

# **Online Polls and Registration Based Sampling: A New Method for Pre-election Polling**

**Michael Barber**  
Princeton University

**Chris Mann**  
University of Miami

**J. Quin Monson**  
**Kelly D. Patterson**  
Brigham Young University

## **Abstract**

This paper presents a new method for pre-election surveys by combining the best features of other pre-election survey methods. This survey is administered online to a probability sample drawn from a voter registration list. Respondents are randomly selected from the voter file and sent an invitation letter with a link and an access code for an online poll. Two methods of probability sampling are used: a Simple Random Sample (SRS) and a Probability Proportionate to Size (PPS) sample. The PPS sample is drawn using a regression model employing variables available in the voter file to produce a probability of voting for each individual in the voter file. Preliminary results indicate that the methodology is very accurate compared to election results. A counterintuitive possibility of this research is that unlike most surveys where lower response rates increase the chance of non-response error, the low response rates in these surveys may actually decrease the chance of non-response error. Low response rates could improve the coverage of the population of likely voters because the time and effort to go online and complete the questionnaire mimics the act of voting and may require similar levels of political interest.

Presented at the annual meeting of the American Association of Public Opinion Research,  
Chicago, IL, May 13-16, 2010

## **Introduction**

Pre-election surveys are among some of the most visible work conducted by survey researchers in the eyes of the public. Because of the difficulty in accurately identifying likely voters, accurate pre-election polling is also among the most difficult tasks in survey research methodology. One expert on pre-election surveys concludes, "one of the weakest design features of most [pre-election] polls is their inability to identify correctly likely voters, especially in low turnout elections" [Crespi 1988, 178]. The process involves both science and art. It is scientific when theory-driven methods are employed, but it is also artistic in the sense that creativity and ingenuity are required in a field with an uncertain and constantly evolving methodology.

This paper presents a new method for pre-election surveys by combining several of the best features of other pre-election survey methods. This new method is administered online to a probability sample drawn from a voter registration list. Respondents are randomly selected from the voter file and sent an invitation letter via U.S. Mail with a link and an access code for an online poll. Two methods of probability sampling are used and compared: a Simple Random Sample (SRS) and a Probability Proportionate to Size (PPS) sample. The PPS sample is drawn using a regression model employing variables available in the voter file to produce a probability of voting for each individual in the voter file. We examine data from two elections that cover a variety of electoral situations including a low turnout primary and a high turnout statewide election.

Our results indicate that the methodology is very accurate compared to final election results. When the SRS and PPS sampling strategies are compared, the PPS sample is especially accurate in the low turnout elections where identifying an accurate likely electorate has historically proven to be difficult. A counterintuitive possibility of this research is that unlike most surveys where lower response rates increase the chance of non-response error, the low response rates in these surveys may actually decrease the chance of non-response error. Low response rates could improve the coverage of the population of likely voters because the time and effort to go online and complete the questionnaire closely resembles the act of voting and may require similar levels of political interest.

## Coverage Error and Models of Pre-election Polling

The difficulty with pre-election polling lies in constructing a sampling frame from a target population for an event that has not yet occurred. Put differently, identifying likely voters in a pre-election survey requires constructing a sampling frame of potential voters that accurately reflects a target population of future voters. Thus, identifying likely voters is essentially a problem of potential coverage error.

Coverage error occurs when the sampling frame does not correspond well with the target population and is defined as the "mathematical difference between a statistic calculated for the population studied and the same statistic calculated for the target population [Weisberg 2005, 205]. Efforts to assess poll accuracy, such as Martin, Traugott, and Kennedy's [2005] measure of "predictive accuracy" or the election reports of the National Council on Public Polls ([www.nccp.org](http://www.nccp.org)), are essentially efforts to measure the unit coverage error of election polls<sup>1</sup>. A false negative occurs when you screen people out who actually will vote and a false positive occurs when you include those people in your estimate who do not actually plan to vote. Substantial error is possible, particularly if turnout is likely to be very low which frequently happens in primary and off-year local elections.

Private pollsters that do the majority of preelection surveys are often reluctant to share their methods for identifying likely voters and reducing the potential coverage error because they exist in a competitive environment where being more accurate than the next pollster provides a competitive advantage [Crespi 1988, 9; Voss, Gelman, and King 1995, 118]. With comparatively less to build on from a community of scholars, preelection polling remains somewhat of an art form, requiring much analytic creativity and risk taking.

Preelection polling has typically been done using the telephone with RDD samples. Screening questions are employed to help identify likely voters. These may include eligibility to vote (registration), past voting behavior, intention to vote in the next election, interest in the election, knowledge about voting and the election, and political efficacy [Crespi 1988, 79].

---

<sup>1</sup> Determining likely voters in a telephone survey extends the coverage problem to one of within-unit coverage because the sampling unit for a typical telephone survey is the household and not the individual voter [Lavrakas 1993, 118].

The application of the screening questions to the creation of likely voter models can be subdivided into two basic categories: deterministic and probabilistic (Burden 1997). In a deterministic model, respondents are either used in survey estimates as likely voters or not. The threshold may include a single question or a combination of several questions. In the model used by Gallup, a cut point is set using contemporary estimates of turnout for the upcoming election<sup>2</sup>.

Probabilistic models involve using survey questions and perhaps other available information to assign a probability of voting or a weight to each individual survey respondent. Warren Mitofsky developed a method based on a probabilistic model for CBS News [Voss, Gelman, and King 1995; see also Traugott and Tucker 1984; Petrocik 1991; Freedman and Goldstein 1997]. Instead of discarding survey responses that do not meet a certain threshold, Mitofsky's method uses the entire sample, including respondents who are not registered to vote, and attaches weights for each respondent's probability of voting. The first step involves a series of questions related to voter registration, voting history, attention to the campaign, and mobility. For a national sample, state election laws are used to weight unregistered respondents who can no longer register to vote at zero. The remaining respondents are divided into twelve categories based upon responses to the screening questions. Using postelection survey data from recent elections, probabilities of voting are calculated for each of the twelve categories and appropriate weights are then attached.

Probabilistic weights can also be constructed by relying entirely on self-reports of likely voting (Burden 1997)<sup>3</sup>. Using self reports to calculate a probability is based on the theory that when someone reveals their intention to engage in an activity, that is the best predictor

---

<sup>2</sup>The methodology employed by Gallup has remained more or less the same since first implemented by Perry, but the items included in the scale have varied from year to year. It is essentially a deterministic model, except that partial weighting is used for a small number of respondents at the threshold. See Perry [1960 and 1979, 320-22] for a description of the development and earlier uses of the scale. See Saad [1997] for a listing of available items from 1952-1996 and Traugott and Tucker [1984] for the construction of the scale in 1980.

<sup>3</sup>For example, Burden (1997, 1167) lists the following question wording: "Just as the weatherman talks about the percent chance of rain tomorrow, what is the percent chance, from 0 to 100, that you will go and vote?" A similar approach utilizes an eleven point scale developed for models that predict consumer purchasing behavior [Juster 1960] and is based upon the idea that a respondent's own stated probability of voting can be used to weight survey responses in creating a probable electorate [Hoek and Daves 1997].

of their future behavior [Fishbein and Ajzen 1975]. Studies that examine the voter validation data available through the American National Election Studies (ANES) show that respondents who are false negatives are not a significant source of error: those who say they do not plan to vote usually keep their word. False positives are the major source of error, largely due to social desirability effects and misreporting by respondents who vote regularly but not always [Petrocik 1991; Silver, Anderson, and Abramson 1986, 615].

One of the most common questions in pre-election polling screens involves self-reporting turnout in a recent past election. This approach relies on the assumption that past behavior is a strong predictor of future behavior. As Gerber, Green, and Shachar [2003] demonstrate through a randomized field experiment, voting in one election substantially increases the probability that an individual will vote again in the next election. The idea that voting is "habit-forming" helps form the basis for a proposal by Green and Gerber [2006] to apply probabilistic methods in pre-election surveys using Registration Based Sampling (RBS). A major advantage of RBS methods compared to the more typical pre-election survey methods outlined above is that a wealth of information in voter files can be used to forecast individual-level voter turnout. In a comparison of traditional RDD methods with their method of probabilistic RBS sampling (also conducted by telephone), Green and Gerber find that their RBS sampling method performs as well or better in terms of its predictive accuracy.

A past limitation of RBS sampling methods that inhibited widespread adoption and evaluation was the unevenness of quality and availability of voter registration files. Until recently, many states did not even compile a statewide voter registration list. Indeed, anyone claiming to have a national list possessed a data file with many records that were badly out of date. The Help America Vote Act of 2002 (HAVA) required all states to have a statewide voter registration file in place by 2006. While the quality is still uneven, most states have the high quality voter files required for RBS.

In addition to a variety of sampling methods, differences in survey mode are also appearing in pre-election polls. Online surveys present possibilities for use in pre-election polling. However, estimates obtained from online methods in general can contain significant bias and present some real challenges [Couper 2000]. For example, Chang and Krosnick [2009]

report that online surveys that use nonprobability methods are not as representative demographically when compared to a benchmark like the Current Population Survey (CPS), even after weighting. They also report that online respondents tend to be stronger partisans with higher levels of political knowledge. These differences are especially acute in the nonprobability sample. Indeed a recent comprehensive report on online panels issued by the American Association for Public Opinion Research concludes, "Researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values" [AAPOR 2010, 52].

However, assuming a probability method is followed, the differences introduced by using an online survey methodology for a pre-election survey (especially in a low turnout election) where the coverage problem calls for identifying likely voters, could actually improve the survey's accuracy. In the Chang and Krosnick study, even the probability sample online survey showed higher levels of partisanship and political knowledge. However, partisanship and political knowledge are likely correlated with voter turnout, especially in a low-stimulus primary election. In other words, the online sample may not be as representative of the general population as a standard RDD sample because it includes more politically active people, but it could be more representative of the target population of likely voters because the self-selection process for voting and survey participation are similar. Likewise, if the online population is biased towards higher levels of education, income, and other socio-economic differences, these are precisely the same biases that distinguish voters from nonvoters. The sole exception is age, where older people are less likely to spend time online but are more likely to vote.

Another important source of potential error is nonresponse. Conventional wisdom suggests that high response rates lead to reduced nonresponse error and higher accuracy. However, generally speaking low responses rates have not led to significant levels of bias in surveys [Groves 2006; Keeter et al. 2006]. The same appears to be true of online surveys [Yeager et al. 2009]. Visser et al. [1996] analyze a set of election polls conducted in Ohio that includes the Columbus Dispatch Poll. The Dispatch poll is conducted entirely by mail using an RBS sampling method and has a relatively low response rate (averaging 25%). Despite its relatively low response rate, it has displayed uncanny accuracy. Visser et al. speculate that the reason for the accuracy is that the decreased response rate actually improves the survey

coverage of the population of likely voters. The time and effort required to fill out and mail back the questionnaire require similar levels of political interest normally correlated with the act of voting. When combined with a design that uses RBS and a questionnaire that also replicates the ballot (especially by not allowing voters to be undecided on vote choice), the low response rate of the Columbus Dispatch Poll appears to actually reduce coverage error and produce accurate election predictions.

But can extremely low response rates still yield accurate estimates? Peytchev, Baxter, and Carley-Baxter [2009] make an important theoretical distinction between the level of effort and the type of effort made to reduce nonresponse. Simply making additional contacts may actually serve to both increase the response rate and increase the nonresponse error at the same time because the additional effort may bring more survey respondents into the sample with the wrong characteristics. In a pre-election survey context where the coverage problem is so important, it becomes difficult to untangle the consequences of low response rates on coverage error and nonresponse error simultaneously. In light of the findings by Visser et al. with the Columbus Dispatch Poll, additional contacts with the registered voters selected for the survey to encourage them to respond would undoubtedly have increased the response rate, but may have also increased the coverage error by increasing the false positives who returned the survey and who failed to actually vote on Election Day.

The surveys described below are designed to use the advantages of RBS, online, and a self selected likely voter pool to increase the accuracy of pre-election poll estimates at a fraction of the cost of traditional pre-election polling methodologies.

### **Theorizing A New Method**

By identifying the best features of each of the pre-election polling methods discussed above, we can begin to develop a new method that incorporates these features. Random digit dialing has a long history and is still frequently used because it uses a sampling method that is truly random and thus does not introduce bias into the estimates [Yeager et al. 2009]. However, its weakness for pre-election polling is that the target population of likely voters is substantially different than the RDD sampling frame of all phone numbers. By incorporating

potential respondents' demographic information, registration based sampling allows us to create a more accurate sampling frame, thus helping to better identify likely voters before ever speaking with a respondent in a survey [Green and Gerber 2006]. Using mailed surveys is less expensive than conducting a survey via telephone or in person and the coverage of the target population appears to benefit from low response rates [Visser et al. 1996]. Finally, online surveys, while often drawing an unrepresentative sample, are extremely cost effective thereby making it possible to conduct more of them. By using a piece of each of these methods, we create an inexpensive way to conduct accurate pre-election polls.

The first key to such accuracy is generating a sample that accurately reflects the likely electorate. Using a digital voter file provides us an excellent starting place for identifying those who are most likely to vote. However, as we will discuss later, it is important to not select only those voters who are the most likely to vote as determined by a particular statistical model. Consider a simple example. Suppose that voters in a state are assigned a predicted probability of voting based on demographics such as age, gender, voting history and registration status. Now consider those who are least likely to vote. While each individual is expected not to vote, in the aggregate, many of these people will show up on Election Day. If enough of these people do vote, and their candidate or policy preferences are systematically different from those who have a higher likelihood of voting, then any pre-election survey should sample several of these people. Issues, candidates, weather, and the voters themselves change from election to election, and as a result basing future turnout on previous voting should not be the only way to develop an accurate sample. In general, every election will contain a mixture of high and low probability voters, and an accurate pre-election poll will account for this mixture.

Furthermore, while statistical models can predict turnout quite accurately, we also want to include a respondent's self report of how likely they are to vote to act as an additional screen when identifying who to include in the sample. We hypothesize that voters who complete the survey are more interested in politics in general and more likely to participate in the election than those who fail to complete the survey. By mailing the survey to voters and asking them to then complete a survey online, we effectively gauge the voter's level of interest and willingness to make an effort to participate. Thus, the combination of a traditional mailed survey with an online survey creates more burdens for the respondent

than traditional phone or in-person interviewing methods. We hypothesize that this effort closely mimics the effort needed to vote and will act as a likely-voter screen. As a result, we should not be concerned with a low response rate. In fact, we would expect to have a low response rate since we expect and want many of those we sampled (particularly those who are uninterested in the upcoming election) to not complete the survey.

## **Data and Methods**

In order to sample voters, we created a model that assigned a predicted probability of voting to each potential voter in the state. We created this model using the publicly available Utah voter file. This file contains information regarding registration status, partisan affiliation, age, address, and previous voting history for every registered voter in the state. The file is updated every election and includes whether or not each person in the file turned out to vote. Additionally, any changes in the person's registration status (i.e. first time registration date, change of party affiliation, legislative districts) are updated as needed. Using this information, we create likely voter models for the 2008 primary election and the 2008 general election using logit regression analysis.

Our predictive model relies on the most recent election that resembles the upcoming election in which we are interested. We construct our model using the appropriate historical election, then apply the model to the current voter information to project the probability of turnout for each individual voter. Our dependent variable in each model is a dichotomous measure of whether or not the citizen voted in the most recent election that resembles the election we are surveying. In the 2008 3<sup>rd</sup> Congressional District Republican primary election we use the 2006 GOP primary in the 3<sup>rd</sup> district as the dependent variable. The 2006 GOP primary is the most recent Congressional primary election and thus best resembles the 2008 primary. Turnout in 2006 was 11.6% in the 3<sup>rd</sup> Congressional District. To verify our conjecture of the similarity of the two elections, we calculated the voting rate in the 3<sup>rd</sup> District in 2008 following the election from official state results. 9.6% of registered voters turned out in the 3<sup>rd</sup> District, which is similar to the 2006 rate. In the 2008 general election survey we used the 2004 general election as the dependent variable. We chose to go back to 2004 instead of using the 2006 general election because there are clear differences in

turnout rates and characteristics of the electorate when comparing midterm and presidential general elections. Thus, we felt 2004 was a more appropriate choice. In 2004, 73.3% of registered voters turned out, and in 2008 turnout was 67.8%. 2008 turnout was slightly lower. However, the rate in the 2006 general was much lower at 44.7%, confirming our decision to use 2004 instead of 2006.

The independent variables in both models are taken from the voter file, including previous voting record, age, time from most recent change in registration status. When developing the model using the previous election as the dependent variable, we adjust the independent variables for that election (e.g. calculating age at the time of the election used as the dependent variable). Table 1 displays the coefficients and standard errors for each variable in the model. See Appendix A for a description of the variables used.<sup>4</sup>

#### Primary Election Model

$$Y_i \in \{\sim Vote_{2006 \text{ Primary } 3rd \text{ CD}} = 0, Vote_{2006 \text{ Primary } 3rd \text{ CD}} = 1\}, Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\begin{aligned} \text{logit}(\pi_i) = & \beta_0 + \beta_1 Prim_i + \beta_2 Gen_i + \beta_3 Age_i + \beta_4 Republican_i \\ & + \beta_5 Registration \text{ Quintiles}_i + \beta_6 Age * Reg \text{ Quintiles}_i \end{aligned}$$

#### General Election Model

$$Y_i \in \{\sim Vote_{2004 \text{ General}} = 0, Vote_{2004 \text{ General}} = 1\}, Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\begin{aligned} \text{logit}(\pi_i) = & \beta_0 + \beta_1 Prim_i + \beta_2 Gen_i + \beta_3 PresPrim08 + \beta_4 Age_i + \beta_5 Age^2 \\ & + \beta_6 Republican_i + \beta_7 Democrat_i + \beta_8 MinorParty_i \\ & + \beta_9 Registration \text{ Quintiles}_i + \beta_{10} Age \text{ Quintiles}_i + \beta_{11} Age \text{ Quintile} \\ & * Reg \text{ Quintile}_i \end{aligned}$$

---

<sup>4</sup> Malchow (2008), Chapter 11 describes a similar process of modeling turnout used by political consultants.

Table 1 About Here

### Models of Predicted Probability of Voting

The distributions of individual voter turnout probabilities are shown below in Figure 1. As we would expect, the distribution of probabilities in the primary election model is skewed to low values. In the general election model where turnout is higher than the primary election, the distribution of predicted probabilities is much more uniform. We will not spend time discussing the coefficients and their respective statistical significance as any regression model with as many observations as we have is bound to produce extremely small standard errors. However, we should note, as shown in the regression table that the sign of each coefficient is in the expected direction. Previous primary election voting has the largest impact on the predicted probability of voting in the primary election with previous general election voting exhibiting the same properties in the general election model. Partisan affiliation and age also increase the likelihood of voting as expected.

Figure 1 About Here

### Distribution of Predicted Probability for Primary and General Elections

After generating the predicted probabilities of voting for each person in the voter file, we then sampled 10,000 people to invite to participate in the surveys. We divided the sample into two parts and then used two different methods of sampling. In the primary election we sampled 8000 cases using a simple random sample (SRS) and 2000 cases using a probability proportionate to size (PPS) sample. The PPS sample uses the predicted probability of voting as a weight when sampling. Thus, those with higher probabilities of voting are more likely to be included in the sample.<sup>5</sup> We will discuss the implications of this sampling strategy in the results section of the paper. In the general election we drew a sample of 10,000 with 5,000 being selected using SRS and 5,000 selected using PPS sampling<sup>6</sup>.

---

<sup>5</sup> We used the Stata command "samplepps" to draw the PPS sample.

<sup>6</sup> Because we did not want to sample two people in the same household, we sampled more than 10,000 people each time and eliminated cases that contained the same address. We eliminated duplicates by sorting duplicate addresses alphabetically by first name and took the

## Survey Recruitment

Once we had selected our sample we sent letters inviting those sampled to participate in the survey by accessing a website. Each letter contained a unique ID code that we had assigned to each respondent that they were required to input in order to take the survey. The survey asked various questions regarding likelihood of voting, candidate preference for several offices, and demographic questions. See Appendix B for a copy of the letter that was sent in both the primary and general elections. A copy of both survey instruments is available upon request. In the primary election the letters were sent 10 days prior to the upcoming election. In the general they were sent 7 days in advance of the election. Response to the survey was closed on the morning of Election Day. In the primary election 653 people began the survey and in the general election 641 people started. Thus our response rates using the AAPOR Response Rate 1 formula were 6.2% and 6.4% respectively. When we remove those letters that were undeliverable (AAPOR Response Rate 2), our response rates increase to 6.5% and 6.7%.

## Verification

After the election, we verified that our model based on past voting behavior and demographics accurately predicted whether or not the person voted in the 2008 primary or 2008 general election. Figure 2 below displays a bar graph of the mean predicted probability of voting compared to the voting rate of those same people who actually voted. All of the respondents in the voter file for whom a predicted probability was calculated are included in the comparison. We have grouped the respondents into 10 percentage point bands based on their predicted probability of voting. This allows us to make a much more useful comparison of the model's accuracy than a traditional sensitivity/specificity analysis would allow. For example, those people whose predicted probability of voting was between 0 and

---

name that was closest to the end of the alphabet. We are confident that names are randomly distributed in the alphabet and by selected the name closest to Z we are not biasing the survey. After eliminating duplicate cases, we were left with slightly more than 10,000 cases. Using a random number generator based on a uniform distribution we assigned all cases a numeric value and then sorted the cases based on that value. We then took the first 10,000 cases.

.1 have a mean predicted probability of .0258. This is displayed in the left most black column. Next to this column is the actual voting rate of those same respondents in the 2008 primary (.0467). This is displayed in the left most white column. If the model is accurate, we should expect to see only a small difference between these two columns for each of the 10 bins in each graph. We see that in both elections, the predicted and actual bins diverge, sometimes quite dramatically. However, it is important to note that in most of these cases, these bins are sparsely populated. The difference is minimal near the actual turnout rate in the 2008 primary election, 7.99%. Therefore, a weighted average difference of turnout rates between the predicted probability of voting and the actual turnout is more appropriate. Dividing the voters into 100 different groups based on their predicted probability of voting ([0-.01), [.01-.02), ..., [.99-1.0)), the average weighted deviation (weighted by the number of voters in each bin) between the mean predicted probability of each bin and the actual turnout rate among these bins is -.02, indicating that on average we only slightly over predicted turnout by roughly two percentage points. This average difference is nearly identical to the difference in turnout between the 2006 primary (11.6%) and the 2008 primary (9.6%).

Figure 2 About Here

Mean Predicted Probability of Voting Compared to Actual Turnout Rate in 10 Percentage Point Bins

Figure 3 displays the ROC curves comparing the predictions of the model to the actual turnout in the 2008 primary election in the 3<sup>rd</sup> Congressional district. A ROC curve plots the sensitivity (i.e. the rate of true positives) vs. specificity (i.e. the rate of false positives)<sup>7</sup> for model of binary outcomes as model's predictions are tested while varying the cut-point from 0 to 1. The area under the curve is a measure of the accuracy of the model. A model that performed exactly randomly (e.g. flipping a coin) would follow the diagonal line from the bottom left to the top right corner (known as the "line of no discrimination") and have an area of 0.5. A model that was less accurate at predicting correct outcomes than random would have an area less than 0.5. A perfect model with no false negatives and no false

---

<sup>7</sup> The ROC curve actually plots 1-specificity.

positives would trace the y-axis and the x-axis and have an area of 1. Therefore, ROC curve estimates close to 1 indicate a model that accurately predicts individual voter turnout (a dichotomous outcome) for all potential predicted probabilities of voting. For the primary election, the area under the ROC curve was 0.899. Moreover, the curve itself appears to be quite smooth, indicating that the model performs well across the entire range of predicted probabilities of voting.

Figure 3 About Here

### ROC Curves

The same analysis of the model for the 2008 general (Figure 2) shows the model is again quite accurate for voters around the 2008 General election turnout of 67.8% . However, the model dramatically under-predicts turnout among those that are assigned low probabilities of turning out in this high turnout election. The largest difference between predicted turnout and actual turnout rates is among those who were least likely to vote, however; the model prediction places only 34 out of more than 1.4 million people into this bin. Thus, while the difference is large, when considering its overall impact on a weighted average of the deviations (as done in the primary election model) will be small. Using the same method as in the primary election, we divided the voters into 100 bins according to their predicted probability of voting. The average weighted deviation between the mean predicted probability of each bin and the actual turnout rate among these same bins is .03. As above, recall Figure 1b where these low turnout bands are very sparsely populated for the 2008 General election, these are the bins in which the deviation is largest. On the other hand, the bins in which most of the data are located are where the model is most accurate in its predictions.

The ROC curve comparing the predicted probabilities of turnout in the 2008 general election to the actual voter turnout (Figure 3) also shows a strong performance by the model. The area under the ROC curve is 0.867 and the curve is smooth.

Having established that the model was quite accurate at predicting people's likelihood of voting, we now turn to an analysis of the distributions of predicted probabilities among the

subset of the population sampled in either the PPS or the simple random sample. Figure 3 below displays the predicted probability of voting for those sampled in the PPS and simple random samples for the primary election and the general election. As we would expect, the simple random samples resemble the entire population distribution in both cases. The PPS distributions are weighted more heavily towards those who are more likely to vote. The PPS distributions are different between elections, which is a function of the differences between the underlying distributions from which the PPS samples were drawn.

Figure 4 About Here

#### Distribution of Predicted Probabilities among PPS and SRS for Both Elections

In each election, the distribution of the SRS sample across probability of turning out mirrors the population because the sampling is unrelated to the probability of turning out. However, the PPS samples are weighted by the probability of turning out. In the general election, the differences are quite small because most voters have a high probability of turning out. For the general election, the main difference between Figure 4c (SRS) and 4d (PPS) is that voters on the left side of the graphs with a low probability of turning out are less likely to be included in the PPS sample (Fig 4d). The effect of weighting the PPS sample by the probability of turning out has a more dramatic effect in low turnout elections like the primary. Figure 1a (SRS) shows an extreme skew to the left side of the graph. However, when weighted by the probability of voting, the distribution of the PPS sample is much more uniform. In the PPS sampling process (Figure 1b), the large share of voters on the left is discounted by their low probability of voting, while the small share of voters on the right is inflated because of their high probability of voting. The more uniform distribution seen in Figure 1b is a result of the balancing of the probability based on density of voters and the intensity from the predicted likelihood of voting.

The SRS sample and PPS sample distributions in the primary and general elections are specific to these election contexts. Nevertheless, the differences between the samples in the primary and general elections illuminate why pre-election polling in low turnout elections is so difficult and costly. In a general election, the simple random sample roughly approximates the likely electorate. Techniques such as vote likelihood screens only need to do a small amount of work to refine the sample of responses to be representative of the

likely electorate. In a primary election, a simple random sample is very different from the distribution of likely voters. Likely voter screens and other techniques for shaping the responses to be representative of the likely electorate must do an enormous amount of work. At best, the use of these techniques is costly, in the form of longer screening batteries, discarded responses from unlikely voters, etc. At worst, these techniques are not robust or accurate enough to correctly identify a universe that is representative of the likely electorate in low turnout elections like the primary.

### **Response Rates between Sampling Methods**

While our response rates were 6.4% and 6.7%<sup>8</sup> for the primary and general elections respectively, the response rate within each survey was dramatically different based on the type of sampling that was used to select the potential respondent. As the PPS sample was drawn with unequal probability favoring those who we had predicted would be more likely to vote, we expected that those same people would also be more likely to complete the survey. Both voting and participation in a survey are correlated with levels of interest in, engagement with, and knowledge about the election. This hypothesis proved true in both cases: the completion rate in the primary survey was 6.1% for those sampled using the SRS method and 12.5% for those who were sampled using PPS sampling; in the general election survey the response rate for those sampled by the SRS method was 6.6% while the PPS sampling response rate was 8.5%.

Table 2 shows completion of the survey is significantly correlated with higher predicted probabilities of voting. The first regression shows the results of a logistic model of the 10,000 people sampled for the general election survey with the dependent variable as response to the survey. The second regression shows the same analysis of the 10,000 people in the primary election survey sample.

---

<sup>8</sup> We will use AAPOR response rate 2 formula for all response rates for the remainder of the paper.

## Table 2 About Here

### Correlation between Probability of Voting and Survey Completion

#### **Accuracy of the Polls**

Since the goal of the survey was pre-election forecasting, the most important characteristics of the survey is how accurately its results reflect the actual outcome of the elections. To measure the accuracy, we compare forecasts from our survey to the actual election results.

We assess the accuracy of our forecasts using responses from voters who indicated they would vote on Election Day or had already voted via early voting or absentee/mail ballot. In each survey, respondents were first asked whether or not they had already voted via absentee or early voting. If they had, they were directed to a question that asked who they voted for. This question included an option for "don't remember/did not vote in this election." Those respondents who told us they had not yet voted were asked to indicate on a 10 point scale how likely it was that they would vote in the upcoming election. Those who put "0, no chance I will vote" were directed to demographic questions at the end of the survey. Those who put anything besides "0" were then asked who they planned to vote for in several races. We do not include responses from those who already voted but indicated that they either could not remember who they voted for or did not vote in that specific race.

In the primary election we asked participants to indicate how they voted/would vote in the Republican congressional primary. In the general election, we asked the participants to indicate how they had voted/would vote in six different races: president, governor, U.S. Representative, attorney general, Utah House and Utah Senate.

The following tabulations do not sum to 100 because we do not include respondents who indicated a minor candidate. In the primary, minor candidates were defined as anyone other than Congressman Chris Cannon and challenger Jason Chaffetz. In the general election, minor candidates were defined as anyone except the Republican nominee and the Democratic nominee in each race.

The column "Predictive Accuracy" is a measure of survey accuracy proposed by Traugott, Martin, and Kennedy (2005)<sup>9</sup>. The measure is calculated using a logged odds ratio between the poll forecast and the actual results of the election. Negative numbers indicate Democratic bias while positive numbers indicate Republican bias. If the confidence interval includes 0 then the poll is not statistically significantly biased. Table 3 below displays the results for both the primary and general election surveys. Table 4 splits the responses by sampling method and compares the predictive accuracy measure for each race. Note that in Table 3 every prediction's confidence interval on the predictive accuracy measure includes zero, indicating that each of our predictions was quite accurate given the bounds of sampling variation. Unfortunately, there was not a race that was particularly close such that we could truly see how well our predictions held up in a competitive race in which a pre-election poll would shed light on the likely outcome. However, while the margin of the Chaffetz/Cannon race is quite large, it was equally unexpected. Much of the media coverage leading up to the primary election suggested that the two candidates were nearly tied. In fact, a public telephone poll using an RDD sample conducted by a major Utah newspaper at the same time as our survey suggested that Cannon was in the lead, although the margin of error was too large to be definitive. Given these prior beliefs, the accuracy of our poll is quite encouraging.

Tables 3 and 4 About Here

### Accuracy of Poll in Primary and General Elections

#### Comparison of Accuracy by Sampling Method

Table 4 allows us to compare the predicted accuracy measure between the two different sampling methods. The point estimates of the PPS sample are more accurate than the SRS sample in all but two of the races. Despite larger standard errors because of the smaller sample, the pattern of greater accuracy from the PPS sample suggests that the PPS sample gives us a more accurate sampling frame for the likely electorate. It appears to be working as

---

<sup>9</sup> Predictive accuracy,  $A$ , is a measure developed by Martin, Traugott, and Kennedy (2005), and is described in full there. The formula for computing  $A$  is  $A_{ijk} = \log \left[ \frac{(r_{ijk})}{(d_{ijk})} \right] / \left[ \frac{(R_{ijk})}{(D_{ijk})} \right]$  for poll  $i$  in race  $j$  in state  $k$ .  $(r,d)$  represent predicted vote shares for the two major parties and  $(R,D)$  are actual vote shares.

an effective screen for likely voters by relying on past voting behavior to predict engagement with the current election rather than relying on responses to vote intention questions.

## **Conclusion**

Overall, these results suggest that sampling based on a respondent's statistical predicted probability of voting combined with self reported interest in the election (as measured through actually going online and taking the survey) holds promise for providing more accurate pre-election forecasts. As discussed earlier, the survey response rate was higher in both cases among those that were selected using PPS sampling. This suggests that PPS sampling is doing the work expected of a likely voter screen would work in a traditional telephone interview. However, the result comes at a significantly lower cost and are based on verifiable past behavior rather than responses which are notoriously unreliable.

Moreover, the results are accurate across a variety of races and different types of elections which suggests that this method hold promise for expansion into other areas of polling that are difficult due to low response rates.

We note that the PPS sample was not more accurate than the simple random sample in every race (see Table 4). Its advantage in the general election disappears in less contested races (with less campaign information, name recognition, etc). The low information and less contested races (Attorney General, congressional districts 1 & 2) have the largest inaccuracies and no advantage between the PPS and simple random samples. As an avenue for future research, we would like to investigate why this may be the case. One hypothesis we have is that sampling based on previous voting behavior may be superior in cases where likely voters are also highly informed about the candidates, but if habitual voters are as uninformed or uninterested as unlikely voters, the difference between sampling methods may disappear.

## **Appendix A - Description of Variables Used in Likely Voter Models**

### **[Gen04]**

This is our dependent variable for the general election model. This variable is a dummy variable indicating whether or not the person voted in the 2004 general election. 73% of registered voters turned out in this election. We used this variable instead of the 2006 general election, which was more recent, because we felt that the different dynamics between presidential and midterm elections were great enough to warrant using the most recent presidential election rather than the most recent general election.

### **[Prim06]**

This is our dependent variable for the primary election model. This variable is a dummy variable indicating whether or not the person voted in the 2006 Republican primary election in the 3<sup>rd</sup> Congressional District. 9.5% of registered voters in the 3<sup>rd</sup> Congressional District turned out in this election.

### **[Prim]**

This variable is a count of the number of times the person has voted in the four most recent primary elections (2000, 2002, 2004, 2006). We did not include the 2008 primary in the model. This was an oversight on our part. When we constructed the model, we did not have a voter file that included the 2008 primary. When we did receive an updated file, we forgot to update the model to include the 2008 primary. Thus, this variable ranges from 0 to 4 in both the primary and general election models.

### **[Gen]**

This variable is a count of the number of times the person has voted in the five (Primary model) or three (General model) most recent general elections (1998, 2000, 2002, 2004, 2006). In the general election model we did not include the 2004 general election because we used that election as our dependent variable. To avoid having part of an independent variable that is a function of the dependent variable we excluded 2004 from this covariate.

This variable ranges from 0 to 5 in the primary election model and from 0 to 3 in the general election model. We chose not to include 1998 in the general election model after deciding that after 10 years, many people in the current voting file were not voting in 1998.

### **[PPrim08]**

This variable is a dummy variable indicating whether or not the person voted in the 2008 presidential primary. We did not include this in the prim index because we felt that the dynamics of the presidential primary (i.e. the Romney Campaign's popularity in Utah) were different enough to warrant its inclusion as a separate variable. Turnout in this race was 32.4%.

### **[Age]**

This is a continuous variable that indicates the person's age at the time the voter file was downloaded from the state's server. We created the variable by subtracting the election date from the person's birth date as recorded in the voter file and dividing by 365. This method does not take into account leap years. However, given a 100 year old voter, the most that our method could be off because of leap year would be 25 days (100/4 years per leap year).

### **[Age2]**

This variable allows age to affect the dependent variable in a nonlinear manner. We hypothesized that increasing age would also increase the likelihood of voting, but only to a certain point, whereupon increasing age may actually decrease the probability of voting due to health issues that come with old age. By squaring the age variable, the age and age<sup>2</sup> variables can affect the dependent variable in a parabolic way.

### **[Rep]**

This variable is a dummy variable that is coded 1 if the person is a registered Republican. Approximately 40% of our observations are included in this category.

### **[Dem]**

This variable is a dummy variable that is coded 1 if the person is a registered Democrat. Approximately 8% of our observation are included in this category.

### **[Oth]**

This variable is a dummy variable that is coded 1 if the person is registered with some other party besides the Republican or Democratic Party. This accounts for 2% of our observations. By including rep, dem, and oth in the model, our omitted comparison group is voters who are registered as unaffiliated voters. There is an Independent Party in Utah, so it is important to clarify between Independent and unaffiliated. A common misconception is that of the dominance of the Republican Party in Utah. In fact, the majority of voters are unaffiliated, 51% of our observations are included in this category.

### **[Reg5]**

The voter file contains two different registration dates for each person. The first is an original registration date. This variable was incorrectly coded in the voter file so we did not use it in our likely voter model. The second variable, which we did use, indicates the date that the voter most recently changed their registration status. This could be because they moved to a different part of Utah or changed their party affiliation. This happens often, however, we suspect that this occurred with unusual frequency as many Democrats and Unaffiliated voters switched their registration to be registered Republican in order to vote for Mitt Romney in the February Presidential Primary. The variable is called reg5 because we broke the original variable into quintiles. In quintiles the variable no longer has values that are dates: the observations take on the values of 1 through 5 based on the quintile they belonged to in the original registration variable. The variable is now ordinal rather than interval/ratio.

### **[Age5]**

This variable is very similar to the Reg5 variable above. We took the age variable and created an ordinal variable, Age5, that indicated the quintile that the voters age would fall into. The reason for creating this variable was to interact age with registration date. Because registration date indicates the most recent change in registration rather than the original registration date, we hypothesized that a recent change in registration date could indicate different things about the voter's likelihood of voting. A middle-aged, very active voter who is prompt in updating his/her registration should have a high probability of voting in the

election. However, a young voter who has recently registered for the first time should have a lower probability of voting than the person previously described. However, the value of the registration variable would be the same. They both changed their registration status recently. By interacting registration quintiles with age quintiles, we can allow the registration date to have more flexibility in the model, making it possible to distinguish between the old, active voter and the young, newly registered voter. This increases the accuracy of the model (The more flexibility, the more accuracy).

## Appendix B – Invitation Letters Sent to Sample



### CENTER FOR THE STUDY OF ELECTIONS AND DEMOCRACY

*Brigham Young University*

June 13, 2008

As the June 24, 2008 U.S. House of Representatives primary election approaches I invite you to participate in a special edition of the *Utah Voter Poll*, conducted by the Center for the Study of Elections and Democracy at Brigham Young University.

You were selected at random from a list of all registered voters in the Utah 3rd Congressional District. Your participation is very important to us and will help make the survey accurate. This online survey takes less than 10 minutes to complete and your answers are completely confidential.

To begin the survey:

- Enter the following URL into any web browser: <http://utahvoterpoll.org>
- Click on “CLICK HERE TO BEGIN SURVEY”
- Enter your five-digit “Access Code” located just below the return address on the outside of the envelope we sent with this letter.

To ensure that only voters who have been invited can participate in the survey we have provided a unique access code. If you cannot find or cannot read the access code on the outside of the envelope, please email [utahvoterpoll@byu.edu](mailto:utahvoterpoll@byu.edu) or call 801-422-5237 so that we can help. For other questions, responses to a list of “Frequently Asked Questions” are located on our web site at <http://utahvoterpoll.org/faq.htm>

The survey is available now. Please begin and finish the survey before it closes at midnight on Monday, June 23<sup>rd</sup>. Thank you very much for helping with this important study.

Sincerely,



J. Quin Monson, PhD  
Center for the Study of Elections and Democracy  
Brigham Young University

P.S. Your participation is very important to us. Please take the survey before midnight on June 23<sup>rd</sup>.



CENTER FOR THE STUDY OF  
ELECTIONS AND DEMOCRACY

*Brigham Young University*

October 27, 2008

As the November 4, 2008 U.S. general election approaches I invite you to participate in a special edition of the *Utah Voter Poll*, conducted by the Center for the Study of Elections and Democracy at Brigham Young University.

You were selected at random from a list of all registered voters in the state of Utah. Your participation is very important to us and will help make the survey accurate. This online survey takes less than 10 minutes to complete and your answers are completely confidential.

To begin the survey:

- Enter the following URL into any web browser: <http://utahvoterpoll.org>
- Click on "CLICK HERE TO BEGIN SURVEY"
- Enter your five-digit "Access Code" in the space provided. If required, enter your congressional district number as well. Both are printed just below the return address on the outside of the envelope sent with this letter.

To ensure that only voters who have been invited can participate in the survey we have provided a unique access code. If you cannot find or cannot read the access code on the outside of the envelope, please email [utahvoterpoll@byu.edu](mailto:utahvoterpoll@byu.edu) or call 801-422-5237 so that we can help. For other questions, responses to a list of "Frequently Asked Questions" are located on our web site at <http://utahvoterpoll.org/faq.htm>

The survey is available now. Please begin and finish the survey before it closes at midnight on Monday, November 3<sup>rd</sup>. Thank you very much for helping with this important study.

Sincerely,

J. Quin Monson, PhD  
Center for the Study of Elections and Democracy  
Brigham Young University

P.S. Your participation is very important to us. Please take the survey before midnight on November 3<sup>rd</sup>.

## References

- American Association of Public Opinion Research. 2010. "AAPOR Report on Online Panels." Prepared for the AAPOR Executive Council by a Task Force operating under the auspices of the AAPOR Standards Committee.  
[http://www.aapor.org/AM/Template.cfm?Section=AAPOR Committee and Task Force Reports&Template=/CM/ContentDisplay.cfm&ContentID=2223](http://www.aapor.org/AM/Template.cfm?Section=AAPOR_Committee_and_Task_Force_Reports&Template=/CM/ContentDisplay.cfm&ContentID=2223), Accessed, May 6, 2010.
- Burden, Barry. 1997. "Deterministic and Probabilistic Voting Models." *American Journal of Political Science* 41:1150-69.
- Chang, LinChiat and Jon Krosnick. 2009. "National Surveys via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73:641-678.
- Couper, Mick P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64:464-94.
- Crespi, Irving. 1988. *Pre-election Polling: Sources of Accuracy and Error*. New York: Russell Sage Foundation.
- Fishbein, Martin and Icek Ajzen. 1975. *Belief, Attitude, Intention, and Behavior*. Reading, MA: Addison-Wesley Publishing.
- Freedman, Paul and Ken Goldstein. 1997. "Building a Probable Electorate from Preelection Polls: A Two Stage Approach." *Public Opinion Quarterly* 60:574-87.
- Gerber, Alan S., Donald P. Green, and Ron Shachar. 2003. "Voting May be Habit-Forming: Evidence from a Randomized Field Experiment." *American Journal of Political Science* 47:540-50.
- Green, Donald and Alan Gerber. 2003. "Using Registration-Based Sampling to Improve Pre-Election Polling." *Paper presented at American Association of Public Opinion Research Annual Conference*. Nashville, TN.
- Gerber, Alan S. and Donald P. Green. 2006. "Can Registration-Based Sampling Improve the Accuracy of Midterm Election Forecasts?" *Public Opinion Quarterly* 70:197-223.
- Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys" *Public Opinion Quarterly* 70:646-75.
- Hoek, Janet and Robert P. Daves. 1997. "Turnout Prediction: A Comparison of Methodologies." Presented at the annual meeting of the American Association for Public Opinion Research, Norfolk, VA.

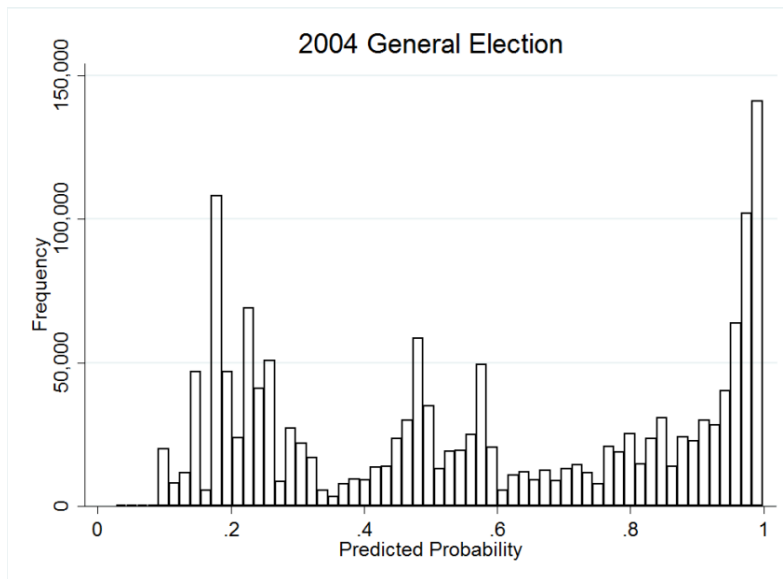
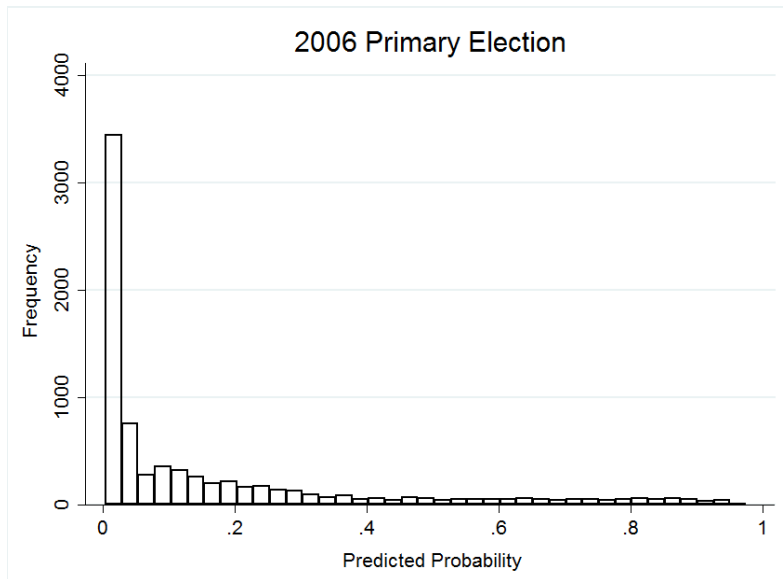
- Juster, F. Thomas. 1960. "Prediction and Consumer Buying Intentions." *American Economic Review* 50 (2):604-22.
- Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best and Peyton Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70:759-779.
- Lavrakas, Paul J. 1993. *Telephone Survey Methods: Sampling, Selection, and Supervision*. 2nd ed. Newbury Park, CA: Sage Publications.
- Malchow, Hal. 2008. *Political Targeting*. Predicted Lists Publishing.
- Martin, Elizabeth, Michael Traugott, and Courtney Kennedy. 2005. "A Review and Proposal for a New Measure of Poll Accuracy." *Public Opinion Quarterly*, 69, no. 3: 342-369.
- Perry, Paul. 1960. "Election Survey Procedures of the Gallup Poll." *Public Opinion Quarterly* 24:531-42.
- Perry, Paul. 1979. "Certain Problems in Election Survey Methodology." *Public Opinion Quarterly*. 43:312-25.
- Petrocik, John R. 1991. "An Algorithm for Estimating Turnout as a Guide to Predicting Elections." *Public Opinion Quarterly* 55:643-47.
- Peytchev, Andy, Rodney K. Baxter, and Lisa R. Carley-Baxter. 2009. "Not All Survey Effort is Equal: Reduction of Nonresponse Bias and Nonresponse Error." *Public Opinion Quarterly* 73: 785-806.
- Saad, Lydia K. 1997. "An Historical Analysis: Presidential Candidate Preferences According to Likelihood to Vote." Presented at the annual meeting of the American Association for Public Opinion Research, Norfolk, VA.
- Schwartz, Doug and Clay Richards. 2003. "The Big East Two: A Comparison of RBS and RDD Polls in the 2002 Elections in New York and Pennsylvania." *Paper presented at American Association of Public Opinion Research Annual Conference*. Philadelphia, PA.
- Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80:613-24.
- Traugott, Michael W. and Clyde Tucker. 1984. "Strategies for Predicting Whether a Citizen Will Vote and Estimation of Electoral Outcomes." *Public Opinion Quarterly* 48:330:43.

- Traugott, Michael. 2005. "The Accuracy of the National Preelection Polls in the 2004 Presidential Election." *Public Opinion Quarterly*, 69, no. 5: 642-654.
- Visser, Penny S., Jon A. Krosnick, Jesse Marquette, and Michael Curtin. 1996. "Mail Surveys for Election Forecasting? An Evaluation of the *Columbus Dispatch* Poll" *Public Opinion Quarterly* 60:181-227.
- Voss, D. Stephen, Andrew Gelman, and Gary King. 1995. "Preelection Survey Methodology: Details from Eight Polling Organizations, 1988 and 1992." *Public Opinion Quarterly* 59:98-132.
- Weisberg, Herbert F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press.
- Yeager, David S., Jon Krosnick, LinChiat Chang, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2009. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples" Stanford University.  
<http://comm.stanford.edu/faculty/krosnick/Mode%2004.pdf>, Accessed, May 6, 2010.

## Figures and Tables

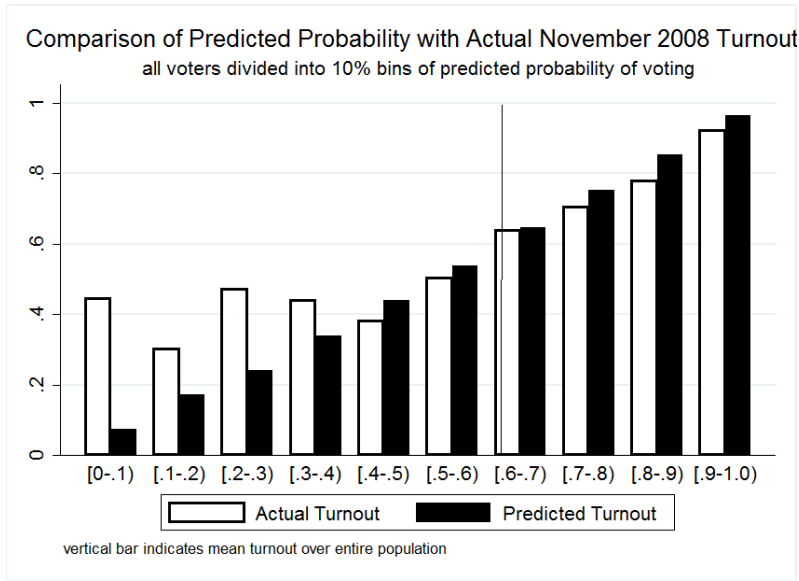
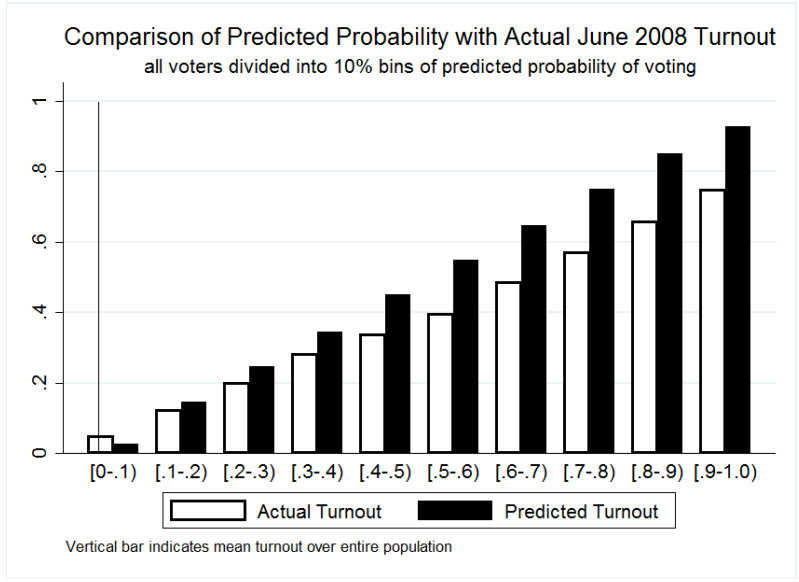
### Figures 1a and 1b:

#### Distribution of Predicted Probabilities for Primary and General Elections

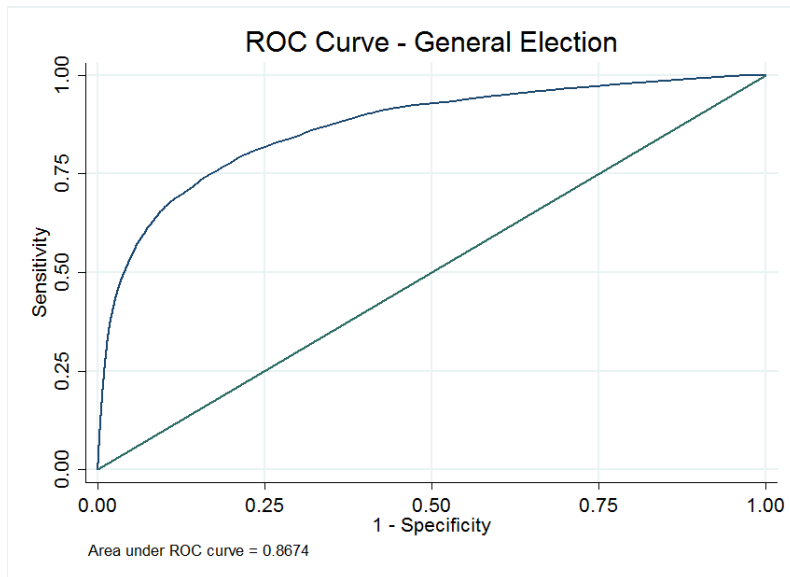
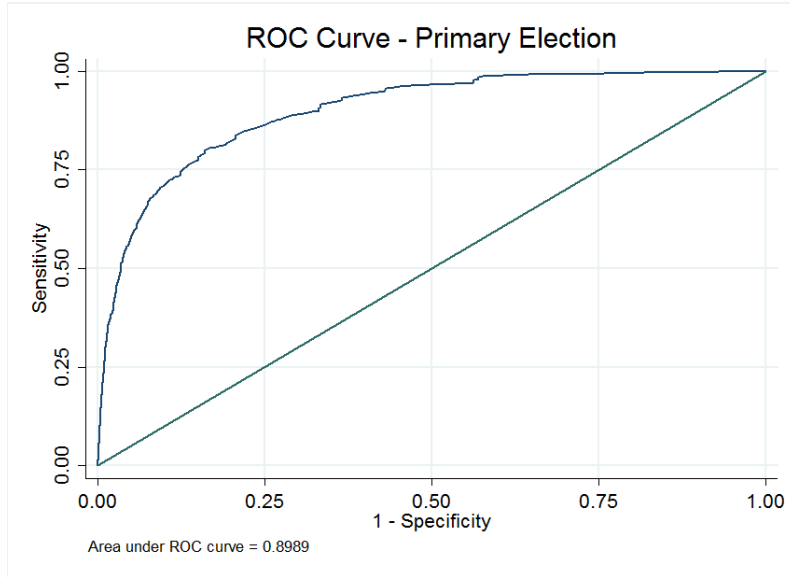


Figures 2a and 2b:

Mean Predicted Probability of Voting Compared to Actual Turnout Rate in 10 Percentage Point Bins

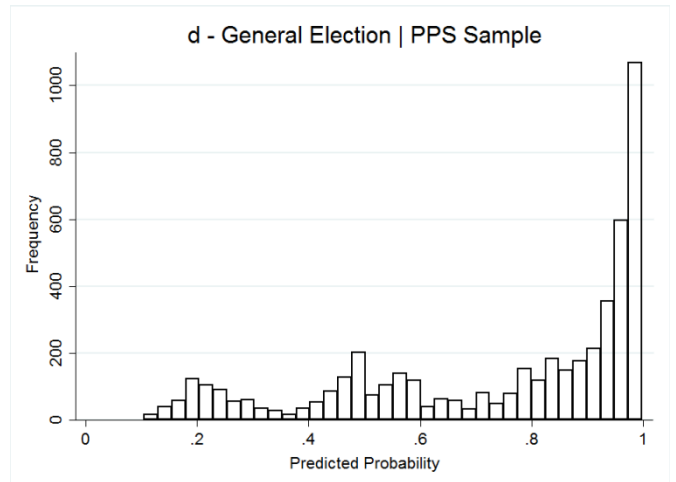
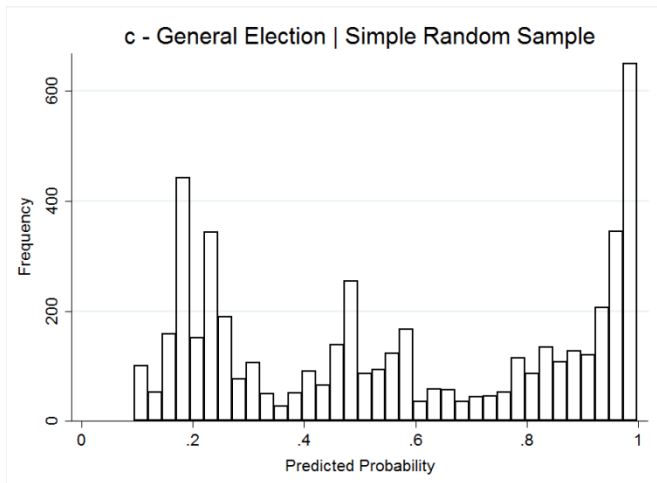
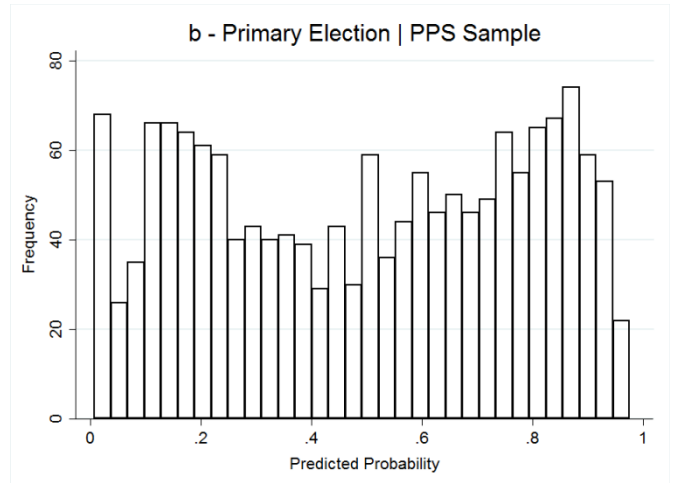
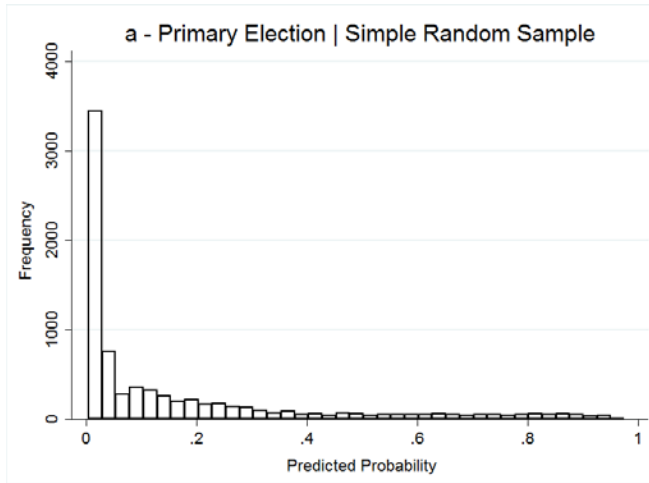


Figures 3a and 3b – ROC Curves for Each Model



**Figures 4a-4d:**

**Distribution of Predicted Probabilities for Primary and General Elections by Sampling Method**



Tables 1a and 1b:

Regression Results of Likely Voter Model with 2006 Primary and 2008 General Elections as Dependent Variables

Logistic Regression of Probability of Voting in 2006 Primary Election

Dependent Variable: Voted in 2006 Primary Election

	Coefficient	St. Error	P-Value
Prim	0.608	0.006	***
Gen	0.286	0.004	***
Re-expressed Age	0.013	0.000	***
Republican	0.728	0.038	***
Reg Quintile 2	-0.648	0.049	***
Reg Quintile 3	-1.125	0.067	***
Reg Quintile 4	-0.911	0.070	***
Reg Quintile 5	-1.028	0.043	***
Republican * Reg Quintile 2	2.702	0.054	***
Republican * Reg Quintile 3	2.141	0.071	***
Republican * Reg Quintile 4	2.023	0.074	***
Republican * Reg Quintile 5	1.577	0.050	***
Constant	-4.582	0.037	***
N = 323123			
Pseudo R2 = .38		Log Likelihood = -91077.752	
*** P-value < .01			

Logistic Regression of Probability of Voting in 2004 General Election

Dependent Variable: Voted in 2004 General Election

	Coefficient	St. Error	P-Value
Prim	0.381	0.006	***
Gen	1.394	0.003	***
Pprim08	0.597	0.006	***
Age	-0.005	0.000	***
Age Squared	0.000	0.000	***
Republican	0.374	0.005	***
Democrat	0.181	0.008	***
Other Party	-0.008	0.014	
Reg Quintile 2	0.268	0.111	**
Reg Quintile 3	1.684	0.111	***
Reg Quintile 4	0.278	0.111	**
Reg Quintile 5	0.514	0.111	***
Age Quintile 2	-0.320	0.111	***
Age Quintile 3	0.327	0.111	***
Age Quintile 4	0.841	0.111	***
Age Quintile 5	0.649	0.112	***
Reg Quint 2 * Age Quint 2	0.098	0.112	
Reg Quint 2 * Age Quint 3	0.022	0.112	
Reg Quint 2 * Age Quint 4	-0.257	0.112	**
Reg Quint 2 * Age Quint 5	-0.114	0.112	
Reg Quint 3 * Age Quint 2	0.407	0.112	***
Reg Quint 3 * Age Quint 3	-0.328	0.111	***
Reg Quint 3 * Age Quint 4	-0.832	0.112	***
Reg Quint 3 * Age Quint 5	-0.629	0.112	***
Reg Quint 4 * Age Quint 2	0.303	0.112	***
Reg Quint 4 * Age Quint 3	-0.490	0.112	***
Reg Quint 4 * Age Quint 4	-1.054	0.112	***
Reg Quint 4 * Age Quint 5	-1.092	0.112	***
Reg Quint 5 * Age Quint 2	0.431	0.112	***
Reg Quint 5 * Age Quint 3	-0.402	0.112	***
Reg Quint 5 * Age Quint 4	-0.904	0.112	***
Reg Quint 5 * Age Quint 5	-0.859	0.112	***
Constant	-1.636	.11	***

N = 1483426

Pseudo R2 = 0.3407

Log Likelihood = -660129.7

\*\*\* P-value<.01 \*\* P-value<.05

**Table 2:**

**Correlation between Survey Completion and Predicted Voting in the Election**

**Logistic Regression**

Dependent Variable - Began Online Survey in General Election

	Coefficient	Robust Standard Error	
Predicted Probability	2.025	0.181	***
Constant	-4.181	0.149	***
N=9978			

Dependent Variable - Began Online Survey in Primary Election

Predicted Probability	1.806	0.126	***
Constant	-3.256	0.064	***
N=9877			

\*\*\* P-value < .001

**Table 3. Final\* Preelection Poll Estimates of the Outcome of the 2008 GOP Primary and General Election**

\*COMBINED PPS AND SRS SAMPLES

					Predictive Accuracy (A)	
<b>3rd Congressional District Primary</b>	<b>Cannon</b>	<b>N</b>	<b>Chaffetz</b>	<b>N</b>	(A)	95% C.I.
Actual Results	40.22%	19,255	59.78%	28,618		
Poll Prediction	43.15%	192	54.38%	242	0.165	(-.025, .354)
Margin of Error $\pm$ 4.7%						
<hr/>						
<b>2008 General Election</b>	<b>Republican</b>		<b>Democrat</b>			
<b>President</b>	<b>McCain</b>		<b>Obama</b>			
	63.09%	571,115	33.92%	307,016		
Poll Prediction $\pm$ 4.1%	61.10%	377	31.44%	194	0.04	(-.13, .22)
<b>Governor</b>	<b>Huntsman</b>		<b>Springmeyer</b>			
	77.90%	700,565	19.49%	175,301		
Prediction $\pm$ 4.1%	78.22%	474	17.49%	106	0.11	(-.10, .32)
<b>Attorney General</b>	<b>Shurtleff</b>		<b>Hill</b>			
	69.92%	621,773	26.37%	234,466		
Prediction $\pm$ 4%	70.83%	425	27.17%	163	-0.02	(-.20, .16)
<b>US District 1</b>	<b>Bishop</b>		<b>Bowen</b>			
	65.12%	190,185	30.29%	88,450		
Prediction $\pm$ 7.4%	60.54%	112	34.05%	63	-0.19	(-.50, .12)
<b>US District 2</b>	<b>Dew</b>		<b>Matheson</b>			
	34.73%	113,342	63.13%	206,007		
Prediction $\pm$ 6.7%	33.48%	75	62.95%	141	-0.03	(-.31, .25)
<b>US District 3</b>	<b>Chaffetz</b>		<b>Bennion</b>			
	66.06%	180,218	27.84%	75,941		
Prediction $\pm$ 7.1%	66.50%	133	27.50%	55	0.019	(-.30, .33)

**Table 4.**  
**Comparison of Predictive Accuracy (A) Measure for PPS sample vs. SRS**

	<b>PPS</b>		<b>SRS</b>		<b>Winner</b>
<b>3rd District Primary</b>	0.083	(-.23,.39)	0.211	(-.03,.45)	PPS
N	161		271		
<b>2008 General Election</b>					
<b>President</b>	-0.086	(-.32,.14)	0.217	(-.05,.48)	PPS
N	314		255		
<b>Governor</b>	0.076	(-.20,.35)	0.15	(.03,.47)	PPS
N	324		254		
<b>Attorney General</b>	-0.142	(-.38,.09)	0.149	(-.13,.43)	PPS
N	330		257		
<b>US District 1</b>	-0.561	(-.96,.16)	0.35	(-.17,.87)	SRS
N	98		77		
<b>US District 2</b>	-0.36	(-.76,.04)	0.328	(-.07,.73)	SRS
N	119		97		
<b>US District 3</b>	-0.057	(-.47,.35)	0.122	(-.37,.61)	PPS
N	107		81		