The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis.

This electronic document was made available from www.rand.org as a public service of the RAND Corporation.

Skip all front matter: Jump to Page 1 ▼

Support RAND

Browse Reports & Bookstore

Make a charitable contribution

For More Information

Visit RAND at www.rand.org

Explore the RAND Corporation

View document details

This report is part of the RAND Corporation research report series. RAND reports present research findings and objective analysis that address the challenges facing the public and private sectors. All RAND reports undergo rigorous peer review to ensure high standards for research quality and objectivity.

Methodology of the RAND Continuous 2012 Presidential Election Poll

Arie Kapteyn, Erik Meijer, and Bas Weerman

RAND Labor and Population

For more information on this publication, visit www.rand.org/t/rr858

Preface

This paper was originally published in 2012 as part of the RAND working paper series. The RAND Continuous 2012 Presidential Election Poll implemented an innovative way to forecast the results of the U.S. presidential election. The sample for the study consisted of members of the RAND American Life Panel, which is an Internet panel that is representative of the U.S. population of 18 years and older. The 2012 presidential election poll was initiated and financed by RAND. This research was conducted in RAND's Labor and Population unit. For inquiries related to RAND Labor and Population, please contact Krishna Kumar, Director, RAND Labor and Population, at kumar@rand.org.

Table of Contents

Tables

Introduction

The RAND Continuous 2012 Presidential Election Poll implements an innovative way to forecast the results of the U.S. presidential election that will be held on November 4, 2012. Its two main innovations are that respondents are asked to express the percent chance of voting for each candidate and that the study is conducted within a panel. The innovations have been pioneered by Delavande and Manski (2010), and they have shown that this is a successful method of eliciting respondents' likely voting behavior. Traditional election polls ask respondents who they would vote for if the election were held today. They also often ask respondents how likely it is that they will vote, with typical answer categories "definitely vote," "probably vote," probably not vote," "definitely not vote."

Delavande and Manski (2010) call these "verbal questions." They call the questions asking for the percent chance of voting and the percent chance of voting for each candidate if one would vote "probabilistic questions." They administered both verbal and probabilistic questions in the American Life Panel (ALP) in 2008, and they asked respondents after the elections about their actual voting. They found that the verbal and probabilistic questions showed a large amount of agreement. In early August, the probabilistic questions were better predictors of actual voting than the verbal questions, but in late October, the verbal questions were better predictors. Also, combining current and previous answers was found to improve predictions of actual voting over just current answers.

The current study expands on the Delavande and Manski approach. New results become available every day, and they are automatically incorporated in daily graphs of results, which are posted on https://mmicdata.rand.org/alp/index.php?page=election. The current document describes the methodology behind the graphs in detail.

Sample

The sample for the study consists of members of the RAND ALP, which is an Internet panel that is representative of the U.S. population of 18 years and older. Representation is attained by using standard survey sampling methods such as random digit dialing and address sampling and not requiring prior Internet access. RAND provides members who did not previously have Internet with Internet access. See https://mmicdata.rand.org/alp/index.php?page=panelcomposition for more information about the sampling and composition of the ALP.

In June 2012, all 5,301 ALP members who are U.S. citizens were asked whether they would be willing to participate in this study. More than 3,500 of them agreed to participate. Upon consent to participation, panel members are invited once a week to answer a limited number of questions about the election. They receive $2 for each completed survey. Every day, one-seventh

of the panel members are asked to respond to the questions discussed below. Typically, respondents answer the survey on the day they are invited, but we allow them to delay by up to six days. After seven days, they are invited for the next survey and they cannot return to the previous one. Thus, respondents receive their invitations always on the same day of the week. The first invitations were sent on July 5, 2012.

We constructed the graphs with daily samples based on a "rolling window." Specifically, each daily sample is based on a seven-day window, consisting of the past seven days, counting from the day it is assembled. So for a sample assembled on, say, Tuesday, July 17, the sample window is Tuesday, July 10, up to and including Monday, July 16. This balances the desire to follow the developments daily and to ensure high statistical precision by having as large a sample as possible, with a stable composition.

To construct the daily sample, we select the survey(s) answered within the seven-day sample window by each participant, if there are any. Due to irregular timing, it is possible that a respondent may have answered two surveys within the sample window. In this case, both are included. For simplicity, they are treated as independent observations. This may lead to slight underestimation of the standard errors. Because most respondents complete the survey on the day they are invited, and there can be at most two observations for a person, this is a minor issue.

The "previous" survey (for computing changes in predicted voting behavior) is the most recent survey answered before the one included in the sample. This will typically be the survey in the prior week, but because of missed surveys or irregular timing, this previous survey may have been earlier, or within the current sample window. Also, if the actual previous survey was filled out inconsistently, it is dropped from the sample (see Section 2) and the "previous" survey for computational purposes is the most recent consistently answered one.

Data Cleaning

In every survey except the first one, there are three questions. Question Q1 asks for the percent chance that the respondent will vote in the election; Q2 is a question consisting of three components: Q2a is the percent chance of voting for Obama if the respondent would vote; similarly, Q2b is the analogous percent chance of voting for Romney, and Q2c is the analogous percent chance of voting for another candidate. Question Q3 asks about the respondent's prediction of who will win the election, again with three components: Q3a is the percent chance that Obama will win, Q3b is the percent chance that Romney will win, and Q3c is the percent chance that another candidate will win. See the appendix or https://mmicdata.rand.org/alp/elections.php for the exact question wording and screen layout.

The first time panel members respond, they are asked two preliminary questions about their voting behavior in the 2008 U.S. presidential elections: Question I1 asks whether the individual voted in the 2008 Presidential election, and if yes, I2 asks whether the individual voted for Obama, McCain, or another candidate.

For each individual, the order of the Republican and Democratic candidates is randomized. However, the same order is used for all surveys and all questions answered by a respondent, in order not to confuse the respondent.

The questionnaire attempts to enforce consistent answers—that is, answers should be percentages between 0 and 100, and the components of Q2 and Q3 should add to 100% . (In Q2 and Q3, we treat blanks as zeros. This happens often, especially if the respondent's preferred candidate receives a value of 100%.) However, it is not possible to enforce consistency completely. For example, a respondent can terminate the survey prematurely. Fortunately, inconsistencies are rare. We remove interviews with inconsistencies from the sample prior to computing derived variables (in particular, changes in predicted voting behavior) and statistics.

Reweighting

As with nearly all survey data, we use weights to ensure that the sample matches the population of interest on a number of characteristics that are potentially related to outcomes of interest. For the current study, we compute weights in two stages. In the first stage, we compute a base weight for each individual. This base weight, which we will denote by w_{0i} here, with i the index for the individual, matches the distribution of socio-economic and demographic characteristics of the sample of panel participants to population distributions as estimated from the BLS/Census Bureau's Current Population Survey (CPS). Variables that are matched are sex, age categories, race-ethnicity, education, household size, and family income. A more detailed description of the weighting method employed can be found on the dedicated page on the ALP website (https://mmicdata.rand.org/alp/index.php?page=weights).

The second stage of weighting is done for each daily sample separately. This consists of reweighting the daily sample (poststratification) such that its recall of voting behavior in 2008 matches known population voting behavior in 2008. This is based on the premise that the best predictor of future voting behavior is past voting behavior, and that any discrepancies in composition with respect to past voting behavior thus are likely to give biased predictions of voting behavior in the 2012 election.

Information about voting in 2008 is obtained from the answers to I1 and I2 in the first survey. Unfortunately, more than 20% of the respondents did not answer I1 and I2. However, almost half of these respondents participated in an earlier election survey, conducted immediately after the presidential elections in November 2008 (ms52; see https://mmicdata.rand.org/alp/index.php?page=data&p=showsurvey&syid=52). Hence, we can use the information from that survey to obtain information about voting behavior in 2008. (For respondents who answered these questions both in 2008 and in 2012, we have found that more than 90% of the answers are the same, thus giving further confidence in the reliability of this information.), Additionally, a few current participants who did not answer I1 and I2 were too

young to vote in 2008, and thus we know that they must have been nonvoters. This leaves a group of about 6.5% of the participants for whom we do not know their voting behavior in 2008.

We then classify each individual uniquely in one of the following groups (poststrata): (G1) 18–21 years old (in 2012), not eligible to vote in 2008; (G2) 22+, did not vote in 2008; (G3) voted for Obama in 2008; (G4) voted for McCain in 2008; (G5) voted for another candidate in 2008; (G6) 22+, unknown voting behavior in 2008.

The following information is employed to determine the poststratified weights:

- As estimated from the CPS, 7.4% of U.S. citizens 18 and over are 18–21 years old. Hence, after reweighting, G1 should be 7.4% of the sample.
- Voter turnout in 2008 can be obtained from the estimates provided by the United States Election Project at George Mason University (McDonald, 2012): the voting eligible population was 212,702,354 and turnout for highest office (president) was 131,304,731; hence, the turnout rate was 61.73% and $100 - 61.73 = 38.27\%$ of the eligible population were nonvoters.
- Obama received 52.92% of the votes in 2008, McCain received 45.66% of the votes, and other candidates jointly received the remaining 1.42% of the votes (http://www.fec.gov/pubrec/fe2008/2008presgeresults.pdf). Combining this with the information about the turnout rate, it follows that $0.6173 * 52.92 = 32.67\%$ of the eligible population voted for Obama, $0.6173 * 45.66 = 28.19\%$ voted for McCain, and $0.6173 * 1.42 = 0.88\%$ voted for another candidate.

Let W_{0g} be the sum of the base weights in group g among the participants (g = G1–G6), and let $P_U = W_{06}/(W_{02} + W_{03} + W_{04} + W_{05} + W_{06})$ be the fraction unknown among participants who were old enough to vote in 2008. Then the target stratum size for group 1 (18–21) is 0.074 (7.4%), the target stratum size for group 6 (unknown) is $P_U * (1 - 0.074)$, and the target stratum sizes for groups G2–G5 are computed such that their relative sizes are 38.27 : 32.67 : 28.19 : 0.88. Table 1 gives the exact formulas for all target stratum sizes.

Table 1. Target Stratum Sizes

Group	Target Stratum Size
G1 (18–21)	0.074
G2 (nonvoter)	$(1 - 0.074) * (1 - P_U) * 0.3827$
G3 (Obama)	$(1 - 0.074) * (1 - P_U) * 0.3267$
G4 (McCain)	$(1 - 0.074) * (1 - P_U) * 0.2819$
G5 (other)	$(1 - 0.074) * (1 - P_U) * 0.0088$
G6 (unknown)	$(1 - 0.074) * P_U$

Let S_g denote the target stratum size of stratum g and consider a specific sample, t, say. (As indicated above, this includes observations from seven days.) Let W_{tg} be the sum of the base weights in group g among the observations in sample t. Define the correction factors $F_{tg} = S_g/W_{tg}$.

The (poststratified) weight variable to be used for observation i in the sample is $w_{ti} = w_{0i} F_{tg}$ if respondent i is in group g. Thus, the sum of the poststratified weights in group g is S_g, as desired.

Note that the sum of the weights across the whole sample is

$$\sum_i w_{ti} = \sum_g \sum_{i \in Gg} w_{ti} = \sum_g \sum_{i \in Gg} w_{0i} F_{tg} = \sum_g W_{tg} \left(S_g / W_{tg} \right) = \sum_g S_g = 1$$

This simplifies many of the formulas below, because this sum would otherwise appear in the denominators of the expressions. In the following, we drop the t subscript and simply write w_i, where it is understood that this is a weight variable that is specific to the daily sample t under consideration.

Voter Turnout

Voter turnout, the percentage of the voting eligible population (VEP) who will vote, is predicted as the weighted average of the percent chance of voting by each individual:
$$VT = \sum_i w_i Q1_i .$$

Popular Vote (Weekly Poll)

The percent chance that individual i will vote for candidate j (= Obama, Romney, other; or a, b, c) is $Pj_i = Q1_i \times Q2j_i/100$. Hence, the percentage of the VEP that will vote for candidate j is predicted as $FVj = \sum_i w_i Pj_i$. These percentages add up to the predicted voter turnout. The predicted fraction of the votes for candidate j is obtained by dividing these numbers by the predicted voter turnout, so that they add up to 100%:
$PVj = 100 \times FVj / VT = 100 \times \left[\sum_i w_i Pj_i \right] \Big/ \left[\sum_i w_i Q1_i \right]$. Note that this can be interpreted as a weighted average of $Q2j_i$, where the weights w_i are adjusted by the voting probabilities $Q1_i$, so that the adjusted weight variable is $(w_i Q1_i)$.

Predicted Winners

Individual i believes that the percent chance that candidate j wins the election is $Q3j_i$. The overall estimate of the percent chance that candidate j wins the election is its weighted average
$$PWj = \sum_i w_i Q3j_i .$$

Numbers of Obama Supporters and Romney Supporters

We define an Obama supporter as someone who either votes for Obama or a nonvoter who would have voted for Obama if he or she had voted. Romney supporters are defined analogously.

5

With this definition, Q2a$_i$ is the percent chance that individual i will be an Obama supporter, and Q2b$_i$ is the percent chance that individual i will be a Romney supporter. The estimated percentage of the VEP that will be supporters of candidate j is $\text{SUP}j = \sum_i w_i Q2j_i$.

Voter Turnout Among Obama and Romney Supporters

We can also predict voter turnout separately for Obama and Romney supporters, even though we only have probabilistic measures of the latter. Voter turnout among supporters of candidate j is equal to (100 times) the number of individuals who vote for candidate j divided by the number of supporters of candidate j:

$$\text{VT}j = 100 \times \text{FV}j / \text{SUP}j = 100 \times \left[\sum_i w_i Pj_i \right] / \left[\sum_i w_i Q2j_i \right].$$

This can be interpreted as a weighted average of the percent chance of voting as expressed in Q1, where the weights w_i are adjusted by the probability of being a supporter of candidate j, as measured by Q2j_i, so that the adjusted weight variable is (w_i Q2j_i).

Changes (Shifts)

Day-to-day or week-to-week changes in the aggregate predictions can be computed easily as the difference of the predicted totals. For example, if on August 5, Obama is at 49.1% and on August 6, he is at 50.5%, he gained 1.4 percentage points.

A unique characteristic of our study is that the panel nature allows us to study changes in preferences for the candidates among the same respondents. For example, we can evaluate whether a net increase in preference for Romney is due to a small number of individuals switching from Obama to Romney or a large number switching from Obama to Romney and a slightly less large number switching in the opposite direction.

For the computation of these shifts, we first combine the probability of nonvoting and the probability of voting for an "other" candidate: P0$_i$ = (100 - Q1$_i$) + Pc$_i$.

Let "−1" denote the previous value of a variable and Δ the difference between the current value and the previous value. Thus, the change in the likelihood that individual i will vote for Obama is $\Delta\text{Pa}_i = \text{Pa}_i - \text{Pa}_{i,-1}$.

Compute ΔPj_i for j = 0, a, b. If (only) one of these three is negative, this becomes the "donor" and the other two become the "recipients." We can then compute $L_i(j, k)$, the shift from candidate j to candidate k (treating nonvoting and voting for "other" jointly as a single candidate). For example, if $\Delta\text{Pa}_i = -10\%$, $\Delta\text{Pb}_i = +6\%$, and $\Delta\text{P0}_i = +4\%$, we interpret this as 6% going from candidate a (Obama) to candidate b (Romney) and 4% from a to 0 (nonvoting+other). Thus, L_i(a, b) = 6%, L_i(a, 0) = 4%, L_i(b, 0) = 0, L_i(b, a) = 0, L_i(0, a) = 0, and L_i(0, b) = 0.

Similarly, if two out of three are negative, both of these are donors feeding into the third: if $\Delta\text{Pa}_i = -10\%$, $\Delta\text{Pb}_i = -6\%$, and $\Delta\text{P0}_i = +16\%$, we interpret this as 10% going from a to 0 and 6%

from b to 0, with $L_i(a, b) = 0$, $L_i(a, 0) = 10\%$, $L_i(b, 0) = 6\%$, $L_i(b, a) = 0$, $L_i(0, a) = 0$, and $L_i(0, b) = 0$.

The aggregate flows are the weighted sample means: $\bar{L}(j, k) = \sum_i w_i L_i(j, k)$.

Standard Errors and Shaded Areas

Standard errors for estimators based on poststratified weights can be obtained using the methods and formulas given in StataCorp (2009a, p. 1026; 2009b, pp. 50–51, 160–162). Here we apply these to the specific estimators we use.

Several statistics discussed above—for example, the predicted percent chance that a candidate will win the elections—are of the form $\bar{X} = \sum_i w_i X_i$, where X is a reported or derived variable. Standard errors of such averages are computed as follows. First, define (and compute) the group-specific means $\bar{X}_g = \sum_{i \in Gg} w_i X_i / \sum_{i \in Gg} w_i$. Next, for each observation, compute the deviation from the group mean: if observation i is in group g, then $e_i(X) = X_i - \bar{X}_g$. Then the standard error of the mean is

$$se(\bar{X}) = \sqrt{\sum_i \left[w_i e_i(X) \right]^2}.$$ (1)

Other statistics, such as the estimate of the popular vote for a candidate, are ratios of means of the form $R = 100 \times r = 100 \times \bar{Y} / \bar{X} = 100 \times \sum_i w_i Y_i / \sum_i w_i X_i$. Their standard errors are

$$se(R) = \frac{100}{\bar{X}} \sqrt{\sum_i \left\{ w_i \left[e_i(Y) - r\, e_i(X) \right] \right\}^2}.$$ (2)

In the graphs, we shade the area within which the difference between Obama and Romney is not statistically significant. This area is centered around the midpoint between the Obama and the Romney graph, and its total (vertical) width is 1.96 times the standard error of the difference between the Obama and the Romney graph. Thus, by construction, at a given point in time (corresponding to a given sample), either both the Obama and Romney graph are in the shaded area, or they are both outside the shaded area. If they are in the shaded area, their difference is not statistically significant at the 5% level, and if they are outside the shaded area, the difference is statistically significant at the 5% level. To be able to construct this shaded area, we therefore need the standard error of the difference between the Obama and Romney graphs.

For statistics of the form $\bar{X} = \sum_i w_i X_i$, such as the predicted percent chance that a candidate will win the elections, let X_{ai} be the Obama variable and X_{bi} be the Romney variable. Then the difference between the Obama average and the Romney average is

$$\Delta \bar{X} = \bar{X}_a - \bar{X}_b = \sum_i w_i X_{ai} - \sum_i w_i X_{bi} = \sum_i w_i (X_{ai} - X_{bi}) = \sum_i w_i \Delta X_i,$$

7

where $\Delta X_i = X_{ai} - X_{bi}$ is the difference between the Obama and Romney variable at the individual level. The rightmost expression is a standard average of a derived variable, and thus its standard error follows from equation (1) with X_i replaced by ΔX_i.

The standard error of the difference between the popular vote for Obama and Romney can be defined in a similar way: define $Y_i = \mathrm{Pa}_i - \mathrm{Pb}_i$ as the difference between the probabilities of voting for Obama and Romney at the individual level. Then the difference between the popular vote for Obama and Romney is of the ratio form, and its standard error is given by equation (2).

The standard error of the difference in voter turnout between Obama supporters and Romney supporters is computed as follows. This difference can be written as

$$\Delta R = R_a - R_b = 100 \times r_a - 100 \times r_b = (100 \times \bar{Y}_a / \bar{X}_a - 100 \times \bar{Y}_b / \bar{X}_b)$$
$$= 100 \times \left(\sum_i w_i Y_{ai} \Big/ \sum_i w_i X_{ai} - \sum_i w_i Y_{bi} \Big/ \sum_i w_i X_{bi} \right).$$

Its standard error is

$$\mathrm{se}(\Delta R) = 100 \sqrt{ \sum_i \left\{ w_i \left[\frac{e_i(Y_a) - r_a e_i(X_a)}{\bar{X}_a} - \frac{e_i(Y_b) - r_b e_i(X_b)}{\bar{X}_b} \right] \right\}^2 } \ .$$

Statistics and Graphs Broken Down by Other Characteristics

We also compute statistics and plot graphs broken down by other characteristics, typically demographics such as sex or ethnicity. Computing statistics and standard errors for these involves only minor adaptations of the formulas in the previous sections. Let D_i be a dummy variable that indicates whether an individual has the characteristic (e.g., is female; $D_i = 1$) or not (male; $D_i = 0$). Then the average of a variable X for individuals with the characteristic is $\bar{X}_{D=1} = \sum_i w_i X_i D_i \Big/ \sum_i w_i D_i = \sum_i w_i Y_i \Big/ \sum_i w_i D_i = \bar{Y} / \bar{D}$, with Y implicitly defined. This has the ratio form for which the standard errors can be computed as in equation (2). Similarly, the popular vote for a candidate takes the form

$$R = 100 \times \bar{Y}_{D=1} / \bar{X}_{D=1} = 100 \times \sum_i w_i Y_i D_i \Big/ \sum_i w_i X_i D_i = 100 \times \sum_i w_i Z_i \Big/ \sum_i w_i V_i \ ,$$

with Z and V implicitly defined, and its standard errors also follow from equation (2). Note that the respondents who do not have the characteristic ($D_i = 0$) do contribute to the computation of the standard errors.

For computing statistics for the "other" characteristic (male instead of female), we simply redefine D and apply the same formulas.

Appendix: Questionnaire

Below are the questions exactly as the respondents see them on the screen. The order of the candidates is randomized so that about half of the respondents always see questions with first Obama and then Romney, whereas the other half of the respondents always see questions with first Romney and then Obama.

Several messages are embedded in the survey program. They are listed below. It should be obvious what responses trigger the messages below:

You already completed this survey today. Thank you for your participation!

Please make sure the percentages are between 0% and 100%.

Please answer the questions before pressing Next.

Please make sure the total adds up to 100%.

What is the percent chance that you will vote in the Presidential election?

[   ] %

If you do vote in the election, what is the percent chance that you will vote for Obama? And for Romney? And for someone else? Please provide percent chances in the table below:

Barack Obama (Democrat)	[ ]	%
Mitt Romney (Republican)	[ ]	%
Someone else	[ ]	%
Total	[ ]	%

What do you think is the percent chance that each of the candidates for president will win the election?

Barack Obama (Democrat)	[ ]	%
Mitt Romney (Republican)	[ ]	%
Someone else	[ ]	%
Total	[ ]	%

RAND
American Life
Panel

References

Delavande, A., & Manski, C. F. (2010). Probabilistic polling and voting in the 2008 presidential election: Evidence from the American Life Panel. *Public Opinion Quarterly*, *74*, 433–459. doi:10.1093/poq/nfq019

McDonald, M. P. (2012). 2008 General Election Turnout Rates. *United States Elections Project*. Available from http://elections.gmu.edu/Turnout_2008G.html (accessed August 13, 2012).

StataCorp (2009a). *Stata base reference manual: Release 11*. College Station, TX: StataCorp.

StataCorp (2009b). *Stata survey data reference manual: Release 11*. College Station, TX: StataCorp.