Heady Days Are Here Again

Online polling is rapidly coming of age

By Humphrey Taylor and George Terhanian

Whether we like it or not, online survey research will be a big part of our future. But these are very early, if heady, days for online research. Despite evidence suggesting that online surveys can produce credible, trustworthy information under many circumstances, we still have much to learn before we know how good online surveys can be.

Those of us who work and live in the Internet Revolution understand that the "internet clock" runs much faster than a normal clock. Two years ago, we knew almost nothing about how to conduct online surveys. Today, we know much more—so much more that we occasionally fall into the trap of thinking we have all the answers. When this happens, we try to heed Warren Mitofsky's advice that, "There's a lot of room for humility in polling... Every time you get cocky you get it wrong."[1]

While we have learned a lot about internet polling, we only have a fraction of the knowledge that will be accumulated in a year or two.

One test of the credibility of any new data collection method hinges on its ability to reliably and accurately forecast voting behavior. For this reason, last fall we attempted to estimate the 1998 election outcomes for governor and the US Senate in 14 states on four separate occasions using internet surveys. We mounted this experiment in public for a myriad of reasons, including our desire to advance the collective understanding of our survey research colleagues who, at times, tend to dismiss internet-based research without any empirical evidence. Put bluntly, this sort of categorical dismissal is imprudent and irresponsible. As Walter Lippman wrote, without empirical evidence, we may "leave matters to the unwise... those who... greatly imperil the future... by leaving great questions to be fought out between ignorant change on the one hand, and ignorant opposition to change on the other."[2]

So how did we do? In our October 19-21 and October 29- 31 surveys, we correctly projected the winner in 21 of 22 (95%) races (see Table 1). In comparison, 49 of 52 (94%) telephone polls that were published in *Hotline* or posted on the CNN website for these same races made correct projections. Of course, "percent correct" is a crude, potentially misleading, indicator of the accuracy of election forecasts. For this reason, we computed two additional performance indicators: the average "spread" error and the average "candidate" error for our projections.

In the October 29-31 survey, we projected that Charles Schumer would defeat Alfonse D'Amato, 49% to 44%, or by 5 percentage points. Schumer actually defeated D'Amato, 54% to 45%, or by 9 points. We therefore misforecast the spread between these candidates by 4 points—our "spread" error for this race. For our 22 projections in the October 29-31 surveys, our average "spread" error was 6.8 points, about a half a percentage point more than the average "spread" error for the 52 telephone polls (6.2 points).

Our estimate of Schumer's vote share was off by 5 points (54% vs. 49%), while for D'Amato, we missed by 1 point (45% vs. 44%). Our average "candidate" error in New York, then, was 3.0 points (6 divided by 2). For our 22 projections, the average "candidate" error was 4.4, about a half percentage point less than the average "candidate" error for the 52 telephone polls (5.0). (Note that we did not reallocate "don't know" responses to the main candidates.)[3]

Although the accuracy of our forecasts surprised many colleagues in the survey research community—some of whom work within our own organization—we did not do nearly as well as we would have liked. But we certainly learned a lot. Above all else, we know that we need to learn much more about many sampling issues, about how to improve the weighting of online samples, and about the method effects of surveying on the internet.

"Our survey research colleagues, at times, tend to dismiss internet-based research without any empirical evidence. Put bluntly, this sort of categorical dismissal is imprudent and irresponsible."

The first sampling issue to confront is the unavailability of a comprehensive list of e-mail addresses of the internet population. Anyone who wants to conduct a telephone survey can obtain a more or less comprehensive listing of all residential telephone numbers—even if that listing also includes many business numbers and an increasing number of unallocated numbers. Unfortunately, this is not the case for e-mail addresses. Even if such a listing existed, sending e-mails to potential respondents of online surveys who have not agreed beforehand to participate in online research

Table 1

1998 election forecasting performance indicators	Harris Online Election Survey of: 10/19-21	10/29-31	11/1-2	11/2-3	Final Telephone Surveys
Percent correct	95%	95%	91%	100%	94%
Average spread error	7.6	6.8	6.5	6.3	6.2
Average candidate error	5.3	4.4	4.3	4.0	5.0

would look like "spamming" and be considered unacceptable by many people.

Online researchers must therefore depend on lists of e-mail addresses they purchase or build themselves. We are not aware of any listing that includes more than a small minority, in most cases tiny minorities, of e-mail addresses.

Another equally daunting problem is the variety of e-mail lists organizations can choose from in selecting an online sample. Some organizations have used e-mail addresses obtained from people surveyed by telephone; some use addresses obtained through advertising; still others use samples drawn from America Online and Prodigy subscribers. We have recruited our own panel through multiple sources including banner advertisements and sweepstakes that have run across the web, the Harris/Excite daily poll, product registrations on Excite and Netscape, and telephone surveys. Although we have made a concerted effort to build as diverse a sample of internet users as possible, our actual sample is by no means representative, in a statistical sense, of the US internet population or the total US population.

With different online research organizations using completely different lists, the weighting required to bring each sample into line with the total population, or even the total online population, is also different.

Because so little is known about the reliability (and unreliability) of online surveys, it is very important to run parallel telephone and online surveys, asking the same questions, during the same time frame, of comparable populations. By comparing the data we are learning rapidly about the differences between our telephone and our online surveys—differences in sampling (and when, if, and how to compensate for these by weighting), and in method effects.

Without these data from parallel studies the learning curve would be much slower. For example, every data collection method, including online surveys, has different method effects. Our experience suggests our online surveys tend to produce much higher "not sures" and different responses to attitude scales than do our telephone surveys.

These differences do not mean that telephone surveys are "right" or online samples are "wrong." It does mean that we must exercise extreme care when comparing attitudinal variables (for example, for trending purposes) in online and telephone surveys.

From a statistical perspective, the best way to make fair comparisons between two groups is to mount a controlled experiment in which members of a population are assigned at random to either a treatment or control group. Random assignment ensures that the resulting two groups are equivalent apart from chance. Of course, what is desirable is not always feasible. Under many circumstances, it is difficult or impossible to employ random assignment. For instance, it is not possible on ethical grounds to mount a randomized experiment to estimate the effect of cigarette smoking on mortality rates. Nor is it possible to randomly assign a sample of all US adults to a telephone or online survey to compare these methods—only 45% of US adults use the internet, and only about 3% of internet users belong to the Harris Poll Online community. The inability to employ random assignment in these instances does not diminish the need to compare smokers with non-smokers or telephone surveys with online surveys. But it does make these tasks more challenging.

Rubin and Rosenbaum, building on the work of Cochran and Cox, and Campbell and Stanley, developed a sturdy statistical substitute for random assignment called "propensity score adjustment" for occasions when random assignment is difficult or impossible.[4] The propensity score is a single, summary measure that represents the probability of belonging to one group rather than another, where the probability is indicated by some observable characteristics of the individual. The observable characteristics used to estimate a propensity score typically include demographic and contextual information.

The propensity score has many excellent features. Notably, it efficiently and economically accounts and adjusts for differences in the observable characteristics of individuals who belong to one group rather than another.

This statistical work bears directly on our continued efforts to validate findings from our online surveys. After we mount parallel telephone and online surveys, for example, we attempt to weight our online survey in the interest of balancing the observable characteristics of the telephone and online respondents. It is quite easy to select a handful of characteristics to balance. For instance, we know that we must overweight online respondents who are aged 65 and older because 5% of all online users are aged 65 and older whereas 17% of all US adults are 65 and older. When we employ a traditional rim-weighting approach, we introduce more and more error into our online estimates as we balance more and more characteristics. For this reason, we generally select but a few characteristics to adjust.

When we take a propensity score approach, however, we are able to select many more variables for adjustment without increasing the error of our estimates. (Remember, a propensity score is a single, summary measure that represents a collection of variables.) Of course, a propensity score is only as good as the variables that were included in the model to estimate it in the first place, and understanding what variables to include is hard work.

All surveys—indeed, all censuses—miss some people. Whether or not it is possible to project from a sample survey, or a census, to the total population depends on several factors, including, of course, our pre-existing knowledge of differences between the sample, the sampling frame (i.e., the list), and the total population.

Telephone surveys miss, among others, people who live in homes without telephones, people who avoid surveys and people who refuse to be interviewed. Fortunately, demographic and other weighting can be used to overcome most of these differences for measuring some, but not all, variables.

However, no amount of weighting, no matter how sophisticated, can ever compensate for variables which are zero percent, or one hundred percent, of the sample or the population (but not both). However much you weight zero, it is still zero. And, in practice, this is also true for variables which are close to zero or one hundred percent of either the sample or the universe.

Obviously, if you want to survey people who do not have computers or who are not online, you cannot do that online.

At the 1999 AAPOR conference in St. Petersburg, Florida, Linda Piekarski of Survey Sampling gave a dramatic and scary presentation of trends in telephone sampling. She talked about the growth of technological barriers (e.g., cellular phones, caller ID, answering machines, multiple lines), the negative impact of telemarketing, declining contact rates, increasing refusal rates, the declining proportion of working residential numbers—and more. She challenged the research industry to reverse this trend, but there seems to be no good reason why all these alarming trends should not continue. As they do, telephone surveys will have less and less right to describe their methods as probability sampling. With the continuing decline of response rates to telephone surveys—not to mention the very low response rates of people who are asked to join audience measurement or mail panels—most research is already conducted with convenience samples, respondents who, unlike the majority, have given researchers permission to survey them. Like online surveys, telephone pollsters will increasingly be using convenience sampling.

Most American survey researchers regard quota sampling, one type of convenience sampling, as an unacceptable methodology proved unreliable by the Truman-Dewey election of 1948. The reality is more complicated. It was not quota sampling but the timing of the polls in 1948 (all of which were completed long before election day) which was the problem, as the polls failed to pick up the late surge in Truman's support. Never mind the facts, however. Quota sampling was seen as a culprit and was dumped by American pollsters who have used methods which they have described, and defended, as probability sampling ever since.

Not so in Europe, where the overwhelming majority of in-person surveys have been, and are still, conducted using quota sampling. In Europe, academics, as well as commercial, social scientists have used and defended quota sampling. Even if they have not done a great job of showing why or how quota sampling works so well, the record of the pollsters using quota sampling in Europe (particularly France and Germany) in hundreds of elections has convinced almost all skeptics there.

The relevance of the European experience is that well designed, well conducted quota sampling "works" very well for measuring public attitudes, voting behavior and voting intentions. If it works in Europe, why not here in the US?

While the first reaction of many critics may still be that only probability sampling is acceptable, we use and have come to accept research based on convenience sampling for several highly visible activities. These include the Nielsen TV ratings, mail panels, and The Conference Board's Consumer Confidence Index.

Statisticians, econometricians, epidemiologists, sociologists, and government agencies rely on convenience samples at times, particularly when it is difficult or impossible to employ random assignment to create comparable treatment and control groups as a precursor to estimating the effect of particular policies, practices, conditions or methods. The US General Accounting Office, for instance, employed a propensity score

approach a few years back (because random assignment was impossible) to compare mastectomy and breast conservation for the treatment of breast cancer.[5]

Whether we recognize it or not, we have come to accept that many well-designed, well-conducted surveys using convenience sampling "work." They meet the empirical tests if not the theoretical ones.

We believe that online research will be a huge part of the survey research industry's future. It is by far the fastest growing part of our commercial research business.

Will it be the appropriate methodology for all research? Of course not; we still use in-person surveys today for many purposes where they are better than telephone surveys. Do we know all we need to know about how to do good online research? Absolutely not; we recognize tough problems and major challenges related to sampling, weighting and method effects. But online researchers are learning fast and will keep on learning fast.

This is an unstoppable train, and it is accelerating. Those who don't get on board run the risk of being left far behind.

Endnotes

[1] Associated Press, March 3, 1999.

[2] Walter Lippman, "The Savannah Speech," in C. Rossiter and J. Lare (eds.), *The Essential Lippman* (New York: Random House, 1963). Originally published in 1933.

[3] Readers should also note that we published the results of the first two surveys on our website prior to Election Day, but we did not publish the later survey findings.

[4] P. R. Rosenbaum and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 1983, Vol. 70, No. 1, pp. 41-55. D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally & Co., 1963). W. G. Cochran and G. M. Cox, *Experimental Designs* (2nd ed.) (New York: John Wiley and Sons, 1957).

[5] US General Accounting Office, "Breast Conservation Versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and in Randomized Studies" (Washington, DC: USGAO, 1995).

See also:

Gordon Black and George Terhanian, "Using the Internet for Election Forecasting," www.pollingreport.com, 1998.

Humphrey Taylor, "Reading The Electorate: Internet Polls Were Shown To Work," *America at the Polls, 1998* (Storrs, CT: Roper Center for Public Opinion Research, 1999).

Humphrey Taylor is chairman, Louis Harris and Associates; George Terhanian is director of internet research, Harris Black International.