Back to the Future of Online Polling

No Witchcraft Here

By Humphrey Taylor and George Terhanian

"I cannot lay my hands on the memo I wrote years ago [as a graduate student]," Warren Mitofsky writes in the June/July *Public Perspective*, "but if I could, certainly it would not read much differently than my objections to the internet data collection being conducted today.... Forty years later, little has changed."

We share many of the concerns that Mitofsky raises about internet data collection in his essay. For instance, Mitofsky cites the infamous *Literary Digest* survey of 1936 as evidence that large sample sizes—one feature of internet surveys—will not, by themselves, eliminate sample biases. We agree. In fact, we often cite the *Literary Digest* poll to make that exact point. We part ways with Mitofsky, however, on the issue of weighting. We suspect, for instance, that thoughtfully weighted *Literary Digest* poll data would have produced an accurate forecast of the '36 presidential election (The *Digest* data were unweighted). Doubtless, Mitofsky would disagree with us, for we seem to subscribe to completely different philosophies on the issue of weighting.

Mitofsky argues, for instance, that no matter how one weights an online survey (i.e., a non-probability or convenience sample), one cannot correct for the biases which arise from the difference between the online sample and the population of interest. Forty years ago, we might have made a similar argument. However, the field of statistics has witnessed remarkable change since Mitofsky's days as an able graduate student. More recently, statisticians have developed sturdy statistical theories and techniques that explain how to eliminate or greatly reduce the biases associated with non-probability samples.[1] At Harris, we are indebted to these statisticians, for we depend routinely on their work.

We do so in the following manner. Each month, we run parallel telephone and online surveys to understand whether and how respondents and responses differ. After we complete the interviewing process, we statistically balance the observable characteristics of the telephone and online respondents by employing techniques, notably, "propensity score adjustment," that are designed to account and adjust for self-selection bias. The object of our efforts is to develop and refine a weighting

...continued next page

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

Miscalls Likely in 2000

By Warren J. Mitofsky

The arguments offered by Humphrey Taylor and George Terhanian for the Harris Black polling on the internet would certainly make a skeptic take note of the diligent work they have done. Indeed, the zeal with which they face the issues makes it seem reasonable for them to have bet what might be the future of their company on the success of this massive venture. However, this is a rejoinder, not a commendation.

Taylor and Terhanian have made the error in the public polls look more unreasonable than it really was. They measured the average error in the survey numbers by calculating the difference between each candidate and the vote. However, their average error calculation is misleading. Their calculation does not eliminate the "undecided" or "don't know/no answer" categories from the public polls before doing the arithmetic to calculate the polling error. The effect of this oversight is to compare apples and bananas. When the undecideds are allocated, the error in the public polls is 5.2 for the governor polls and 4.4 for the Senate polls. Without the allocation the errors in the public polls appear larger by 0.4 and 1.0 respectively.[1]

If we look at the error in the spread between the top two candidates, we find that only about one-third (37%) of the Harris Black internet polls had an error within the customary plus or minus three points on a candidate. A majority (51%) of the 1998 state telephone polls met this mark, and three-quarters (77%) of the 1996 state telephone polls did.

I would agree with Taylor and Terhanian that it is more important to name the right winner and that the size of the victory, while of consequence, is a secondary concern. For the record, about two-thirds of the contests on a typical election night are won by large margins. The worst outcome to bad estimates for landslide races is that one would think a contest was close when it was not. The crucial contests are the closer races. Large polling errors here can make a narrow margin look like a potential landslide or even a loss for the candidate who wins the election. Internet polling will cause havoc with these contests.

Of the 22 races Harris Black estimated, only seven were close enough to make the outcome less than a certainty. They

...continued next page

Warren J. Mitofsky is president, Mitofsky International.

routine that works on those occasions when we are unable to mount parallel telephone surveys. This is not witchcraft; it is painstaking, theory-driven work.

When we describe our approach to colleagues within the survey research community, very few maintain, as Mitofsky states repeatedly, that the techniques, as well as the theory underlying them, are hollow.

That durable theory would emerge in response to a complex problem (i.e., probability sampling is often impossible) should come as no surprise. Statisticians have long applied their skills in creative ways to solve complex problems. For instance, during World War II, America's top statisticians spent their days developing algorithms to improve the precision of America's bombers, according to the great statistician Fred Mosteller.[2] Thereafter, some of these same statisticians worked long hours to develop techniques to approximate the randomization of probability sampling, primarily because the desire to make fair comparisons seems universal (albeit difficult or impossible at times). At one time or another, for example, most of us attempt to compare self-selecting groups, whether they be smokers and non-smokers, patients who receive one surgical treatment and those who receive another, or high school dropouts and graduates. Unfortunately, some of us seem to draw neither on theory nor on common sense when we do so. This gives non-probability sampling its bad name.

An example of an improper and misleading comparison from Mitofsky is not hard to find. In his otherwise thoughtful essay, he compares the Harris Election '98 polls that were mounted in 14 states to telephone polls that were conducted in all states, not only in 1998 but in 1996 as well. This comparison is improper because it does not control sufficiently for state or year. It is misleading because it overestimates the comparative accuracy of the telephone polls. A proper comparison would not have supported Mitofsky's sweeping conclusion that "there is no valid purpose to the current internet enterprise."

In conclusion, we would like to stress that the design, and particularly the weighting, of online sample surveys is a work in progress. We still have much to learn about this research and are working feverishly with our panel of 4 million respondents to learn more. Only time will tell if our optimism about the future use of online surveys is justified—time and, of course, evidence that online surveys work as well as or better than telephone surveys.

Endnotes

[1]R.F. Boruch and G. Terhanian, "'So What?' The Implications of New Analytic Methods for Designing NCES Surveys," in Gary Hoachlander, Jeanne E. Griffith, and John H. Ralph (eds.), *From Data to Information: New Directions for the National Center for Education Statistics*, (Washington, DC: US Department of Education, National Center for Education Statistics, 1996), NCES 96-901.

[2]Personal communication, November 1995.

overestimated the victory margin in six of the seven by more than the margin of victory. The same size polling errors in the opposite direction would have given the wrong winners. Harris Black did not take much of a gamble predicting the winners for the remaining contests. Eleven were won by margins of from 18 to 40 percentage points. Even their biggest internet poll error (15 points) in these contests still made the winner look like a shoo-in. Of the other four safe races, the Harris Black poll result made two appear very close.

The conclusion I draw is that many miscalls are likely in 2000. If Harris Black takes on the closer races in 2000, given the size of the 1998 errors, I expect a lot of wrong calls. They did no really close races in 1998. The few that were moderately close had relatively large errors. And even if they get lucky in the next election and there are not many more wrong winners, their polls will likely mislead their audience about the spread in the final vote.

At AAPOR's 1999 annual conference I suggested to Terhanian that he and Taylor needed a theory under lying their use of the internet for conducting polls. Without a theory the enterprise is built more on wishful thinking than a solid foundation. In their *Public Perspective* article Taylor and Terhanian cite "propensity score adjustment" in an article by Rubin and Rosenbaum. What the article says is that it is possible to remove bias between units not assigned at random to different treatments in an experiment. The claim in the Harris/Terhanian article is that propensity score adjustment can be used to reduce the bias of a sample selected from among those internet users willing to sign up with them as panelists. I do not claim to be an expert on propensity scores. What I did get from the Rubin/Rosenbaum article is that a propensity score may achieve the effect Taylor and Terhanian are looking for. On the other hand, there is the distinct possibility that they will be increasing or overcorrecting the bias dramatically. The use of propensity scores makes a great many assumptions. It is not clear that the Harris Black internet polls satisfy these assumptions adequately.

Endnote

[1]Taylor and Terhanian refer to 52 state telephone polls for governor and senator in 1998. In the final edition of *Hotline* before the election there were 113 state polls. Of these, 49 were in the 22 states in which Harris Black conducted their internet polls. All calculations in this rejoinder are based on 113 polls, as Taylor and Terhanian do not identify the 52 telephone polls they evaluated. The internet polls used in my evaluation were the last ones publicly available before the election. If there are others, then this whole discussion may be at cross-purposes.