

THE RECORD OF INTERNET-BASED OPINION POLLS IN PREDICTING THE RESULTS OF 72 RACES IN THE NOVEMBER 2000 U.S. ELECTIONS

By:

Humphrey Taylor, John Bremer, Cary Overmeyer, Jonathan W. Siegel, George Terhanian

Abstract

The authors describe the use of Internet-based surveys to predict the results of 72 races in the U.S. elections of November 2000. The authors describe their results, the methods they used, including the use of both demographic and “propensity score” weighting to correct for substantial biases in the raw, unweighted data, and the investments which were needed to achieve these results. They also caution readers not to assume that identical methods can be used with equal success in other countries or to measure all other variables, and stress the need for continuing research to improve online research methods in the future.

Last winter, the International Journal of Market Research published a paper *Does Internet Research Work?* (Taylor, 2000) which described the work that Harris Interactive was doing to develop Internet-based opinion polls in the United States. It described the early stages of the development and the use of demographic and propensity score weighting to correct for biases in raw data collected from samples drawn from the millions of people who had opted in to the “Harris Poll Online Panel,” had agreed to be interviewed on the Web, and had voluntarily provided our firm with their e-mail addresses.

The November 2000 U.S. elections provided us with a unique opportunity to test the accuracy of our online survey methods and, specifically, our ability to predict elections. In total, our online research in these elections covered 72¹ different races as follows:

¹ The Arizona race for Senate has been excluded from the analyses presented here due to the inadvertent omission of one candidate from the survey and the inadvertent inclusion of another candidate on the survey.

- the nationwide vote for president
- the statewide votes for president in 38 states
- the statewide votes for senator in 26 states
- the statewide votes for governor in seven states

We did not cover all 50 states because we were not confident that we could develop adequate online samples in 12 smaller states. Specifically, we felt that some significant sub-populations, such as rural blacks in Mississippi, Alabama, Arkansas and South Carolina were not adequately represented in our database in these states.

THE RESULTS

The accuracy of our online polling efforts in these 72 races exceeded our most optimistic expectations.

1. In the national vote for president, only two polls correctly showed that the two candidates tied (the final results were 48% for both Bush and Gore) – the Harris Interactive online poll and the Harris Interactive telephone poll. However, it is worth noting that almost all the polls did very well in this election.
2. Our predictions for the presidential election votes in 38 states were also remarkably accurate. The Research Business Report (200) has published a detailed comparison of our results with the averages for all the telephone polls in these states which concluded: 'Harris Interactive was off an average 1.8 points for Gore and 2.5 for Bush predictions. Phone polling was off 3.9 on the average Gore percentage and 4.4 for Bush' and 'According to (our) RBR estimates, Harris Interactive's average predicted Bush percentage and average predicted Gore percentage were twice as accurate as its general telephone polling competition.'
3. Our predictions for the 26 Senate races we covered were also quite accurate. We made only one incorrect forecast and our average error was 2.2% for the two main candidates.
4. Our predictions for seven governors' elections were even more accurate. Our average error was 1.9% on the two main candidates.

If we had conducted only the one online poll on the national vote for the presidential election, it would have been easy to believe that our success was just luck. Indeed, when so many national polls are as close to the final result as they were in these elections, our having the candidates tied, and being 'the best poll' was surely luck in part. But when we have 72 different elections to compare, it is obvious that our accuracy cannot be explained by luck or statistical accident. We believe we have clearly demonstrated that Internet polls can be designed and executed to measure voting intentions with great accuracy.

A BRIEF SUMMARY OF OUR METHODOLOGY

The key elements of the methods used in our final online election surveys to predict the national vote and the results of 71 elections in 38 states, were:

- the development and building of a large national online panel of willing respondents

- the testing and development of a sampling and weighting approach (to adjust for the different likelihood, or propensity, of different respondents to be in our online or telephone samples), throughout 1999 and 2000.
- the conducting of two similar, but smaller, surveys in September and October, as 'trial runs' for the main event.
- Interactive online interviews conducted between October 31 and November 6 with a total of 240,666 adults who were, based on the answers, categorized as 'likely voters.'
- The weighting of the data in the 39 samples (the national sample and the 38 state samples) using both demographic weights and a propensity score reflecting the probability of respondents being online.

The *demographic weights* were virtually identical to those used in our nationwide Harris Interactive telephone poll (which was the only other poll to have Bush and Gore tied in its final prediction). They included sex, age, education, and race/ethnicity.

The *propensity score* was developed based on questions measuring behavior and attitudes which, as our proprietary research among parallel online and telephone surveys had taught us, were substantially different in samples drawn from our online database, even after demographic weighting.

As to which questions to use to develop propensity scores, we do not believe we have discovered the holy grail. We will continue to test and develop new or better questions to be used in propensity score weighting to reduce the biases in our online samples, not just for political surveys but for marketing, social and other research. In these election surveys we used questions which measured alienation, readership, participation and investment, which were selected on a strictly empirical bases from hundreds of questions which we tested. Comparing our online and telephone surveys we found that weighting by propensity scores using these questions did the most to reduce biases efficiently.

An earlier paper (Taylor 2000) about the use of propensity score weighting described the use of questions measuring health status, political party identification and the number of telephone lines as effective in reducing the biases in our online surveys. The parallel testing in telephone and online samples of many possible questions caused us to change the questions previously used for propensity score weighting.

It would not have been possible to produce accurate predictions for these elections without using *both* the demographic weights and the propensity score measure.

The Impact of Both Demographic and Propensity Score Weighting

The tables in this section show the national presidential election vote and the three different kinds of state elections, with the impact of both demographic and propensity score weighting on reducing the errors in our forecasts.

For the nationwide presidential vote the raw, unweighted data produced an error Bush lead of 9%. This was reduced to 4% by demographic weighting and disappeared when the propensity some weights were added (Table 1).

Table 1
The Nationwide Presidential Election Vote

	Harris Interactive Date			Result %
	Raw Unweighted %	Demographically Weighted %	Propensity Score Weighted* %	
George W. Bush	51	49	47	48
Al Gore	42	45	47	48
Ralph Nader	4	4	4	3
Pat Buchanan	1	1	1	0
Other candidates	2	1	1	1
Lead (Spread)	9	4	0	0
Error on Lead	9	4	0	NA
Average Error on 2 Main Candidates	4.5	2.0	1	NA

NOTE: All numbers rounded to the nearest percentage point.

* Includes both demographic and propensity score weighting.

For the presidential vote in 38 states, for which we produced 38 separate forecasts, the raw, unweighted data produced an average error on the lead of 7.1 percentage points, which was reduced to 4.0% by demographic weighting and to 3.4% by propensity score weighting (Table 2).

TABLE 2

The Impact of Weighting on 38 States on the Error on the Gap Between Bush and Gore

STATE	RESULT	Raw, Unweighted		Demographic (All Weighted)		Propensity Score Weighted*	
	(Gap Between Bush and Gore)	Gap	Absolute Error	Gap	Absolute Error	Gap	Absolute Error
Alaska	31	29	2	27	4	23	8
Arizona	6	9	3	6	0	12	6
California	-12	-2	10	-6	6	-6	6
Colorado	8	10	2	8	0	5	3
Connecticut	-18	-9	9	-7	11	-11	7
Florida	0	6	6	-1	1	-3	3
Georgia	12	27	15	11	1	8	4
Hawaii	-18	0	18	-2	16	-15	3
Idaho	40	34	6	40	0	38	2
Illinois	-12	-1	11	-6	6	-6	6
Indiana	16	21	5	16	0	17	1
Iowa	0	1	1	0	0	6	6
Kansas	21	20	1	19	2	17	4
Kentucky	15	18	3	15	0	12	3
Maine	-5	-7	2	-2	3	-8	3
Maryland	-16	-1	15	-11	5	-17	1
Massachusetts	-27	-21	6	-19	8	-26	1
Michigan	-5	+2	7	-5	0	-5	0
Minnesota	-2	-1	1	3	5	-3	1
Missouri	3	14	10	6	3	3	0
Nebraska	29	29	0	25	4	19	10
Nevada	4	11	7	7	3	3	1
New Hampshire	1	9	8	8	7	-2	3
New Jersey	-16	-8	8	-13	3	-14	2
New Mexico	0	23	23	15	15	-5	5
New York	-24	-16	8	-16	8	-15	9
North Carolina	13	20	7	9	4	5	8
Ohio	4	10	6	6	2	3	1
Oklahoma	22	29	7	28	6	22	0
Oregon	0	1	1	-1	1	-7	7
Pennsylvania	-4	-1	3	-5	1	-5	1
Tennessee	4	21	17	10	6	2	2
Texas	21	36	15	23	2	18	3
Utah	40	38	2	39	1	40	0
Virginia	8	16	8	9	1	6	2
Washington	-6	0	6	2	8	1	7
West Virginia	6	11	5	11	5	4	2
Wisconsin	0	4	4	2	2	0	0
Average Absolute Error	-	-	7.1	-	4.0	-	3.4

NOTE: All numbers rounded to the nearest percentage point.

* Includes both demographic and propensity score weighting.

For the state elections for senator in the 26 races covered, the average error on the lead was 9.4 percentage points in the unweighted data, 5.8% with only demographic weighting and 4% percentage points with propensity score weighting (Table 3).

Table 3

The Senate Election Votes in 26 States

	Harris Interactive Date		
	Raw Unweighted %	Demographically Weighted %	Propensity Weighted %
Average absolute error on the lead	9.1	5.8	4.0
Average absolute error on two main candidates	4.6	3.1	2.2

In the races for governor in seven states, the unweighted data produced, on average, a 6.5% error, which was reduced to 3.5% by demographic weighting and to 2.9% by propensity score weighting (Table 4).

Table 4

The Gubernatorial (Governors') Election Votes in 7 States

	Harris Interactive Data		
	Raw Unweighted %	Demographically Weighted %	Propensity Weighted %
Average absolute error on the lead	6.5	3.5	2.9
Average absolute error on two main candidates	3.3	2.6	1.8

ESSENTIAL INVESTMENTS WHICH MADE THESE RESULTS POSSIBLE

One myth that needs debunking is that 'online research is cheap.' Bad online research may well be cheap. A 'call-in' online poll, with no attempt to select a sample, no weighting and no real interactivity is cheap. But our experience is that good online research costs real money. Of course, the marginal cost of data collection is very inexpensive. There are no interviewers, and only very modest telephone bills, to be paid. In the research described here, the actual data collection costs of surveying 240,666 likely voters were modest. But the capital costs of setting up everything needed to do good online research are substantial.

Our successes in predicting the 2000 elections were only possible because our firm made four substantial investments:

- (1) *Investments in building a big online panel of willing respondents.* It would have been possible to predict the *national* vote with a small database. However, we needed a much larger database to conduct the substantial number of interviews with 'likely voters' in the 38 selected states. (Parenthetically we need a much bigger database – currently over seven million worldwide – in order to be able to conduct many surveys, or many topics, for many clients, every month).
- (2) *Investments in new systems, hardware and software.* Investment in technology was necessary to be able to survey large numbers of people with 'better-than-CATI' interactivity (for other research this may involve the use of streaming video, audio or other stimuli).
- (3) *Investment in people and skills.* We needed to hire not only many new people with systems and statistical skills but also people who, because of their background in academic, medical or educational research, were experienced in working with non-probability samples.
- (4) *Investment in research methods.* In order to develop, test and improve both the selection of variables to be used for propensity weighting and the specific weights, it was necessary to run over one hundred parallel online and telephone surveys.

Had we been unable to make these four major investments, our election predictions would surely have been much less accurate.

WHAT THESE RESULTS MEAN – AND DO NOT MEAN

The remarkable accuracy of these online polls in predicting the results of 72 races proves that well-designed online polls *can* be a very reliable way to predict elections. Or, more specifically, that online polls in the United States – and probably in other countries when they achieve a high enough Internet penetration – *can* do this.

However, it would be a mistake to assume that all online pre-election surveys will be reliable election predictors. There are enormous differences in the ways different organizations are using the Internet to conduct research. These differences are much greater than the differences in the methods used to conduct telephone surveys.

All users of online (and other) surveys must remember that no amount of weighting can correct for biases in variables which are close to zero or 100% in either the sample or the population not sampled (i.e. those with **or** without telephone, those who use **or** do not use computers, etc.). Fortunately for us, it seems that none of these 100%/0% variables was independently correlated with voting intentions for different political candidates. Therefore, the weighting model worked.

THE IMPORTANCE OF WEIGHTING

These elections were a wonderful way to test the work we had put into developing our online research skills to measure political attitudes and, specifically, to predict elections. These results should also put to rest the main criticism of online research in the United States. Our fiercest critics have argued that

because we do not use probability samples of all households or all adults, online research cannot work. Europeans, however, with a long and successful history of using quota sampling have been more accepting of non-probability methods.

We have known for a long time that all probability samples have significant, and often substantial, biases which we can sometimes be reduced substantially by weighting. The same, we have demonstrated, is true for online polling on political behavior. However, one key to the development of new survey methods is to run very scared – and to keep running scared. We believe, therefore, that it will be a dangerous mistake to assume that, because of our success in the 2000 Elections, any of the following are true:

- (1) The same weighting variables and weights could be used in other countries with equal success.
- (2) The same weighting variables and weights will work equally well in two or four years time, or further ahead, in other U.S. elections.
- (3) The same weighting variables and weights will work equally well with samples drawn from other databases, or by other means, in the USA.
- (4) The same weighting variables and weights will work equally for surveys on all topics, even when using the same database in the USA.

Indeed, it is our strong belief that all of these assumptions, with the possible exceptions of number (2) will prove to be wrong. If so, the ability to conduct reliable online research will depend on a continuing investment in testing and improving different weighting variables. Research into improving online survey methodology will need to continue for a long time to come.

Lincoln Steffens went to the Soviet Union soon after the Russian Revolution and said "I have seen the future and it works." Following these elections that's how we feel about Internet-based research. Let's hope that pride does not come before a fall and that our forecast is better than Steffens'.

References

Taylor, H.J.F. (2000) Does Internet Research Work? *International Journal of Market Research*, 42, 1, pp. 51-63.

Research Business Report (2000) Harris Interactive Uses Election 2000 to Prove its Online Efficacy and Accuracy. November.