily unadjusted Xbox and estimates of two-party Obama support during the 45 days leading up to the 2012 presidential election.*

Note: the unadjusted suggests a landslide victory for Mitt Romney. The dotted line with white-filled circles indicates a consensus average of traditional polls (the daily aggregated polling results from Pollster.com), the horizontal dashed line at 52% indicates the actual two-party vote share obtained by Barack Obama, and the vertical dotted lines give the dates of the three presidential debates.

an alternative method of statistical correction: multilevel regression and poststratification. Poststratification, in general, is a popular method for correcting for known differences between sample and target populations (Little, 1993). The idea is to first partition the population into cells based on combinations of various demographic and political attributes, then to use the sample to estimate the response variable within each cell, and finally to aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population. Using y to indicate the outcome of interest, the poststratification estimate is defined by,

$$\hat{y}^{\text{PS}} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the j-th cell in the population. We can analogously derive an estimate of y for any subpopulation s that is the union of cells (e.g., voter intent in a particular state) by

$$\hat{y}_s^{\text{PS}} = \frac{\sum_{j \in J_s} N_j \hat{y}_j}{\sum_{j \in J_s} N_j}$$

where J_s is the set of all cells that comprise s. As is readily apparent from the form of the poststratification estimator, the key is to obtain accurate cell-level estimates as well as

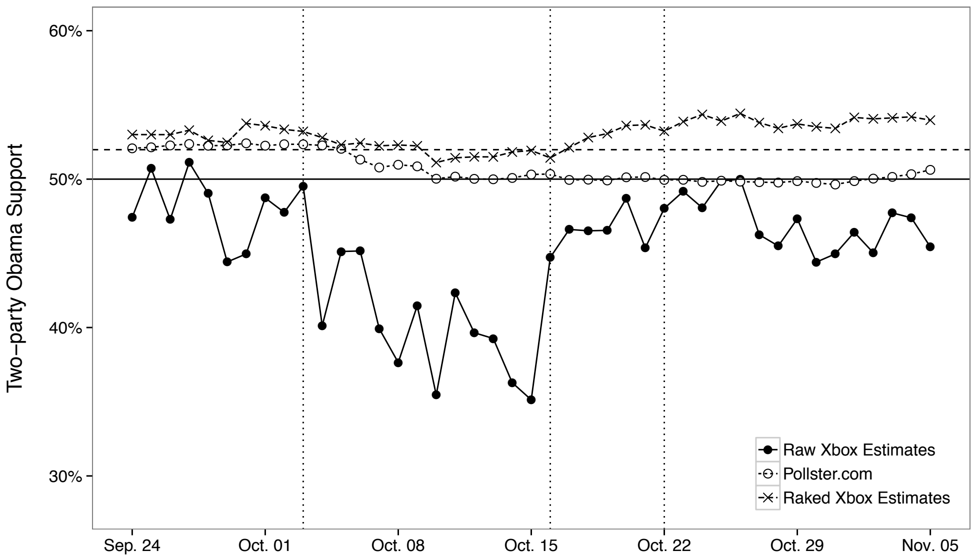


Figure 5.3: *A comparison of daily unadjusted and raked Xbox estimates of two-party Obama support during the 45 days leading up to the 2012 presidential election.*

Note: The dotted line with white-filled circles indicates a consensus average of traditional polls (the daily aggregated polling results from Pollster.com), the horizontal dashed line at 52% indicates the actual two-party vote share obtained by Barack Obama, and the vertical dotted lines give the dates of the three presidential debates. The unadjusted estimates suggest an implausible landslide victory for Mitt Romney. The raked estimates (dashed line with "x"), seem more reasonable, but appear to overestimate support for Obama.

estimates for the cell sizes. One of the most common ways to generate cell-level estimates is to simply average the sample responses within each cell. If we assume that within a cell the sample is drawn at random from the larger population, this yields an unbiased estimate. However, this assumption of cell-level simple random sampling is only reasonable when the partition is sufficiently fine; on the other hand, as the partition becomes finer, the cells become sparse and the empirical sample averages become unstable. We address the sparsity issues by instead generating cell-level estimates via a regularized regression model, namely multilevel regression. This combined model-based poststratification strategy, known as multilevel regression and poststratification (MRP), has been used to obtain accurate small-area subgroup estimates, such as for public opinion and voter turnout in individual states and demographic subgroups (Park, Gelman, & Bafumi, 2004; Lax & Phillips, 2009; Ghitza & Gelman, 2013).

Applying MRP in our setting entails two steps. First, a Bayesian hierarchical model is fit to obtain estimates for sparse poststratification cells; second, one averages over the cells, weighting by a measure of forecasted voter turnout, to get state and national-level estimates. Specifically, we generate the cells by considering all possible combinations of gender (2 categories), race (4 categories), age (4 categories), education (4 categories), state (51 categories), party ID (3 categories), ideology (3 categories) and 2008 vote (3 categories), which partition the population into 176,256 cells. The partisanship variables are strong predictors of vote intention, and so their inclusion is important to generate accurate estimates. However, such poststratification on party identification can be controversial (Pollster.com, 2004). In this case we believe it is appropriate, first because partisanship tends to vary more slowly than vote intentions and political attitudes (Gelman & King, 1993; Cavan Reilly & Katz, 2001), and second because these variables in the Xbox survey were measured only once during the campaign, at the time of the respondents' entry into the panel. Respondent IDs are anonymized, so while we are able to track a respondent across polls we cannot link responses to individual Xbox accounts.

We fit two nested multilevel logistic regressions to estimate candidate support in each cell. The first of the two models predicts whether a respondent supports a major-party candidate (i.e., Obama or Romney), and the second predicts support for Obama given that the respondent supports a major-party candidate. Following the notation of Gelman & Hill (2007), the first model is given by

$$\begin{aligned}\Pr(Y_i \in \{\text{Obama, Romney}\}) = \\ & \text{logit}^{-1}\big(\alpha_0 + \alpha_1(\text{past Democrat vote share in respondent's state}) \qquad (1)\\ & + a_{j[i]}^{\text{state}} + a_{j[i]}^{\text{edu}} + a_{j[i]}^{\text{gender}} + a_{j[i]}^{\text{age}} + a_{j[i]}^{\text{race}} + a_{j[i]}^{\text{party ID}} + b_{j[i]}^{\text{ideology}} + b_{j[i]}^{\text{last vote}}\big)\end{aligned}$$

where α_0 is the fixed baseline intercept, and α_1 is the fixed slope for Obama's fraction of two-party vote share in the respondent's state in the last presidential election. The terms $a_{j[i]}^{\text{state}}$, $a_{j[i]}^{\text{edu}}$, $a_{j[i]}^{\text{sex}}$ and so on—which in general we denote by $a_{j[i]}^{\text{var}}$—correspond to varying coefficients associated with each categorical variable. Here, with a slight abuse of notation, the subscript $j[i]$ indicates the demographic group to which the i-th respondent belongs. For example, $a_{j[i]}^{\text{age}}$ takes values from $\{a_{18-29}^{\text{age}}, a_{30-44}^{\text{age}}, a_{45-64}^{\text{age}}, a_{65+}^{\text{age}}\}$ depending on the cell membership of the i-th respondent. The varying coefficients $a_{j[i]}^{\text{var}}$ are given independent prior distributions

$$a_{j[i]}^{\text{var}} \sim N(0, \sigma_{\text{var}}^2).$$

To complete the full Bayesian specification, the variance parameters are assigned a hyperprior distribution

$$\sigma_{\text{var}}^2 \sim \text{inv-}\chi^2(\nu, \sigma_0^2),$$

with a weak prior specification for the remaining parameters, ν and σ_0. The benefit of using such a multilevel model is that estimates for relatively sparse cells can be improved through borrowing strength from demographically similar cells that have richer data. Similarly, the second model is defined by

$$\begin{aligned}\Pr(Y_i = \text{Obama} \,|\, Y_i \in \{\text{Obama, Romney}\}) = \\ & \text{logit}^{-1}\big(\beta_0 + \beta_1(\text{past Democrat vote share in respondent's state}) \qquad (2)\\ & + b_{j[i]}^{\text{state}} + b_{j[i]}^{\text{edu}} + b_{j[i]}^{\text{sex}} + b_{j[i]}^{\text{age}} + b_{j[i]}^{\text{race}} + b_{j[i]}^{\text{party ID}} + b_{j[i]}^{\text{ideology}} + b_{j[i]}^{\text{last vote}}\big)\end{aligned}$$

and

$$\begin{aligned} b^{\text{var}}_{j[i]} &\sim N(0, \eta^2_{\text{var}}), \\ \eta^2_{\text{var}} &\sim \text{inv-}\chi^2(\mu, \eta^2_0). \end{aligned}$$

Jointly, Equations (1) and (2) (together with the priors and hyperpriors) define a Bayesian model that describes the data. Ideally, we would perform a fully Bayesian analysis to obtain the posterior distribution of the parameters. However, for computational convenience, we use the approximate marginal maximum likelihood estimates obtained from the `glmer()` function in the `R` package `lme4` (Bates, Maechler, & Bolker, 2013). We fit separate models for each day using a four-day moving window, aggregating the data collected on that day and the previous three days, to make cell-level estimates for each of the 45 days leading up to the election.

Having detailed the multilevel regression step, we now turn to poststratification, where cell-level estimates are weighted by the proportion of the electorate in each cell and aggregated to the appropriate level (e.g., state or national). To compute cell weights, we require cross-tabulated population data. One commonly used source for such data is the Current Population Survey (CPS); however, the CPS does not include some key poststratification variables, such as party identification. We thus instead use exit poll data from the 2008 presidential election. Exit polls are conducted on election day outside voting stations to record the choices of exiting voters, and they are generally used by researchers and news media to analyze the demographic breakdown of the vote (after a post-election adjustment that aligns the weighted responses to the reported state-by-state election results). In total, 101,638 respondents were surveyed in the state and national exit polls. We use the exit poll from 2008, not 2012, because this means that in theory our method as described here could have been used to generate real-time predictions during the 2012 election campaign. Admittedly, this approach puts our predictions at a disadvantage since we cannot capture the demographic shifts of the intervening four years. While combining exit poll and CPS data would arguably yield improved results, for simplicity and transparency we exclusively use the 2008 exit poll summaries for our poststratification.

National and state voter support

Figure 5.4 shows the adjusted two-party Obama support for the last 45 days of the election. Compared with the uncorrected estimates in Figure 5.2, the MRP-adjusted estimates yield a much more reasonable timeline of Obama's standing over the course of the final weeks of the campaign. With a clear advantage at the beginning, Obama's support slipped rapidly after the first presidential debate—though never falling below 50%—and gradually recovered, building up a decisive lead in the final days.

On the day before the election, our estimate of voter intent is off by a mere 0.6 percentage points from the actual outcome (indicated by the dashed horizontal line). Voter intent in the weeks prior to the election does not directly equate to an estimate of vote share on election day. As such, it is difficult to evaluate the accuracy of our full time-series of estimates.

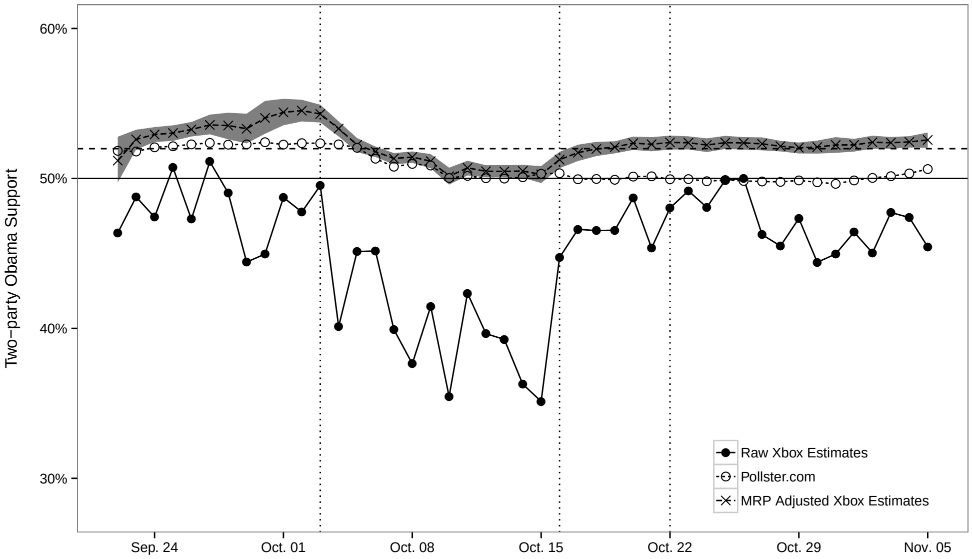


Figure 5.4: *National MRP-adjusted voter intent of two-party Obama support over the 45-day period and the associated 95% confidence bands.*

Note: The horizontal dashed line indicates the actual two-party Obama vote share. The three vertical dotted lines indicate the presidential debates. Compared with the raw responses, the MRP-adjusted voter intent is much more reasonable, and voter intent in the last few days is close to the actual outcome. Notably, during the final days of the campaign, results from Pollster.com (dotted line with white-filled circles) are further off from the actual vote share than are estimates generated from the Xbox data.

However, our estimates are not only intuitively reasonable but are also in line with prevailing estimates based on traditional, representative polls. In particular, our estimates roughly track those from Pollster.com, one of the leading poll aggregators during the 2012 campaign, illustrating the potential of non-representative polling.

National vote share receives considerable media attention, but state-level estimates are at least as important given the role of the Electoral College in selecting the winner (Rothschild, 2015). Forecasting state-by-state races is a challenging problem due to the interdependencies in state outcomes, the logistical difficulties of measuring state-level vote preference, and the effort required to combine information from various sources (Lock & Gelman, 2010). The MRP framework, however, provides a straightforward methodology for generating state-level results. Namely, we use the same cell-level estimates employed in the national estimate, as generated via the multilevel model in Equations (1) and (2), and we then poststratify to each state's demographic composition. Like any other subgroup, we do not need a large number of respondents in any given cell in order to estimate the vote share. In particular, we can generate state-level estimates even for states in which we have few respondents. In this manner, the Xbox responses can be used to construct estimates of voter intent over the last 45 days of the campaign for all 51 Electoral College races.

Figure 5.5 shows two-party Obama support for the 12 states with the most electoral votes.

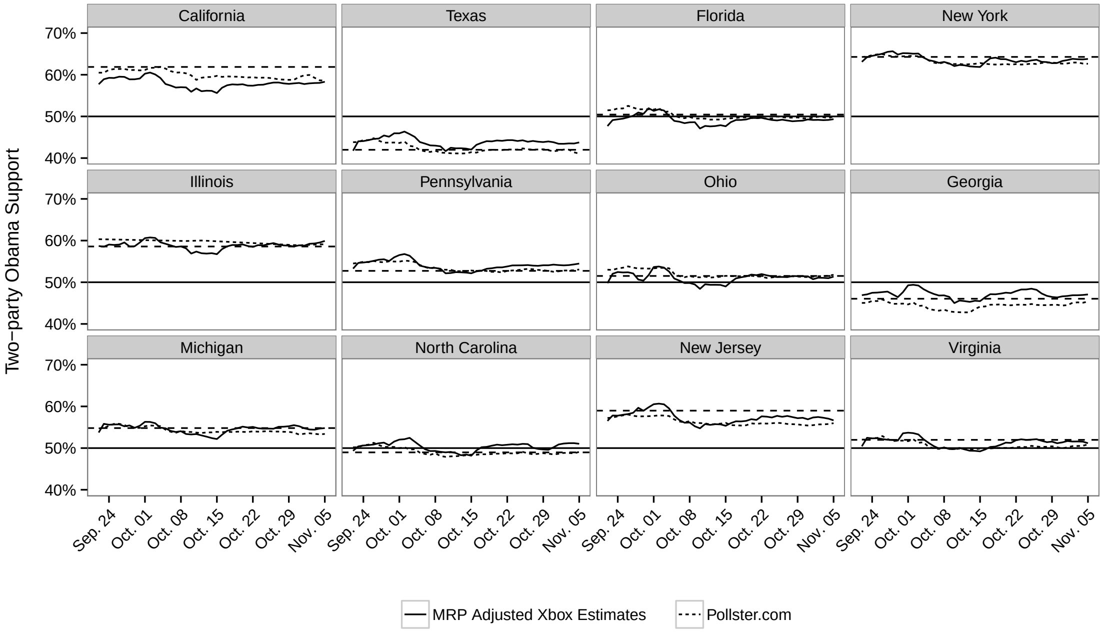


Figure 5.5: *MRP-adjusted daily voter intent for the 12 states with the most electoral votes, and the associated 95% confidence bands.*

Note: The horizontal dashed lines in each panel give the actual two-party Obama vote share in that state. The mean and median absolute errors of the last day voter intent across the 51 Electoral College races are 2.5 and 1.8 percentage points, respectively. The state-by-state daily aggregated polling results from Pollster.com, given by the dotted lines, are broadly consistent with the estimates from the Xbox data.

The state timelines share similar trends (e.g., support for Obama dropping after the first debate), but also have their own idiosyncratic movements, an indication of a reasonable blend of national and state-level signals. To evaluate the accuracy of the MRP-adjusted estimates, we also plot, in dotted lines with white-filled circles, estimates generated by Pollster.com, which are broadly consistent with our state-level MRP estimates. Moreover, across the 51 Electoral College races, the mean and median absolute errors of our estimates on the day before the election are just 2.5 and 1.8 percentage points, respectively.

Voter support for demographic subgroups

Apart from Electoral College races, election forecasting often focuses on candidate preference among demographic subpopulations. Such forecasts are of significant importance in modern political campaigns, which often employ targeted campaign strategies (Hillygus & Shields, 2009). In the highly non-representative Xbox survey, certain subpopulations are heavily underrepresented and plausibly suffer from strong self-selection problems. This begs the question, can we reasonably expect to estimate the views of older women on a platform that largely caters to young men?

It is straightforward in MRP to estimate voter intent among any collection of demographic cells: we again use the same cell-level estimates as in the national and state settings, but poststratify to the desired target population. For example, to estimate voter intent among women, the poststratification weights are based on the relative number of women in each demographic cell. To illustrate this approach, we compute Xbox estimates of Obama support for each level of our categorical variables (e.g., males, females, whites, and blacks) on the day before the election, and compare those with the actual voting behavior of these groups as estimated by the 2012 national exit poll. The Xbox estimates are remarkably accurate, with a median absolute difference of 1.5 percentage points between the Xbox and the exit poll numbers.

Not only do the Xbox data facilitate accurate estimation of voter intent across these single-dimensional demographic categories, but they also do surprisingly well at estimating two-way interactions (e.g., candidate support among 18–29 year-old Hispanics, and liberal college graduates). Figure 5.6 shows this result, plotting the Xbox estimates against those derived from the exit polling data for each of the 149 two-dimensional demographic subgroups Most points lie near the diagonal, indicating that the Xbox and exit poll estimates are in close agreement. Specifically, for women who are 65 and older—a group whose preferences one might a priori believe are hard to estimate from the Xbox data—the difference between the Xbox and the exit poll numbers is a mere one percentage point (49.5% and 48.5%, respectively). Across all the two-way interaction groups, the median absolute difference is just 2.4 percentage points. As indicated by the size of the points in Figure 5.6, the largest differences occur for relatively small demographic subgroups (e.g., liberal Republicans), for which both the Xbox and exit poll estimates are less reliable. For the 30 largest demographic subgroups, Figure 5.6 lists the differences between Xbox and exit poll estimates. Among these largest subgroups, the median absolute difference drops to just 1.9 percentage points.

Insights from high-frequency panels

Arguably the most critical stretch of the final 45 days of the 2012 election was the period immediately following the first presidential debate on October 3. No other time saw such large swings in public opinion for the two major candidates, with Obama losing several points in the polls after what was widely perceived as a poor debate performance. However, such rapid and unexpected movements are hard to capture by conventional polls given the difficulty of quickly collecting a representative sample of respondents. In particular, the Pollster.com average of Obama support remains flat for several days after the debate before slowly declining. In contrast, the continuously running Xbox survey shows an immediate change in support for Obama the day after the debate. Moreover, in addition to the national trends, Figure 5.5 shows how individual states reacted to the debates, a level of granularity that is often prohibitively expensive for traditional, representative polling.

Participants in the Xbox survey were allowed to answer the survey questions up to once per day, and many in fact submitted multiple responses over the course of the campaign. This allows us to construct an ad hoc panel, and in particular, to more closely examine the effects

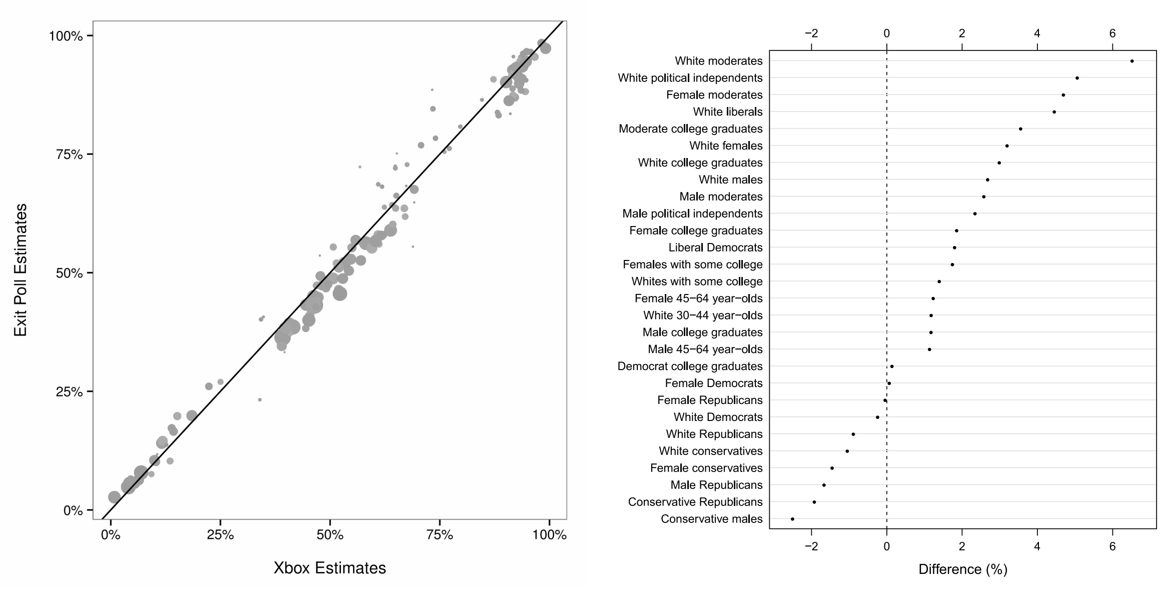


Figure 5.6: Obama support for various demographic subgroups, as estimated from the Xbox data (with MRP adjustment) and national exit polls.

Note: For various two-dimensional demographic subgroups (e.g., women aged 65 or older), the left panel compares two-party Obama support as estimated from the 2012 national exit poll and from the Xbox data on the day before the election. The sizes of the dots are proportional to the population sizes of the corresponding subgroups. The right panel shows the differences between the Xbox and exit poll estimates for the 30 largest such subgroups, ordered by the difference. Positive values indicate the Xbox estimate is larger than the corresponding exit poll estimate. Among these 30 subgroups, the median and mean absolute differences are 1.9 and 2.2 percentage points, respectively.

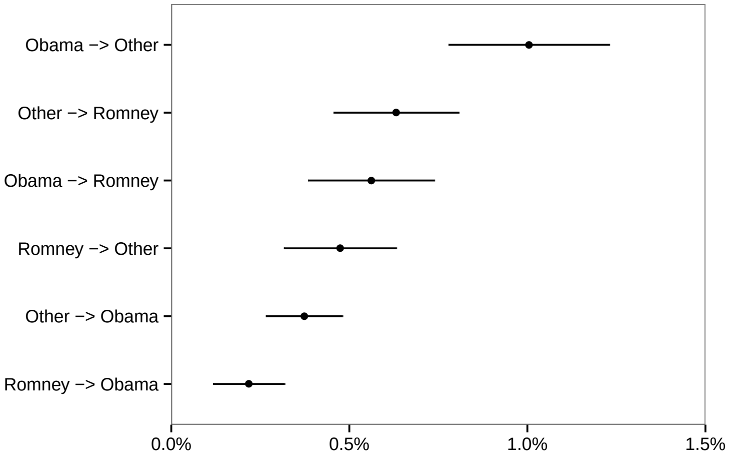


Figure 5.7: *Estimated proportion of the electorate that switched their support from one candidate to another during the one week immediately before and after the first presidential debate, with 95% confidence intervals.*

Note: The majority of switching was not between major party candidates, but between a major party candidate and "other" (i.e., neither Obama nor Romney).

of the first debate on voter sentiment. Specifically, we consider the subset of people who responded to the Xbox survey in the week before and after the first debate, and statistically adjust the responses with MRP to estimate the rate at which individuals switched their support for candidates. Figure 5.7 shows that 3% switched their support. However, the vast majority of the switches were not from one major-party candidate to the other. Instead, we find people primarily moved from Obama to "other" (i.e., neither Obama nor Romney) or from "other" to Romney. In fact, only 0.5% switched from Obama to Romney, and 0.2% from Romney to Obama. The often transient preferences of voters is important for understanding and optimizing campaigns (Vavreck, 2014), but without high-frequency polling it is difficult to reliably catalog such movements.

Conclusion

Measurement of public opinion need not only be accurate, but also timely and cost-effective. In this chapter, we have argued that with proper statistical adjustment, non-representative polling can be all three. We close by discussing three practical challenges facing non-representative polling.

First, implementing our statistical procedure required detailed data. In the face of insufficient demographic information on respondents or inadequate population-level statistics, it would have been difficult to generate accurate forecasts from non-representative data. Further, while much of our procedure is fairly mechanical, selecting the appropriate modeling framework requires some care. Fortunately, at least with the Xbox data, the regression estimates are stable after including only a few key demographic variables (gender, age, state, race and party identification). Moreover, even relatively simple statistical corrections, such as raking, go a long way to de-biasing results, illustrating the robustness of the general approach.

Second, though the marginal cost of collecting data via the Xbox was substantially smaller than with traditional polling methods, there were considerable fixed costs to developing the platform. As non-representative polling becomes more popular, we anticipate it will become increasing easy and inexpensive to conduct such polls. Indeed, in recent related work (Goel, Obeng, & Rothschild, 2015), we have shown that one can conduct accurate non-representative polls on Amazon's Mechanical Turk labor marketplace for one-tenth the cost and time of RDD surveys.

Third, opt-in surveys are potentially more prone to manipulation than polls of randomly sampled respondents. This is an admittedly difficult problem to resolve fully—and one that has been found to afflict prediction markets (Rothschild & Sethi, 2014)—but there are steps one can take to mitigate the issue. For example, one can limit the ability of individuals to vote multiple times by requiring users to log in to the polling system (as with the surveys we administered on Xbox and Mechanical Turk). Similarly, to prevent automated submissions, the survey can incorporate challenge questions that are easy for a human to answer but difficult for a computer (e.g., CAPTCHAs). Multiple voting can also be detected and blocked by monitoring the IP addresses of submissions and looking out for suspicious spikes in traffic. Finally, one can take a hybrid approach to polling in which a pool of individuals is selected for participation (which would limit malicious behavior) but without the requirement they be representative of the target population.

Looking forward, embracing non-representative polling opens a variety of possibilities for gauging real-time public sentiment. For example, in addition to the primary Xbox polls described in this chapter, we interviewed nearly 50,000 people as they watched the three presidential debates, with the survey questions overlaid on the debate broadcast. Standard representative polling will certainly continue to be an invaluable tool for the foreseeable future. However, non-representative polling (followed by appropriate post-survey adjustment) is due for further exploration, both for election forecasting and for social research more generally.

References

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4: Linear mixed-effects models using s4 classes.* Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 0.999999-2)

Battaglia, M. P., Hoaglin, D. C., & Frankel, M. R. (2013). Practical considerations in raking survey data. *Survey Practice*, *2*(5).

Cavan Reilly, A. G., & Katz, J. N. (2001). Post-stratification without population level information on the post-stratifying variable, with application to political polling. *Journal of the American Statistical Association*, *96*, 1–11.

Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science*.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

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

Ghitza, Y., & Gelman, A. (2013). Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, *57*(3), 762–776.

Goel, S., Obeng, A., & Rothschild, D. (2015). Non-representative surveys: Fast, cheap, and mostly accurate. *Working paper*. Retrieved from `http://researchdmr.com/FastCheapAccurate`

Hillygus, D. S., & Shields, T. G. (2009). *The persuadable voter: Wedge issues in presidential campaigns.* Princeton University Press.

Kaufmann, K. M., & Petrocik, J. R. (1999). The changing politics of american men: Understanding the sources of the gender gap. *American Journal of Political Science*, *43*(3), 864–887.

Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the states? *American Journal of Political Science*, *53*(1), 107–121.

Little, R. J. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, *88*(423), 1001–1012.

Lock, K., & Gelman, A. (2010). Bayesian combination of state polls and election forecasts. *Political Analysis*, *18*(3), 337–348.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with post-stratification: state-level estimates from national polls. *Political Analysis*, *12*(4), 375–385.

Pew Research Center. (2014a). *Methodology - our survey methodology in detail.* Retrieved from `http://www.people-press.org/methodology/our-survey-methodology-in-detail/`

Pew Research Center. (2014b). *Methodology - sampling - cell-phones.* Retrieved from `http://www.people-press.org/methodology/sampling/cell-phones/`

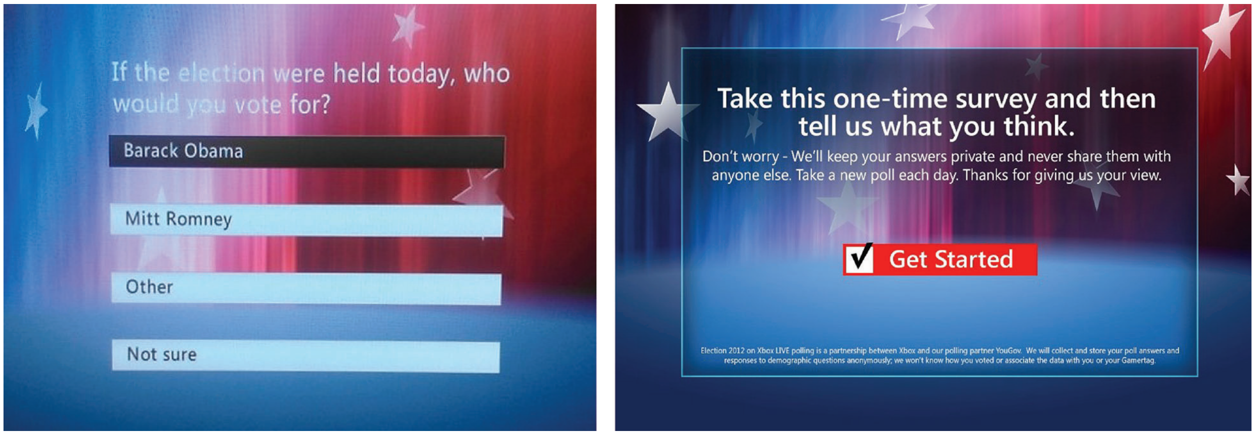
Pollster.com. (2004). *Should pollsters weight by party identification?* Retrieved from http://www.pollster.com/faq/should_pollsters_weight_by_par.php

Rothschild, D. (2015). Combining forecasts for elections: Accurate, relevant, and timely. *International Journal of Forecasting*, *31*(3), 952-964.

Rothschild, D., & Sethi, R. (2014). *Trading strategies and market microstructure: Evidence from a prediction market.* (Working paper)

Vavreck, L. (2014). *The secret about undecided voters: They're predictable.* New York Times. Retrieved from http://www.nytimes.com/2014/11/05/upshot/the-secret-about-undecided-voters-theyre-predictable.html

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*.



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

Take this one-time survey and then tell us what you think.

Don't worry - We'll keep your answers private and never share them with anyone else. Take a new poll each day. Thanks for giving us your view.

Election 2012 on Xbox LIVE polling is a partnership between Xbox and our polling partner YouGov. We will collect and store your poll answers and responses to demographic questions anonymously; we won't know how you voted or associate the data with you or your Gamertag.

Figure 5.8: Screenshots.

Note: 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.

Appendix

Battleground States

For ease of interpretation, in Figure 5.1 we group states into 4 categories: (1) battleground states (Colorado, Florida, Iowa, New Hampshire, Ohio, and Virginia), the five states with the highest amounts of TV spending plus New Hampshire, which had the highest per-capita spending; (2) quasi-battleground states (Michigan, Minnesota, North Carolina, Nevada, New Mexico, Pennsylvania, and Wisconsin), which round out the states where the campaigns and their affiliates made major TV buys; (3) solid Obama states (California, Connecticut, District of Columbia, Delaware, Hawaii, Illinois, Maine, Maryland, Massachusetts, New Jersey, New York, Oregon, Rhode Island, Vermont, and Washington); and (4) solid Romney states (Alabama, Alaska, Arizona, Arkansas, Georgia, Idaho, Indiana, Kansas, Kentucky, Louisiana, Mississippi, Missouri, Montana, Nebraska, North Dakota, Oklahoma, South Carolina, South Dakota, Tennessee, Texas, Utah, West Virginia, and Wyoming).

Questionnaire

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

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 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