Pollsters.com

By Warren J. Mitofsky

When I was a graduate student and working part-time in the market research department at General Mills, I was asked to review the work of two survey research services we subscribed to. These services both kept in regular contact with huge recruited panels of people. The panels could be queried on a moment's notice to respond to the pressing problems of the day. Naively, I suggested that neither service was offering much in the way of useful information. The panels did not represent any larger group, I said, and therefore, there was no reliable way to make them representative of the populace at large.

My analysis was quickly ignored. There were many thousands of people in those panels and they produced cheap samples that were very large, far in excess of anything that could be obtained from other ways of selecting respondents. And yes, Virginia, to these people size mattered. "But what about the *Literary Digest*?" I asked. "They interviewed millions. "That's different," was the reply. To this day, I have never figured out how they differed.

Forty years later, little has changed. Instead of sampling from recruited panels that respond by mail, we now have recruited internet panels that respond by e-mail. I cannot lay my hands on that memo I wrote years ago, but if I could, certainly it would not read much differently than my objections to the internet data collection being conducted today. Online data collection is worse than sampling telephone households in countries that have medium telephone penetration. The people recruited into panels are self-selected, with characteristics that differ from the target population. People on the internet do not represent the adult population of the country, and the internet panelists do not even represent people on the internet. At best, we end up with a large sample representing nothing but itself.

Of course, the people who do these polls know all this, but they believe they have discovered a magic formula for converting their ersatz sample into the real thing. If we can just find the right variables to use in weighting our respondents, their argument goes, we will have useful data. For political polls, the weighting variables have been demographics, as measured by the Census Bureau, and party vote, from probability surveys of voters.

Weighting will improve the reliability of data if the variables used for the adjustment are correlated with the variables being measured. Conversely, if they are not highly correlated, the weighting will make the reliability worse.[1] Because sample surveys usually measure many characteristics, weighting the results improves some estimates, but it can be detrimental to others. One hopes that the key variables are improved while the others are not made too bad.

The assumption made for the internet political polls is that the men and women who respond to the online surveys vote the same way as the men and women who are not in the panel. The same assumption applies to any other variables used to adjust the internet queries. This assumption, however, is flawed. Even if weighting is effective for one set of variables such as voting, there is no way to know that it will hold up for other variables in non-voting studies. Commercial clients who buy this type of data collection for their market research will be severely misled if they think they are getting the same success they might get from the voting studies. The weighting scheme, if it works for voting studies, will not necessarily work for anything else. (So far, the weighting schemes for voting studies have been at best fair. More about that later.)

So what's wrong with the weighting assumptions being used for the internet polls? Imagine an internet sample that has exactly the right mix of men and women, old and young, Republicans and Democrats, and so forth. When weights are applied all survey respondents are counted equally. But the voting results of the internet data may still be at odds with the true population values if the men and women, old and young, Republicans and Democrats not in the internet panel vote differently than those in the panel. It is not just the relationship between men and their vote for a candidate that is important. It is the unknown correlation between the vote of men in the panel and the vote of men not in the panel that is key to using weighting to improve an estimate. For a demographic weighting variable to improve the data it must simultaneously adjust the relative size of the demographic group *and* the missing vote of this group. In order to estimate the missing correlation one would have to conduct a survey or a census of all adults. Then one could learn by how much each demographic group needed to be adjusted. Of course, this defeats the whole purpose of using the internet for conducting the survey in the first place.

> "What bothers me is the willingness to discard the use of sampling frames as a means of selecting a sample and then the feeble attempts at manipulating the resulting bias. That undermines the credibility of the survey process."

Sometimes the internet pollsters run tests where they conduct both an online poll and a probability-based poll in order to identify potential relationships. In turn, these relationships become the basis for adjusting future internet polls. The idea here is that once a relationship is discovered it will continue to exist. But there is no guarantee that a relationship, if discovered, will hold constant. This idea is the empirical justification for doing internet polls, and is the entire basis for the claim of having used a scientific approach. It might work, and then again it might not. One will never know for certain without a means of evaluation each time an internet poll is conducted.

When I suggest these things to people who tout the internet surveys as the wave of the future, their reply has something to do with the low response rates currently achieved by many commercial probability samples.[2] They ask a good question: What is the difference between a probability sample with a low response rate and an internet survey? For starters, the probability sample started with a fair representation of the entire population of interest. There is the possibility under these circumstances that the nonresponders are a random subset of the original sample. And if it is not a random subset, then it is a subset with only a small bias. We have evidence that this is the case with most exit polls. It also seems to hold up for pre-election vote projections conducted close to Election Day. If this were not the case, the pre-election surveys, with their poor response rates, would not regularly come close to the election result.

This would seem to make the likelihood of good estimates from the internet much more remote than estimates from probability surveys with low response rates. The internet surveys simply cannot offer the same possibility. The panel used as a frame to sample respondents is not a probability sample of internet users. Even if it were, internet users are not like non-internet users. For instance, they are better educated, more affluent, and more male. The important characteristics for weighting have not yet been identified. Generalizations from a census of the growing millions in the panel do not even represent internet users. The whole panel represents nothing but itself.

Of course the weakness here is an absence of theory. Without a theory we will find it impossible to generalize from the results of our experiences with low response rate surveys to biased internet data collection. Repeated trials in which we can compare surveys with low response rates and internet queries to election results may give us solace, but they offer little else that is meaningful.

In 1998, Harris Black International (HBI) conducted internet polls in 22 US Senate and gubernatorial races. Gordon Black, chairman of the firm, took bows for his company's performance in these polls in the *Wall Street Journal*.[3] He claimed that "all research is going to migrate to the Internet," that telephone polling as we know it was all but dead, and that anyone who did not agree with him was a dinosaur. We should all be persuaded, he said, because 21 of the 22 polls had the correct winner. He called these internet polls a "scientific revolution" and people like me "defenders of the old paradigm."

Before we all trash our CATI systems it would be well to look at the errors in these polls as compared to other state polls conducted by more traditional methods. Results for all these state polls were taken from the final issues of *Hotline* published prior to the 1996 and 1998 elections.[4]

The HBI online polls were based on samples HBI designated from their database of 3,000,000 names. The state polls were based on a variety of methods, some using random samples and others more dubious methods.[5] Nonetheless, they represented what the public had been getting in their newspapers and on television in recent years.

The performance of the internet polls was fair. While there was only one miscall, one-third of the final online polls (7 of the 22) had an error that was larger than the margin of victory. It would not be surprising under these circumstances for the online polls to have made more miscalls. In comparison, the state polls in 1996 were very good. There was only 1 mistake in the 104 final state polls, and most polls had fairly accurate projections of the victory margins. The 1998 state polls were not nearly as good as those conducted in 1996 and 1992.[6] In fact, they were pretty bad. They had the wrong winner in 13 contests.[7] The 1998 state polls were, however, closer to the mark more often than the HBI internet polls. There were 49 final surveys in the 22 states polled by HBI. Twenty-nine of the state polls were more accurate than HBI. HBI did better than 15 of the state polls. The remaining five had the same error. Table 1 shows that only 37% of the 1998 internet polls had projections estimating the

Table 1

Error in estimating the margin between the first and second candidate (in percentage points)	**Harris Black Internet Polls, 1998**	**State Polls, 1998**	**State Polls, 1996**
≥13%	14%	14%	0%
7-12%	50	35	23
<6%	37	51	77
Number of surveys examined	*(22)*	*(113)*	*(104)*

Note: For 1998, polls estimated Senate and gubernatorial races. For 1996, polls estimated presidential and Senate races.

size of the frontrunner's lead within 6 percentage points of the actual margin. Fifty-one percent of the 1998 state polls and 77% of the 1996 state polls had projections within this range.[8]

The direction of the errors on the internet and state polls is noteworthy. More often, the state polls had errors overstating the margin in races where Democrats won and understated the margin in Republican victories. The HBI polls were just the opposite. They understated Democratic victories and overstated the size of Republican wins. This trend probably reflects the political tilt of people who use the internet today and signed up for the HBI panel. There is no guarantee that this tilt to the panel will continue in the future as more people have internet access. As more people gain internet access new and unpredictable tilts are likely to occur.

The hope of internet pollsters is that in the not too distant future as many people will be using the internet as now use the telephone. However, this will not solve their problem. There still will not be a viable sampling frame available. As things now stand, even if 100% of the population were on the internet there is no way to select a random sample of e-mail addresses. Internet users cannot be sampled directly. HBI samples members of their panel, which is all they ever will be able to do until someone develops either a complete list of e-mail addresses or a scheme for sampling them. As long as samples come from a panel, they will continue to have all the same problems cited above.

It would give me great pleasure to come up with a sound way to use the internet to conduct meaningful surveys. To me, that means a probability sample of respondents would be necessary. It does not mean a self-selected sample, massaged and manipulated to represent something it is not. I thought of sampling people who visit a website, but that sample would only represent users of that site. Another approach would be to use the internet as part of a double-sampling approach. For example, one could select a smaller than usual probability sample of telephone households and a sample of internet users. The two samples could be unduplicated and combined using varying probabilities of selection. The problem is that at this time it is not possible to directly select a probability sample of internet users. A large sample of people could be selected by sampling households, and then interviewing could proceed either on the internet, by e-mail, or by telephone. I have heard of one survey that is proceeding in this way, except they are placing computers in the homes of those who do not have them. The sample in this case could form a panel for future surveys with all the problems and benefits inherent in panels.

There is nothing about my suggestions that would be as inexpensive as drawing a sample from a panel of e-mail addresses. For many people *cheap* is reason enough to justify data of dubious relevance to whatever question they are trying to answer. Truly, it blackens the eye of survey research that established professionals are touting internet surveys without having a theoretical basis for their enterprise. It is not the use of the internet for interviewing that bothers me—that change will come. It is the willingness to discard the use of sampling frames as a means of selecting a sample and the feeble attempts at manipulating the resulting bias that undermine the credibility of the survey process.

I do not want to abandon tested and proven survey methods that are based on a solid foundation of probability theory until we have a new theory-based methodology for doing so. I can see no valid survey purpose to the current internet enterprise. All that will happen will be the accumulation of thousands upon thousands of interviews of dubious merit that will mislead the public and destroy whatever credibility surveys and polls now have. A growing number of survey researchers are unfortunately being led to the rocks like Ulysses' sailors following the Siren call of cheap, but worthless, data.

Endnotes

[1]"Highly correlated" for a ratio estimate means correlations starting at +0.5 will reduce the sampling error of a probability sample over what one would get from a simple unbiased estimate. The higher the correlation, the smaller the sampling error.

[2]An internet pollster suggested to me that response rates for RDD surveys are currently in the 20% range. This seems self-serving, as it is much lower than any reported rates I have heard elsewhere.

[3]John Simons, "Are Political Polls Via Internet Reliable? Yes? No? Maybe?" *Wall Street Journal*, April 13, 1999, p. 1.

[4]*Hotline* is a five-day-per-week newsletter read by journalists and political workers. It excerpts newspaper and television national political news and publishes the results of political polls as they become available.

[5]For a discussion of methods used by state polls in 1992, see Warren J. Mitofsky, "The State of State Election Polls," *Chance*, 1992, 6:1, pp. 9-16.

[6]Mitofsky, "The State of State Election Polls." Though not evaluated in the same way as 1996 and 1998, the 1992 presidential polls were fairly accurate.

[7]The largest errors in calculating the margin between the first and second candidate in the state polls were: OR Sen. -29 points, VT Gov. -25 points, PA Gov. -24 points, OR Gov. -21 points. None of these polls had the wrong winner. The worst of the HBI internet polls were: PA Gov. -15 points, IL Sen, -15 points, which had the correct winners, and the GA Gov. -15 points, which had the wrong winner.

[8]It is worth examining the performance of the state polls to learn if the difference noted here between 1996 and 1998 is a meaningful decline in performance or a presidential year/off-election year phenomenon.

Warren J. Mitofsky is president, Mitofsky International.