International Journal of Forecasting 31 (2015) 980–991

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting elections with non-representative polls

Wei Wang [a,*], David Rothschild [b], Sharad Goel [b], Andrew Gelman [a,c]

[a] Department of Statistics, Columbia University, New York, NY, USA

[b] Microsoft Research, New York, NY, USA

[c] Department of Political Science, Columbia University, New York, NY, USA

ARTICLE INFO

Keywords:
Non-representative polling
Multilevel regression and poststratification
Election forecasting

ABSTRACT

Election forecasts have traditionally been based on representative polls, in which randomly sampled individuals are asked who they intend to vote for. While representative polling has historically proven to be quite effective, it comes at considerable costs of time and money. Moreover, as response rates have declined over the past several decades, the statistical benefits of representative sampling have diminished. In this paper, we show that, with proper statistical adjustment, non-representative polls can be used to generate accurate election forecasts, and that this can often be achieved faster and at a lesser expense than traditional survey methods. We demonstrate this approach by creating forecasts from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform. After adjusting the Xbox responses via multilevel regression and poststratification, we obtain estimates which are in line with the forecasts from leading poll analysts, which were based on aggregating hundreds of traditional polls conducted during the election cycle. We conclude by arguing that non-representative polling shows promise not only for election forecasting, but also for measuring public opinion on a broad range of social, economic and cultural issues.

1. Introduction

At the heart of modern opinion polling is representative sampling, built around the idea that every individual in a particular target population, such as registered or likely US voters, has the same probability of being sampled. From address-based, in-home interview sampling in the 1930s to random digit dialing after the growth of landlines and cellphones, leading polling organizations have put immense efforts into obtaining representative samples.

The wide-scale adoption of representative polling can be traced largely back to a pivotal polling mishap in the 1936 US presidential election campaign. During that campaign, the popular magazine *Literary Digest* conducted a mail-in survey that attracted over two million responses, a huge sample even by modern standards. However, the magazine incorrectly predicted a landslide victory for Republican candidate Alf Landon over the incumbent Franklin Roosevelt. In actual fact, Roosevelt won the election decisively, carrying every state except for Maine and Vermont. As pollsters and academics have pointed out since, the magazine's pool of respondents was highly biased: it consisted mostly of auto and telephone owners, as well as the magazine's own subscribers, which underrepresented Roosevelt's core constituencies (Squire, 1988). During that same campaign, various pioneering

* Corresponding author.
E-mail addresses: ww2243@columbia.edu (W. Wang), davidmr@microsoft.com (D. Rothschild), sharadg@microsoft.com (S. Goel), gelman@stat.columbia.edu (A. Gelman).

http://dx.doi.org/10.1016/j.ijforecast.2014.06.001
0169-2070/

pollsters, including George Gallup, Archibald Crossley, and Elmo Roper, used considerably smaller but representative samples, and predicted the election outcome with a reasonable level of accuracy (Gosnell, 1937). Accordingly, non-representative or "convenience sampling" rapidly fell out of favor with polling experts.

So, why do we revisit this seemingly long-settled case? Two recent trends spur our investigation. First, random digit dialing (RDD), the standard method in modern representative polling, has suffered increasingly high non-response rates, due both to the general public's growing reluctance to answer phone surveys, and to expanding technical means of screening unsolicited calls (Keeter, Kennedy, Dimock, Best, & Craighill, 2006). By one measure, RDD response rates have decreased from 36% in 1997 to 9% in 2012 (Kohut, Keeter, Doherty, Dimock, & Christian, 2012), and other studies confirm this trend (Holbrook, Krosnick, & Pfent, 2007; Steeh, Kirgis, Cannon, & DeWitt, 2001; Tourangeau & Plewes, 2013). Assuming that the initial pool of targets is representative, such low response rates mean that those who ultimately answer the phone and elect to respond might not be. Even if the selection issues are not yet a serious problem for accuracy, as some have argued (Holbrook et al., 2007), the downward trend in response rates suggests an increasing need for post-sampling adjustments; indeed, the adjustment methods we present here should work just as well for surveys obtained by probability sampling as for convenience samples. The second trend driving our research is the fact that, with recent technological innovations, it is increasingly convenient and cost-effective to collect large numbers of highly non-representative samples via online surveys. The data that took the *Literary Digest* editors several months to collect in 1936 can now take only a few days, and, for some surveys, can cost just pennies per response. However, the challenge is to extract a meaningful signal from these unconventional samples.

In this paper, we show that, with proper statistical adjustments, non-representative polls are able to yield accurate presidential election forecasts, on par with those based on traditional representative polls. We proceed as follows. Section 2 describes the election survey that we conducted on the Xbox gaming platform during the 45 days leading up to the 2012 US presidential race. Our Xbox sample is highly biased in two key demographic dimensions, gender and age, and, accordingly, the raw responses disagree with the actual outcomes. The statistical techniques we use to adjust the raw estimates are introduced in two stages. In Section 3, we construct daily estimates of voter intent via multilevel regression and poststratification (MRP). The central idea of MRP is to partition the data into thousands of demographic cells, estimate voter intent at the cell level using a multilevel regression model, and finally aggregate the cell-level estimates in accordance with the target population's demographic composition. One recent study suggested that non-probability samples provide worse estimates than probability samples (Yeager et al., 2011), but that study used simple adjustment techniques, not MRP. Even after getting good daily estimates of voter intent, however, more needs to be done to translate these into election-day forecasts. Section 4 therefore describes how to transform voter intent into projections of vote share and electoral votes. We conclude in Section 5 by discussing the potential for non-representative polling in other domains.

2. Xbox data

Our analysis is based on an opt-in poll which was available continuously on the Xbox gaming platform during the 45 days preceding the 2012 US presidential election. Each day, three to five questions were posted, one of which gauged voter intention via the standard query, "If the election were held today, who would you vote for?". Full details of the questionnaire are given in the Appendix. The respondents were allowed to answer at most once per day. The first time they participated in an Xbox poll, respondents were also asked to provide basic demographic information about themselves, including their sex, race, age, education, state, party ID, political ideology, and who they voted for in the 2008 presidential election. In total, 750,148 interviews were conducted, with 345,858 unique respondents – over 30,000 of whom completed five or more polls – making this one of the largest election panel studies ever.

Despite the large sample size, the pool of Xbox respondents is far from being representative of the voting population. Fig. 1 compares the demographic composition of the Xbox participants to that of the general electorate, as estimated via the 2012 national exit poll.[1] The most striking differences are for age and sex. As one might expect, young men dominate the Xbox population: 18- to 29-year-olds comprise 65% of the Xbox dataset, compared to 19% in the exit poll; and men make up 93% of the Xbox sample but only 47% of the electorate. Political scientists have long observed that both age and sex are strongly correlated with voting preferences (Kaufmann & Petrocik, 1999), and indeed these discrepancies are apparent in the unadjusted time series of Xbox voter intent shown in Fig. 2. In contrast to estimates based on traditional, representative polls (indicated by the dotted blue line in Fig. 2), the uncorrected Xbox sample suggests a landslide victory for Mitt Romney, reminiscent of the infamous *Literary Digest* error.

3. Estimating voter intent with multilevel regression and poststratification

3.1. Multilevel regression and poststratification

To transform the raw Xbox data into accurate estimates of voter intent in the general electorate, we make use of the

[1] For ease of interpretation, in Fig. 1 we group the states into four 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).

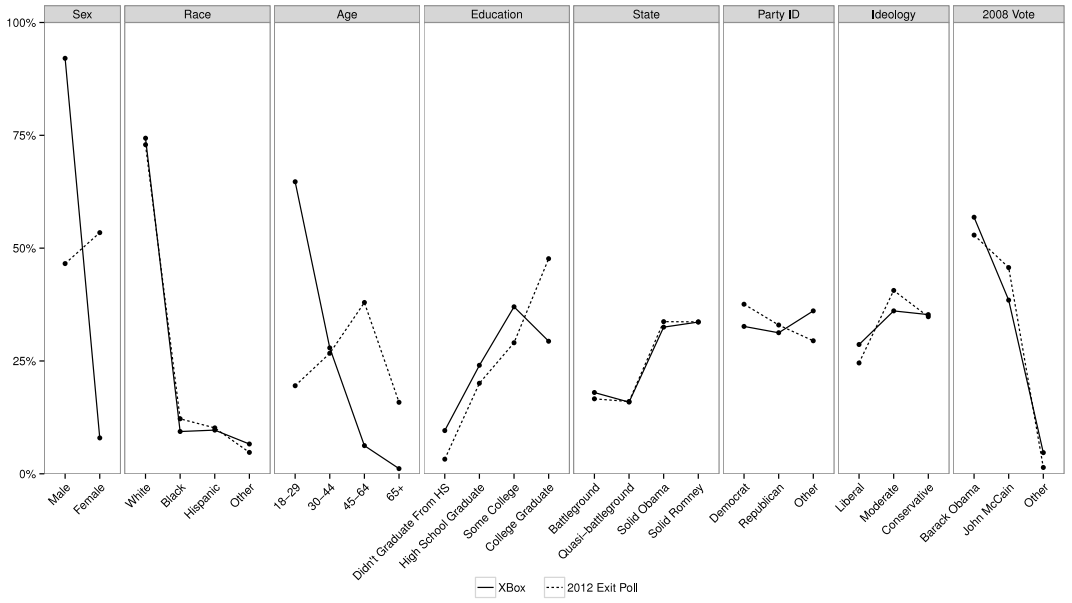

Fig. 1. A comparison of the demographic, partisan, and 2008 vote distributions in the Xbox dataset and the 2012 electorate (as measured by adjusted exit polls). As one might expect, the sex and age distributions exhibit considerable differences.

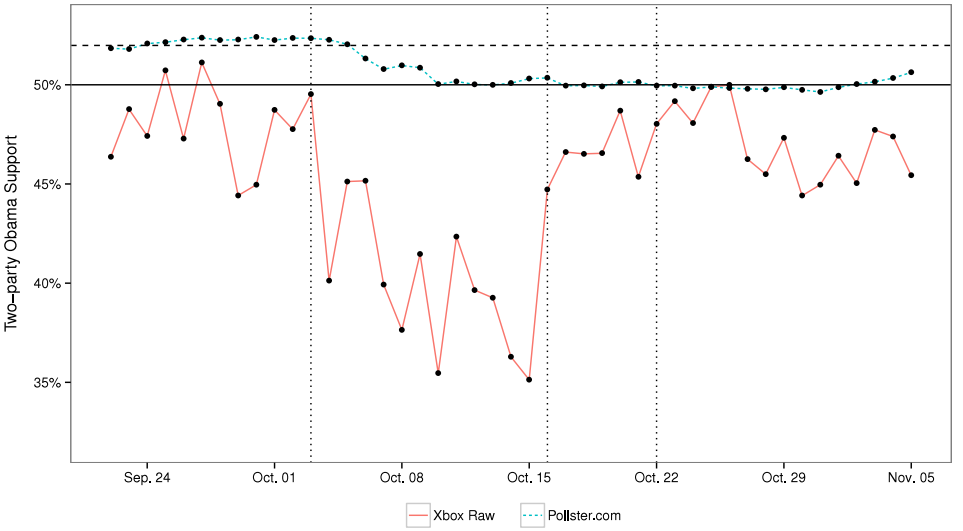

Fig. 2. Daily (unadjusted) Xbox estimates of the two-party Obama support during the 45 days leading up to the 2012 presidential election, which suggest a landslide victory for Mitt Romney. The dotted blue line indicates a consensus average of traditional polls (the daily aggregated polling results from 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. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

rich demographic information that respondents provide. In particular, we *poststratify* the raw Xbox responses to mimic a representative sample of likely voters. Poststratification is a popular method for correcting for known differences between sample and target populations (Little, 1993). The core idea is to partition the population into cells based on combinations of various demographic and political attributes, use the sample to estimate the response variable within each cell, and finally 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

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

where $\hat{y}_j$ is the estimate of y in cell j, and N_j is the size of the jth cell in the population. Analogously, we can derive an estimate of y at any subpopulation level s (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 and estimates of 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 these 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 (Ghitza & Gelman, 2013; Lax & Phillips, 2009; Park, Gelman, & Bafumi, 2004).

More formally, applying MRP in our setting comprises two steps. First, we fit a Bayesian hierarchical model to obtain estimates for sparse poststratification cells; second, we average over the cells, weighting the values 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 sex (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), thus partitioning the data into 176,256 cells.[2] Including the political variables is important because they are strong predictors of vote intentions. Poststratifying based on party identification has sometimes been controversial (Blumenthal, 2004), but we are comfortable with it here, first because it tends to vary more slowly than vote intentions and political attitudes (Cavan Reilly & Katz, 2001; Gelman & King, 1993), and second because party identification and the other background variables in the Xbox survey were measured only once during the campaign, at the time of a respondent's entry into the panel.

We fit two nested multilevel logistic regressions for estimating 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 and Hill (2007), the first model is given by

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

where α_0 is the fixed baseline intercept and α_1 is the fixed slope for Obama's fraction of the 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 we denote in general by $a_{j[i]}^{\text{var}}$ – correspond to the varying coefficients associated with each categorical variable. Here, the subscript $j[i]$ indicates the cell to which the ith 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 ith respondent. The varying coefficients $a_{j[i]}^{\text{var}}$ are given by 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 a multilevel model is that the 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

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

and

$$\begin{aligned}
b_{j[i]}^{\text{var}} &\sim N(0, \eta_{\text{var}}^2), \\
\eta_{\text{var}}^2 &\sim \text{inv-}\chi^2(\mu, \eta_0^2).
\end{aligned}$$

Jointly, Eqs. (1) and (2) define a Bayesian model that describes the data. Ideally, we would perform a fully Bayesian analysis in order to obtain the posterior distribution of the parameters. However, for the sake of 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 run the multilevel model daily using a four-day moving window, aggregating the data collected on that day and the previous three days, so as to produce 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 the cell-level estimates are weighted by the proportion of the electorate in each cell and aggregated to the appropriate level (i.e., state or national). To compute cell weights, we require

[2] All demographic variables are collected prior to a respondent's first poll, thus alleviating concerns that respondents may adjust their demographic responses to be in line with their voter intention (e.g., a new Obama supporter switching his or her party ID from Republican to Democrat).

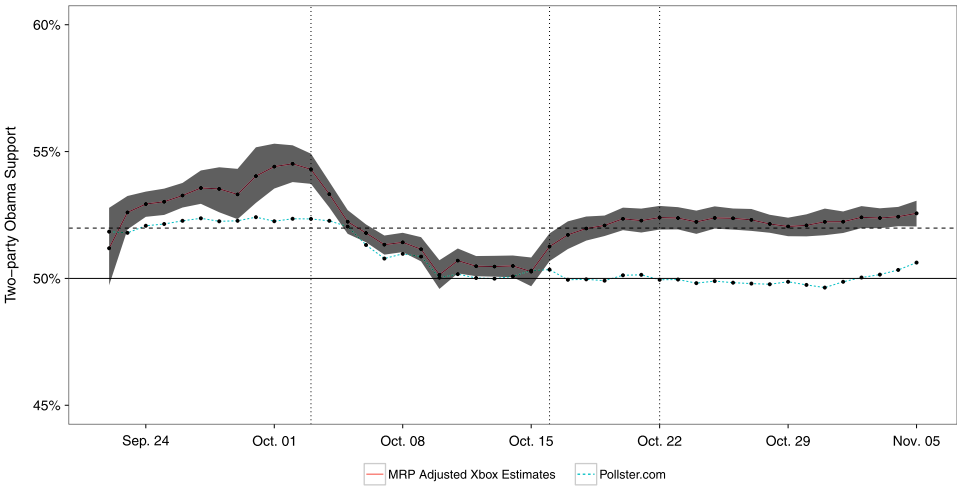

Fig. 3. National MRP-adjusted voter intent of two-party Obama support over the 45-day period, with the associated 95% confidence bands. The horizontal dashed line indicates the actual two-party Obama vote share. The three vertical dotted lines indicate the presidential debates. Compared with the raw responses in Fig. 2, the MRP-adjusted voter intent is much more reasonable, and the voter intent in the last few days is close to the actual outcome. On the other hand, the daily aggregated polling results from Pollster.com, shown by the blue dotted line, are further away from the actual vote share than the estimates generated from the Xbox data in the last few days. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cross-tabulated population data. One commonly used source for such data is the Current Population Survey (CPS); however, the CPS is missing some key poststratification variables, such as party identification. We therefore use exit poll data from the 2008 presidential election instead. Exit polls are conducted outside voting stations on election day, and record the choices of exiting voters; 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 polls from 2008, not 2012, because this means that, in theory, the method we describe here could have been used to generate real-time predictions during the 2012 election campaign. Admittedly, this approach puts our prediction at a disadvantage, since we cannot capture the demographic shifts of the intervening four years. While combining exit poll and CPS data could arguably yield improved results, we limit ourselves to the 2008 exit poll summaries for our poststratification, for the sake of simplicity and transparency.

3.2. National and state voter intent

Fig. 3 shows the adjusted two-party Obama support for the last 45 days before the election. Compared with the uncorrected estimates in Fig. 2, the MRP-adjusted estimates yield a much more reasonable timeline of Obama's standing over the final weeks of the campaign. With a clear advantage at the beginning, Obama's support slipped rapidly after the first presidential debate – though it never fell below 50% – then gradually recovered, building up a decisive lead in the final days.

On the day before the election, our estimate of voter intent is off from the actual outcome (indicated by the dotted horizontal line) by a mere 0.6 percentage points. The voter intent in the weeks prior to the election does not equate directly to an estimate of the vote share on election day—a point we return to in Section 4. As such, it is difficult to evaluate the accuracy of our full time series of estimates. However, not only are our estimates intuitively reasonable, they are also in line with the prevailing estimates based on traditional, representative polls. In particular, our estimates roughly track – and are even arguably better than – those from Pollster.com, one of the leading poll aggregators during the 2012 campaign. We are following what is now standard practice by using an aggregate of nationally reported polls for the comparison, rather than any selected subset of polls that might be judged to be of highest quality. Again, our goal here is not to bash conventional polling but rather to demonstrate how, with sophisticated adjustments, we can match that performance even with a highly non-representative opt-in sample.

The national vote share receives a considerable amount of media attention, but state-level estimates are particularly relevant for many stakeholders, given the role of the Electoral College in selecting the winner (Rothschild, 2013). Forecasting state-by-state races is a challenging problem, due to the interdependencies in state outcomes, the logistical difficulties of measuring state-level vote preferences, 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. Specifically, we use the same cell-level estimates as are employed in the national estimate, that is, those generated via the multilevel model in Eqs. (1) and (2), and then poststratify to each state's demographic composition. 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.

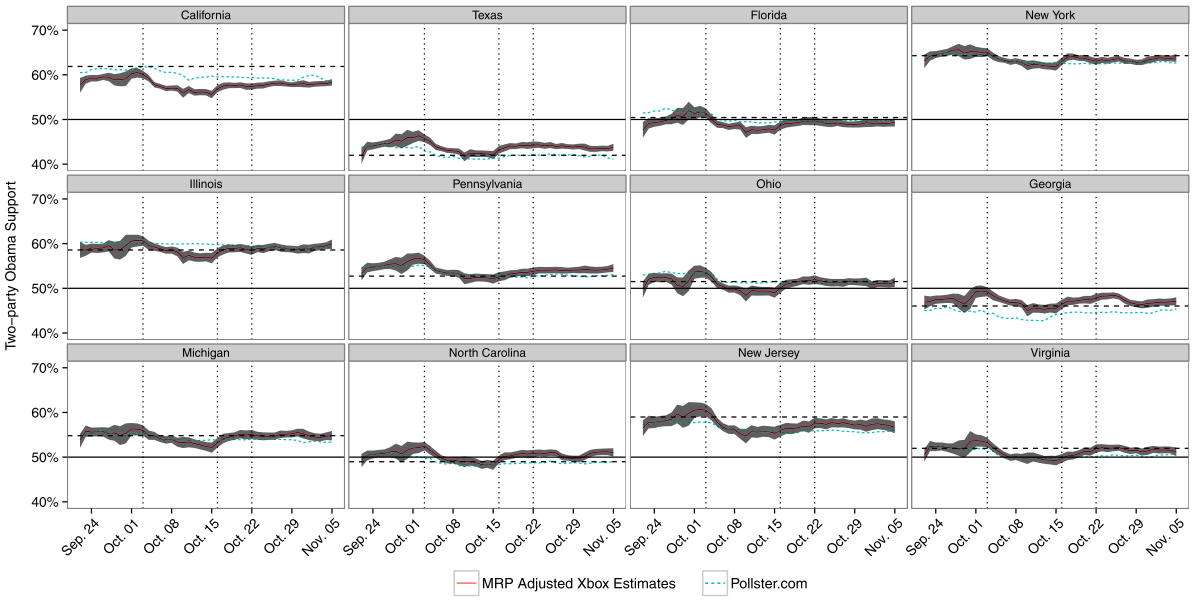

Fig. 4. MRP-adjusted daily voter intent for the 12 states with the most electoral votes, and the associated 95% confidence bands. The horizontal dashed lines in each panel give the actual two-party Obama vote shares in that state. The mean and median absolute errors of the last day voter intent across 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 blue lines, are broadly consistent with the estimates from the Xbox data. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 4 shows the two-party Obama support for the 12 states with the most electoral votes. The state timelines share similar trends (e.g., the support for Obama dropping after the first debate), but also have their own idiosyncratic movements, which is an indication of a reasonable blend of national and state-level signals. To demonstrate the accuracy of the MRP-adjusted estimates, we plot the estimates generated by Pollster.com in Fig. 4 (dotted blue lines); these 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.

3.3. Voter intent for demographic subgroups

Apart from the Electoral College races, election forecasting often focuses on candidate preferences 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 may plausibly suffer from strong self-selection problems. This begs the question, can we reasonably expect to estimate the views of older women based on a platform that caters largely to young men?

With MRP, it is straightforward 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 the 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, Blacks, etc.) on the day before the election, and compare those with the actual voting behaviors of those same groups, as estimated by the 2012 national exit poll. As can be seen in Fig. 5, the Xbox estimates are remarkably accurate, with median absolute differences of 1.5 percentage points between the Xbox and exit poll numbers.[3]

Not only do the Xbox data facilitate the accurate estimation of voter intent across these single-dimensional demographic categories, they also do surprisingly well at estimating two-way interactions (e.g., candidate support among 18–29 year-old Hispanics, and liberal college graduates). Fig. 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.[4] Most points lie close to the diagonal, indicating that the Xbox and exit poll estimates are in agreement. Specifically, for women aged 65 and older – a group whose preferences one might expect a priori to be hard to estimate from the Xbox data – the difference between Xbox and the exit poll is a mere one percentage point (49.5% and 48.5%, respectively). Across all of the two-way interaction groups, the median

[3] The respondents' 2008 votes were not asked on the 2012 exit polls, so we exclude that comparison from Fig. 5.

[4] State contestedness is excluded from the two-way interaction groups, since the 2012 state exit polls are not yet available, and the 2012 national exit polls do not have enough data to estimate state interactions reliably. The 2008 vote is also excluded, as it was not asked in the 2012 exit poll. The "other" race category was dropped because it was not defined consistently across the Xbox and exit poll datasets.

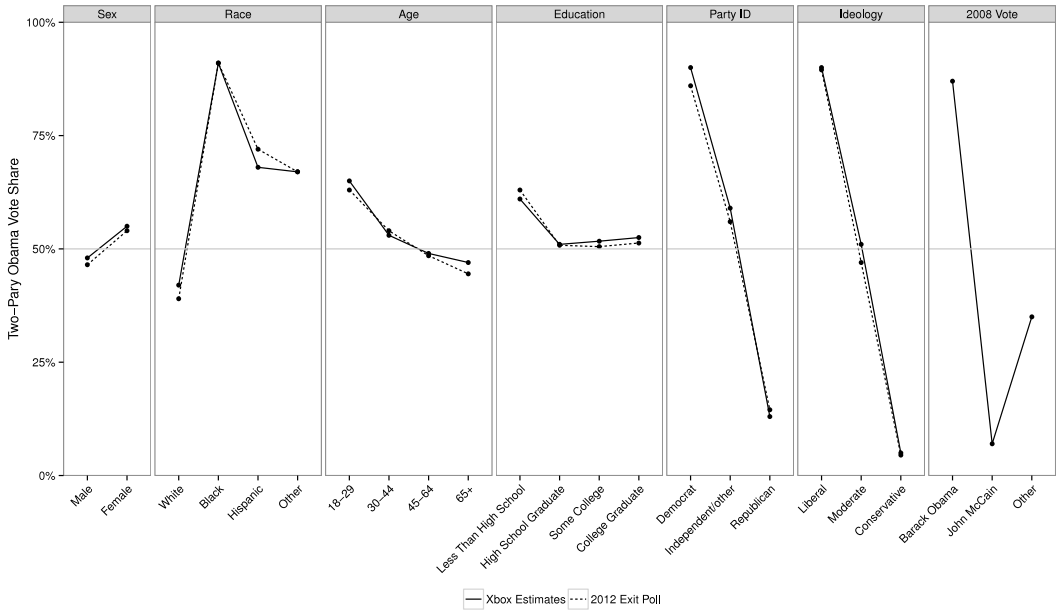

Fig. 5. Comparison of the two-party Obama vote share for various demographic subgroups, as estimated from the 2012 national exit poll and from the Xbox data on the day before the election.

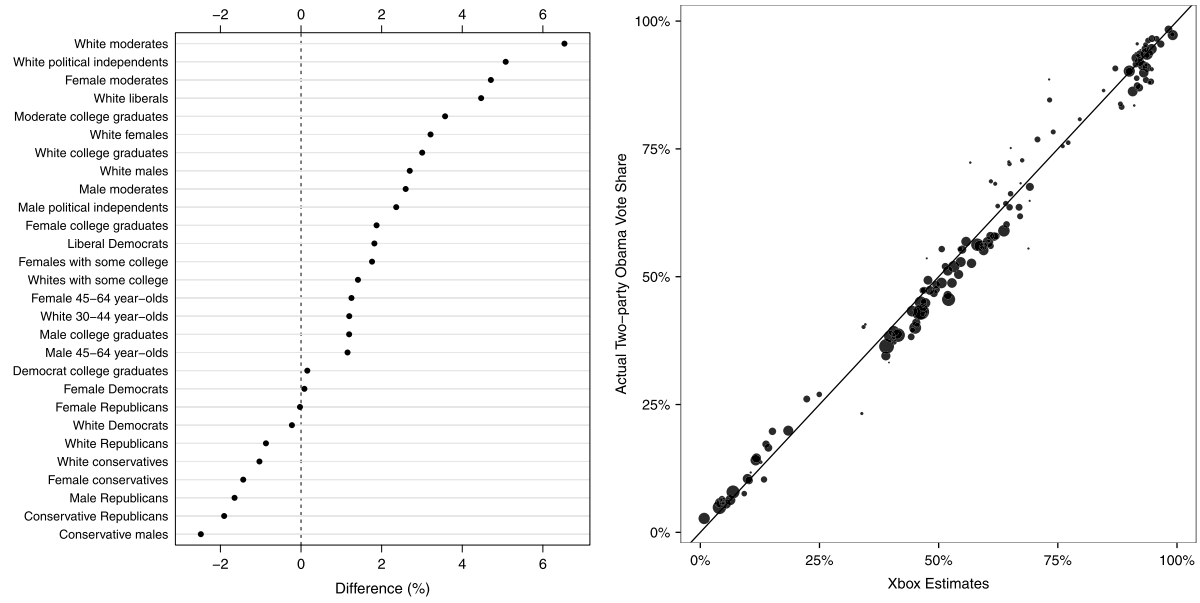

Fig. 6. Left panel: Differences between the Xbox MRP-adjusted estimates and the exit poll estimates for the 30 largest two-dimensional demographic subgroups, ordered by the differences. Positive values indicate that the Xbox estimate is larger than the corresponding exit poll estimate. Among these 30 subgroups, the median and mean absolute differences are 1.9 and 2.2 percentage points, respectively. Right panel: Two-party Obama support, as estimated from the 2012 national exit poll and from the Xbox data on the day before the election, for various two-way interaction demographic subgroups (e.g., 65+ year-old women). The sizes of the dots are proportional to the population sizes of the corresponding subgroups.

absolute difference is just 2.4 percentage points. As the size of the points in Fig. 6 indicates, 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, Fig. 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.

4. Forecasting election day outcomes

4.1. Converting voter intent to forecasts

As was mentioned above, daily estimates of voter intent do not translate directly to estimates of vote share on election day. There are two key factors in this deviation. First, opinion polls (both representative and non-representative

ones) only gauge voter preferences on the particular day when the poll is conducted, with the question typically being phrased as, "if the election were held today". Political scientists and pollsters have long observed that such stated preferences are prone to several biases, including the anti-incumbency bias, in which the incumbent's polling numbers tend to be lower than the ultimate outcome (Campbell, 2008; Erikson & Wlezien, 2008; Rothschild, 2013), and the fading early lead bias, in which a big lead early in the campaign tends to diminish as the election gets closer (Erikson & Wlezien, 2008). Moreover, voters' attitudes are affected by information revealed over the course of the campaign, so preferences expressed weeks or months before election day are at best a noisy indicator of individuals' eventual votes. Second, estimates of vote shares require a model of likely voters. That is, opinion polls measure preferences among a hypothetical voter pool, and thus are accurate only to the extent that this pool captures those who actually turn out to vote on election day. Both of these factors introduce significant complications in forecasting election day outcomes.

To convert daily estimates of voter intent into election day predictions – which we refer to hereafter as *calibrating* voter intent – we compare the daily voter intent in previous elections to the ultimate outcomes in those elections. Specifically, we collect historical data from three previous US presidential elections, in 2000, 2004, and 2008. For each year, we obtain top-line (i.e., not individual-level) national and state estimates of voter intent from all available polls conducted in those elections.[5] From this collection of polling data, we then construct daily estimates of voter intent by taking a moving average of the poll numbers, in a manner similar to the major poll aggregators. We rely on traditional, representative polls to reconstruct historical voter intent; in principle, however, we could have started with non-representative polls if such data had been available in previous election cycles.

We next infer a mapping from voter intent to election outcomes by regressing the election day vote share on the historical time series of voter intent. The key difference between our approach and previous related work (Erikson & Wlezien, 2008; Rothschild, 2009) is that we model state-level correlations explicitly, via nested national and state models and correlated error terms. Specifically, we first fit a national model, given by

$$y_e^{\mathrm{US}} = a_0 + a_1 x_{t,e}^{\mathrm{US}} + a_2 |x_{t,e}^{\mathrm{US}}| x_{t,e}^{\mathrm{US}} + a_3 t x_{t,e}^{\mathrm{US}} + \eta(t, e)$$

where y_e^{US} is the national election day vote share of the incumbent party candidate in election year e, $x_{t,e}^{\mathrm{US}}$ is the national voter intent of the incumbent party candidate t days before the election in year e, and $\eta \sim N(0, \sigma^2)$ is the error term. Both y_e^{US} and $x_{t,e}^{\mathrm{US}}$ are offset by 0.5, so the values run from -0.5 to 0.5 rather than 0 to 1. The term involving the absolute value of voter intent pulls the vote share prediction toward 50%, capturing the diminishing early lead effect. We do not include a main effect for time, since it seems unlikely that the number of days until the election makes any direct contribution to the final vote share; rather, time contributes through its interaction with the voter intent (which we do include in the model).

Similarly, the state model is given by

$$y_{s,e}^{\mathrm{ST}} = b_0 + b_1 x_{s,t,e}^{\mathrm{ST}} + b_2 |x_{s,t,e}^{\mathrm{ST}}| x_{s,t,e}^{\mathrm{ST}} + b_3 t x_{s,t,e}^{\mathrm{ST}} + \varepsilon(s, t, e),$$

where $y_{s,e}^{\mathrm{ST}}$ is the election day state vote share of the state's incumbent party candidate[6] on day t, $x_{s,t,e}^{\mathrm{ST}}$ is the state voter intent on day t, and ϵ is the error term. The outcome $y_{s,e}^{\mathrm{ST}}$ is offset by the national projected vote share on that day, as fit using the national calibration model, and $x_{s,t,e}^{\mathrm{ST}}$ is offset by that day's national voter intent. Furthermore, we impose two restrictions on the magnitude and correlation structure of the error term $\varepsilon(s, t, e)$. First, since the uncertainty naturally decreases as the election gets closer (as t becomes smaller), we apply the heteroscedastic structure $\mathrm{Var}(\varepsilon(s, t, e)) = (t + a)^2$, where a is a constant to be estimated from the data. Second, the state-specific movements within each election year are allowed to be correlated. For simplicity, and as per Chen, Ingersoll, and Kaplan (2008), we assume that these correlations are uniform (i.e., all pairwise correlations are the same), which creates one more parameter to be estimated from the data. We fit the full calibration model with the `gls()` function in the `R` package `nlme` (Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2012).

In summary, the procedure for generating election day forecasts proceeds in three steps:

1. Estimate the joint distribution of state and national voter intents by applying MRP to the Xbox data, as described in Section 3.
2. Fit the nested calibration model described above to historical data in order to obtain point estimates for the parameters, including estimates for the error terms.
3. Convert the distribution of voter intent to election day forecasts via the fitted calibration model.

4.2. National and state election day forecasts

Fig. 7 plots the projected vote shares and pointwise 95% confidence bands over time for the 12 states with the most electoral votes. Though these time series look quite reasonable, it is difficult to assess their accuracy, as there are no ground truth estimates to compare them with in the weeks prior to the election. As a starting point, we compare our state-level estimates to those generated by prediction markets, which are widely considered to be among the most accurate sources for political predictions (Rothschild, 2013; Wolfers & Zitzewitz, 2004). For each state, prediction markets produce daily probabilities of victory. Although Fig. 7 plots our forecasts in terms of the expected vote share, our estimation procedure in fact yields the full distribution of outcomes, and thus, we can likewise convert our estimates into probabilistic forecasts. Fig. 8 shows this comparison, where the prediction market estimate is derived by averaging the two largest

[5] We collected the polling data from Pollster.com and RealClearPolitics.com.

[6] State incumbent parties are defined as the state-by-state winners from the previous election, which is more meaningful in this context than simply using the national incumbent.

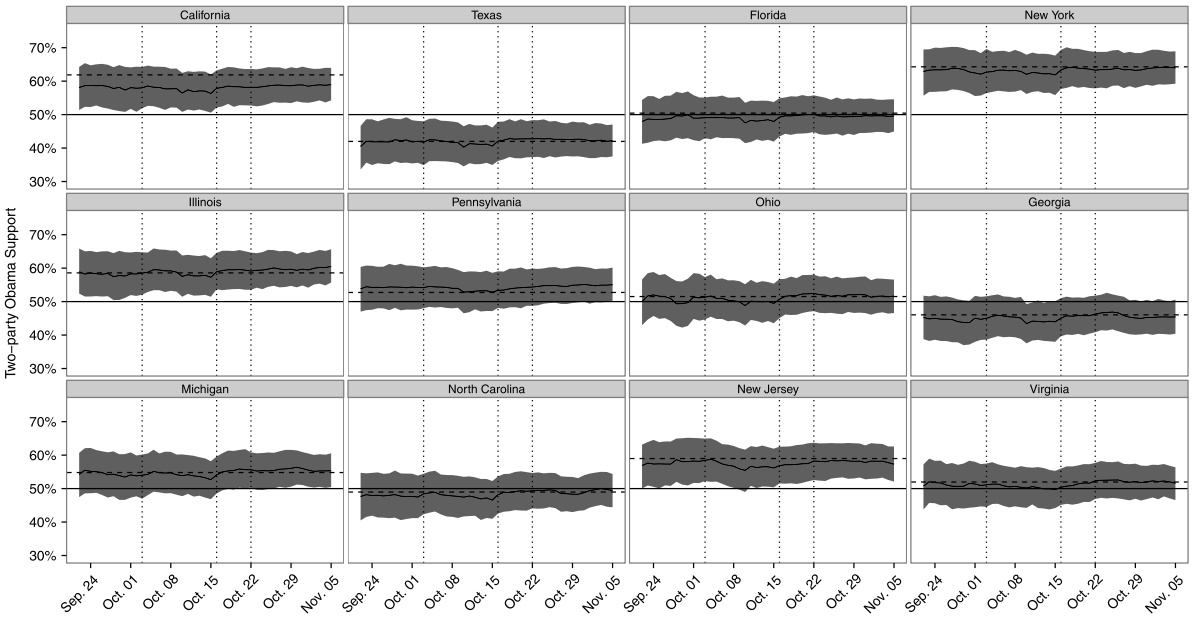

Fig. 7. Projected Obama share of the two-party vote on election day for each of the 12 states with the most electoral votes, with the associated 95% confidence bands. Compared to the MRP-adjusted voter intent in Fig. 4, the projected two-party Obama support is more stable, and the North Carolina race switches direction after applying the calibration model. In addition, the confidence bands become much wider and give more reasonable state-by-state probabilities of Obama victories.

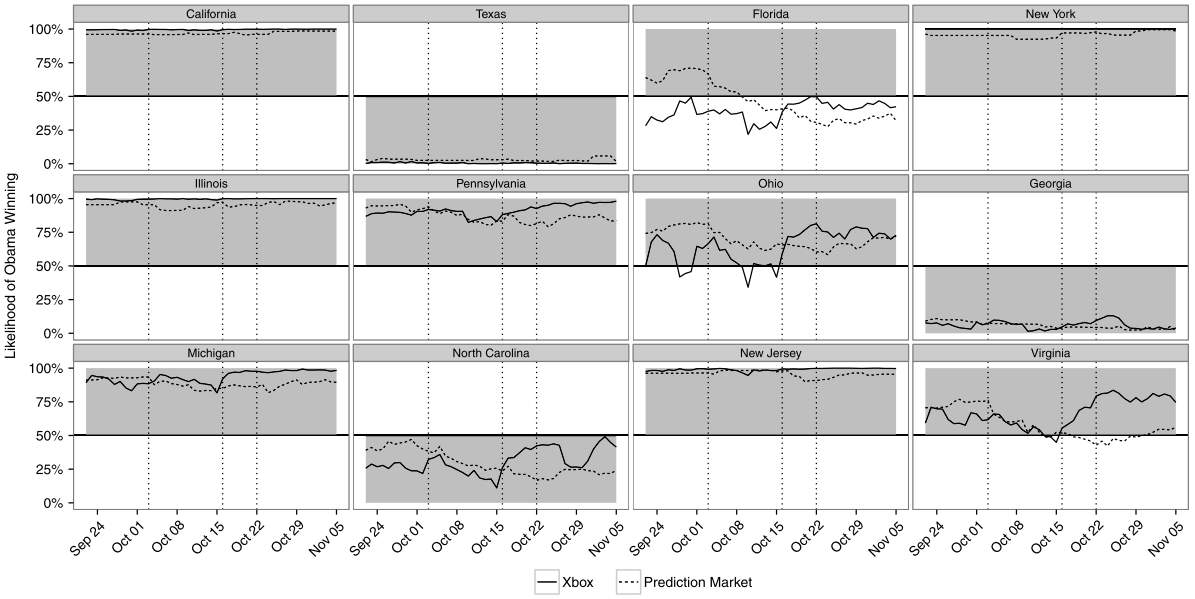

Fig. 8. Comparison between the probabilities of Obama winning the 12 largest Electoral College races based on Xbox and prediction market data. The prediction market data are the average of the raw Betfair and Intrade prices from winner-take-all markets. The three vertical lines represent the dates of the three presidential debates. The shaded halves indicate the direction in which race went.

election markets, Betfair and Intrade. Our probabilistic estimates are largely consistent with the prediction market probabilities. In fact, for races with little uncertainty, such as in Texas and Massachusetts, the Xbox estimates do not seem to suffer from the longshot bias that is common to prediction markets (Rothschild, 2009), and instead yield probabilities which are closer to 0 or 1. For tighter races, the Xbox estimates – while still highly correlated with the prediction market probabilities – look more volatile, especially in the early part of the 45-day period. Since the ground truth is not clearly defined, it is difficult to evaluate which method – Xbox or prediction markets – yields better results. From a Bayesian perspective, if one believes the stability shown by prediction markets, this could be incorporated into the structure of the Xbox calibration model.

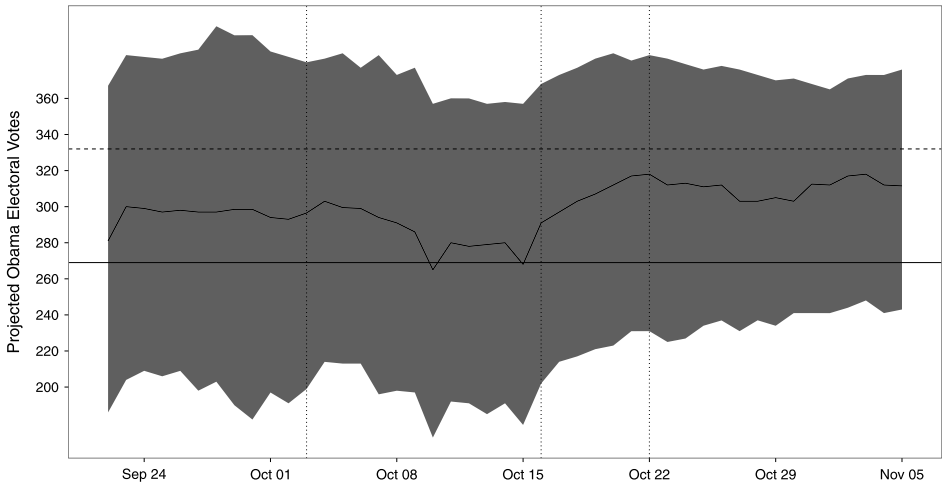

Fig. 9. Daily projections of Obama electoral votes over the 45-day period leading up to the 2012 election, with the associated 95% confidence bands. The solid line represents the median of the daily distribution. The horizontal dashed line represents the actual electoral votes, 332, that Obama captured in 2012 election. The three vertical dotted lines indicate the dates of the three presidential debates.

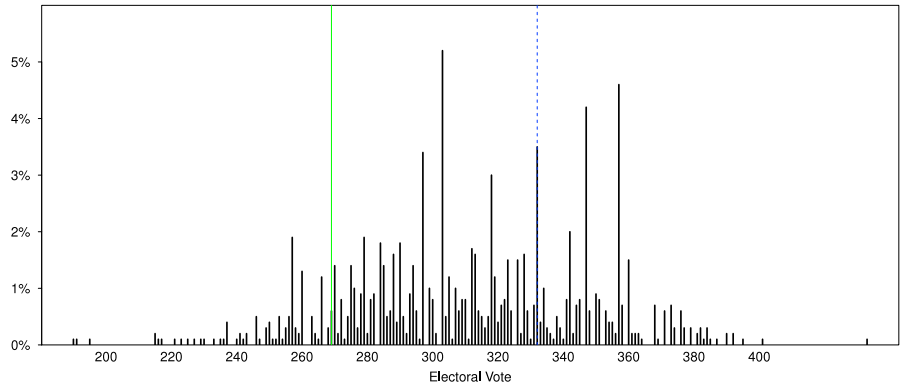

Fig. 10. The projected distribution of electoral votes for Obama one day before the election. The green vertical dotted line represents 269, the minimum number of electoral votes that Obama needed for a tie. The blue vertical dashed line indicates 332, the actual number of electoral votes captured by Obama. The estimated likelihood of Obama winning the electoral vote is 88%. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

With the full state-level outcome distribution, we can also estimate the distribution of Electoral College votes. Fig. 9 plots the median projected electoral votes for Obama over the last 45 days before the election, together with the 95% confidence bands. In particular, on the day before the election, our model estimates that Obama had an 88% chance of victory, in line with the estimates based on traditional polling data. For example, Obama was given a 91% chance of victory according to a method built by Jackman (2005). Zooming in on the day before the election, Fig. 10 shows the full predicted distribution of electoral votes for Obama. While Obama actually captured 332 votes, we estimate a median of 312 votes, with the most likely outcome being 303. Though this distribution of Electoral College outcomes seems reasonable, it does appear to have a higher variance than one might expect. In particular, the extreme outcomes seem to have unrealistically high likelihoods of occurring, which is probably a result of our calibration model not capturing the state-level correlation structure fully. Nonetheless, given that our forecasts are based on a highly biased convenience sample of respondents, the model predictions are remarkably good.

5. Conclusion

Election forecasts need to be not only accurate, but also relevant, timely, and cost-effective. In this paper, we construct forecasts that satisfy all of these requirements using extremely non-representative data. Though our data were collected on a proprietary polling platform, in principle one could aggregate such non-representative samples at a fraction of the cost of conventional survey designs. Moreover, the data produce forecasts that are both relevant and timely, as they can be updated faster and more regularly than standard election polls. Thus, the key aim – and one of the main contributions of this paper – is to assess the extent to which one can generate accurate predictions from non-representative samples. Since there is limited ground truth for election forecasts, establishing the accuracy of our predictions with any precision is

Fig. 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.

difficult. Nevertheless, we show that the MRP-adjusted and calibrated Xbox estimates are intuitively reasonable, and are also similar to those generated by more traditional means.

While our approach performed quite well, it did require detailed data. In the face of insufficient demographic information on respondents, inadequate population-level statistics, or a lack of historical election poll results, it would have been difficult to generate accurate forecasts from non-representative data. Furthermore, while much of our procedure is fairly mechanical, selecting the appropriate modeling framework requires some care. Fortunately, however, at least with the Xbox data, the regression estimates are stable after only a few key demographic variables (sex, age, state, race and party identification) have been included.

The greatest impact of non-representative polling will probably be for smaller, local elections and specialized survey settings, where it is impractical to deploy traditional methods due to cost and time constraints, rather than for presidential elections. For example, non-representative polls could be used in Congressional elections, where there are currently only sparse polling data. Non-representative polls could also supplement traditional surveys (e.g., the General Social Survey) by offering preliminary results at shorter intervals. Finally, when there is a need to identify and track pivotal events that affect public opinion, non-representative polling offers the possibility of cost-effective continuous data collection. Standard representative polling will certainly continue to be an invaluable tool for the foreseeable future. However, 75 years after the *Literary Digest* failure, non-representative polling (followed by appropriate post-data adjustment) is due for further exploration, for election forecasting and in social research more generally.

Acknowledgments

We thank Jon Krosnick for helpful comments and the National Science Foundation (Grant SES-1023189, SES-1023176 and SES-1205516) and the Institute of Education Sciences (Grant IES R305D100017) for partial support of this research.

Appendix. Questionnaire

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, 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 occasions, respondents were directed immediately to answer between three and five daily survey questions, one of which was always the vote intention question (see Fig. A.1).

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

References

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes*. URL: http://CRAN.R-project.org/package=lme4. R package version 0.999999-2.

Blumenthal, M. (2004). *Should pollsters weight by party identification?* URL: http://www.pollster.com/faq/should_pollsters_weight_by_par.php.

Campbell, J. E. (2008). *The American campaign: US presidential campaigns and the national vote (Volume 6)*. Texas A&M University Press.

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.

Chen, M. K., Ingersoll, J. E., & Kaplan, E. H. (2008). Modeling a presidential prediction market. *Management Science*, *54*, 1381–1394.

Erikson, R. S., & Wlezien, C. (2008). Are political markets really superior to polls as election predictors? *Public Opinion Quarterly*, *72*, 190–215.

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*, 762–776.

Gosnell, H. F. (1937). How accurate were the polls? *Public Opinion Quarterly*, *1*, 97–105.

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

Holbrook, A., Krosnick, J. A., & Pfent, A. (2007). *The causes and consequences of response rates in surveys by the news media and government contractor survey research firms* (pp. 499–528). Wiley.

Jackman, S. (2005). Pooling the polls over an election campaign. *Australian Journal of Political Science*, *40*, 499–517.

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*, 864–887.

Keeter, S., Kennedy, C., Dimock, M., Best, J., & Craighill, P. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly*, *70*, 759–779.

Kohut, A., Keeter, S., Doherty, C., Dimock, M., & Christian, L. (2012). *Assessing the representativeness of public opinion surveys*. Pew Research Center for The People & The Press.

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

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

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

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

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team, (2012). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-104.

Rothschild, D. (2009). Forecasting elections comparing prediction markets, polls, and their biases. *Public Opinion Quarterly*, *73*, 895–916.

Rothschild, D. (2013). *Combining forecasts: accurate, relevant, and timely*. Working paper.

Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, *52*, 125–133.

Steeh, C., Kirgis, N., Cannon, B., & DeWitt, J. (2001). Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *Journal of Official Statistics*, *17*, 227–248.

Tourangeau, R., & Plewes, T. J. (2013). *Nonresponse in social science surveys: a research agenda*. The National Academies Press, URL: http://www.nap.edu/openbook.php?record_id=18293.

Wolfers, J., & Zitzewitz, E. (2004). *Prediction markets*. National Bureau of Economic Research.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*, 709–747.