

## Does Internet Research “Work”? Comparing Online Survey Results With Telephone Surveys

Written By:

**Humphrey Taylor**

Chairman, *The Harris Poll*®

### The Fastest Growing Technology in History

Andy Grove, the founder of Intel, says that “in five years’ time, all companies will be Internet companies or they won’t be companies at all.” *The Economist* magazine recently wrote that “The Internet (will) change everything – the way we work, the way we learn and play, even, maybe, the way we sleep or have sex. What is more, it is doing so at far greater speed than the other great disruptive technologies of the 20th century, such as electricity, the telephone and the car.” I will leave others to describe how it will change the way we have sex. But even a modest assessment of how it will change our industry would lead to the conclusion that in a few years time, most successful market and opinion research firms will be conducting much of their research online.

It is a cliché but it is true; the Internet is, by far, the fastest growing technology in the history of the world. At the beginning of 1995, only 7% of all adults in the U.S.A. were online whether from the office, from home, school, library or somewhere else. Less than five years later, that number has risen to almost 50% (49% by our last measure) and looks set to reach about 60% by the end of the year 2000. Who knows when it will stop growing. While much data on Internet use vary greatly and are probably unreliable (and out of date), some reports indicate that Sweden, Denmark and Finland have higher Internet penetration than the United States.

### Does it Work? Is it Reliable?

While online research is the hottest, fastest-growing and most revolutionary development to change our industry since George Gallup demonstrated in 1936 that “scientific” opinion polls were the best way to predict elections, its success will stand or fall by the credibility of the data it generates.

Many academics and statisticians believe that any survey which is not based on probability sampling is unacceptable. They argue that even if something “works” we should not embrace it unless there is a sound statistical/mathematical

theory to explain how and why it works. They forget perhaps that Newton had no theory to explain how gravity worked; he just showed that it did. Medical scientists tell us we know very little about how and why many pharmaceutical products work; we accept and use them because clinical trials show us that they work, even if we do not know why.

Of course, it is dangerous to generalize about the views of academics and statisticians, and they often disagree. There is a deep and fundamental disagreement between the great majority of North American research practitioners and their counterparts in Europe and elsewhere, on the validity of quota sampling for in-person surveys. Never mind that quota sampling has been shown “to work,” and sometimes to work better (for example in predicting election results), than probability sampling. Never mind that it is widely (but not universally) accepted and used by many European academics, statisticians and some government agencies. There is no good theory to explain how and why “it works” and so, to statistical purists (particularly American ones), it is unacceptable.

In my experience, the users and readers of market and opinion survey data divide into three groups:

- The largest (sadly) group for whom “a poll, is a poll, is a poll” regardless of who did it, who paid for it, or how it was done.
- A very small, but very influential, group for whom the only acceptable surveys are those based on pure probability sampling, which achieve an 80% response rate. Unfortunately, that means not accepting 99% of all surveys. [This is not pedantry; a 60% response rate (not to mention much more common 30% to 50% response rates) means that a survey is a “convenience sample” of available “volunteers” whose behaviors, attitudes and motivations, even after demographic weighting, may differ substantially from those of non-respondents and, therefore, the total population.]

- The third group, who are “empirical pragmatists” or “pragmatic empiricists,” who will accept a research methodology (for example, quota sampling) if it consistently produces credible data and which can be shown “to work.” These people determine what the industry will and will not accept.

The future of online polling therefore will depend on the industry’s ability to demonstrate that it too works. That means using online surveys to predict elections and to produce data which are very similar to the data produced by other credible sources, such as census data and the results of other widely accepted surveys.

**Online Research is Fundamentally Different**

The use of the Internet to conduct marketing and opinion research is a much more revolutionary development than the other, more modest, changes I have witnessed in my 36 years in the industry. In the 70s, we began to use the telephone for data collection. In the 80s, we started using CATI and CAPI systems. While these were major advances, they did not fundamentally change our thinking about how to collect and analyze data. They did not fundamentally change the way we designed questionnaires. And while they involved investments in hardware, software and people, most of the cost of data collection was related to the number and length of the interviews. With online research the largest data collection costs are capital costs – setting up the system to do it – not the cost of paying interviewers or telephone companies.

Those who wish to conduct or commission Internet surveys need to recognize that collecting data online is different in several other important ways:

- It is *not based on probability sampling* – but on “volunteer” sampling or “convenience” sampling.
- It is a *visual medium* – allowing respondents to see images, longer text messages, longer lists of response options and, as bandwidth grows, video images.
- It *captures the unedited voice of the respondent* – we have found replies to open-ended questions to be richer, longer and more revealing.
- It *may be more effective in addressing sensitive issues* – adults may be more willing to reveal information about their experiences with sensitive conditions (e.g., anxiety disorders, ovarian cancer, incontinence, erectile dysfunction, depression).

- *Scales may elicit different response patterns* – our experience is that fewer people pick the extremes on scales when they see (as opposed to hear) them.
- *Online surveys may generate more “don’t knows” or “not sures”* – because respondents can see these options – but follow-up questions can be used to reduce this.

And of course,

- *Raw online data substantially underrepresent some groups* – hence the greater importance of weighting than with good in-person or telephone surveys.

**No Listing of Email Addresses**

The first sampling issue to confront is the unavailability of a comprehensive list of email addresses of the Internet population. In most countries, 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 email addresses. Even if such a listing existed, sending emails to potential respondents of online surveys who have not agreed beforehand to participate in online research would look like “spamming” and be unacceptable to many people.

**Weighting – The Big Issue and the Biggest Challenge**

It is still true, albeit less so with each passing month, that the online population is substantially younger and better educated than the public as a whole. It is particularly short of people who did not finish school, people with lower incomes and people over 70. However, substantial numbers (if low percentages) even of these people are online. So, with adequate sample sizes, online data can be weighted to be representative by demographic variables such as age, sex, region, education, occupation, race and ethnicity.

**TABLE 1**  
**Profile of the Online Population in the U.S.A.: As it Looks More Like the Total Population**

	1995	1997	Autumn 1999	Census Data
	%	%	%	%
Black	1	8	10	12
Hispanic	9	8	10	10
Women	21	46	48	52
College Educated	57	38	34	22
Under \$25K Household	15	18	22	27
Aged Over 65	3	3	5	16

Source: Harris Interactive (telephone surveys).

However, experience has shown that raw online survey data have more than just demographic biases. Much work still needs to be done to identify and understand these biases and to determine the most reliable weighting variables to correct for them. That is probably the biggest issue and the biggest challenge. One way to address this challenge is to conduct parallel online and traditional surveys using identical questions.

### “Propensity Weighting”

It is no surprise that certain kinds of people have a greater or lesser propensity to be online and to reply to our surveys. We have adopted the phrase “propensity weighting” to describe the weights used to compensate for biases in online samples that are not adequately corrected by demographic weighting. The phrase is not original; it comes from educational research designed to analyze and explain the impact of different teaching methods (or class size) on academic achievement.

It is similar in concept to the weighting schemes used in epidemiology where non-random populations (e.g. smokers and non-smokers, or patients in different hospitals) are compared. It is relevant wherever different populations, which are not randomly assigned, are being compared. Sometimes this is called “case-mix adjustment.” If we only accepted probability sampling, the only way to determine the health impact of smoking would be to allocate people randomly to be smokers and non-smokers, something even Dr. Mengele did not attempt. Absent random selection, we resort to propensity weighting.

Because of the many biases in raw online survey data, both demographics and propensity weighting are required to correct for the biases we find.

Propensity weighting is a work in progress. The propensity weights we have developed to date are mostly drawn from our parallel telephone and online surveys, based on the assumption that some large differences reflect biases in the online sample because of the greater or lesser propensity of different people to be in our online surveys. Over the next year we expect – based on more comparisons of online and telephone survey data – to settle on five or six standard weighting criteria including attitudinal, behavioral and descriptive variables. So far, variables we have used which seem to work well (i.e. bring online survey data close to telephone survey data) include measures of health status, political party identification and numbers of telephone lines. Several others look very promising but need further testing.

### Weighting Cannot Solve All Problems of Bias

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

Telephone surveys miss, among other people, those who live in homes without telephones, people who are away from home 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. 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. Most telephone and in-person surveys under-represent people who travel a lot and eat out a lot – because they are not at home to be interviewed. Obviously, if you want to do a survey to find out why people do not have computers or are not online, you cannot do that online.

### There Are Many Very Different Types of Online Surveys

There are some dangers in making generalized statements about online research. Online survey methodologies differ greatly. Just as there are many types of in-person and telephone surveys, so there are many types of online surveys, some of which make no claim to be producing representative samples or projectable data. The methodology we have chosen is based on a very large database (not exactly a panel) of people who agree to be surveyed and who give us their email addresses.

At this time, we have some five million such people in our database and we plan to add many more millions, in the United States, in Europe, Japan and elsewhere. For each survey, we invite a sample of people in this database to go to a unique, password-protected website to complete a questionnaire which can include all of the skip patterns, rotations, split samples, etc. that can be administered with a CATI or CAPI program. Some surveys offer financial incentives; others do not.

To be able to conduct many such surveys, to send out the emails, to draw the relevant samples and to continually update the database require a substantial investment in

hardware, software and people. That, not the data collection, is the main cost of online data collection.

**An Enhancing, Enabling and Transforming Technology**

Online research has already been described both enthusiastically, and critically, in many different ways. “A replacement technology” is one such phrase and it will surely replace much of the qualitative and quantitative research work currently done face-to-face or by telephone. However, it will not fully replace other methodologies. Printing did not fully replace handwriting. Radio did not replace newspapers. Television did not fully replace the movies or the radio. We will continue to do in-person and telephone research even if this amounts to a rapidly shrinking part of our work.

However, the Internet enables us to do many things we could not do (or afford to do) before, and greatly enhances the value of our services.

Very specifically, the Internet enables us to:

- Survey **huge** samples of people
- Survey **tiny subsamples** of the population
- Do everything we can do on **CAPI/CATI**
- **Show** lists, still and moving images
- Obtain much richer verbatim replies

AND to do this

- Incredibly **fast**
- At **affordable** costs

For one survey of 103,000 people for our eCommerce Pulse<sup>SM</sup> service, we completed all the interviewing over 10 days. For a survey of 10,000 patients with specific medical conditions, the data collection took one week. Only the Internet enabled us to do this at an affordable cost.

**Online Qualitative Research**

Just as exciting as the prospect of conducting sample survey research online is the ability to conduct both traditional and brand new types of qualitative research on the Internet. These are early, if heady, days and we have surely only scratched the surface of what will be possible. Some initial efforts have been focused on replicating traditional focus group interviews with real-time online focus groups. Others have involved special chat rooms to hold what are, in effect, continuing conversations with individuals and groups who could not easily be brought together at one time or one place.

Because the Internet is a visual medium, it can be used (quantitatively and qualitatively) for concept, advertising and package testing. As bandwidth and computing power grow, it will become an increasingly effective way to test TV commercials and other video productions.

**Forecasting Elections 1998**

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 U.S. election outcomes for 23 races for governor and the U.S. Senate in 14 states.

So how did we do? In our final published (October 29-31) surveys, we correctly projected the winner in 21 of 22 (95%) races. 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.

Actually, our final surveys did rather better – but unfortunately we did not release them until after the election, so we cannot claim them as evidence.

**TABLE 2**  
**Comparisons of Telephone and Online Predictions of 1998 U.S. Elections in 23 Races**

<i>1998 election forecasting performance indicators</i>	<b>Final Online Published Polls</b>	<b>Final Online Unpublished Polls</b>	<b>Final Telephone Surveys</b>
Percent correct winner	95%	100%	94%
Average spread error	6.8	6.3	6.2
Average candidate error	4.4	4.0	5.0

Note: Online data weighed demographically; no propensity weighting.  
Source: Harris Interactive (October 1998).

Although the accuracy of our forecasts surprised some people – 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.

**Comparing Parallel Telephone and Online Surveys of Political Attitudes and Behavior**

In order to learn more about biases in our online surveys and how to weight to correct them, we have conducted many dozen parallel telephone and online surveys using the same questionnaires and conducting the fieldwork at the same time. These are all nationwide U.S. surveys.

Tables 3, 4 and 5 show some of the similarities and differences between the telephone data (weighted demographically) and online data (weighted demographically, with and without some propensity weights). In general, the use of three modest propensity weights succeeds in bringing the online closer to the telephone survey data. However, the results show some interesting differences and suggest that additional weights need to be developed.

In Table 3, the use of demographic weighting alone was sufficient to bring almost all of the online variables very close to the replies in the parallel telephone survey. On these variables, additional propensity weights made virtually no difference.

In Table 4, several of the online results were substantially different from the telephone survey results after only demographic weighting. For most of these variables, the propensity weights succeeded in substantially reducing these gaps.

**TABLE 3  
Comparison of Parallel Telephone and Online Political Survey Data With and Without Propensity Weighting**

	Telephone Survey	ONLINE SURVEY	
		Demo-graphic Weighting	Demo-graphic and Propensity Weighting
<b>Base</b>	<b>1009</b>	<b>13946</b>	<b>12543</b>
	%	%	%
Presidential Approval	58	56	59
Congressional Democratic Approval	43	40	40
Congressional Republican Approval	37	35	35
Trent Lott Approval	29	28	28
Dennis Hastert Approval	27	27	27
Al Gore Approval	39	38	40
Party ID – Democrat	34	37	37
Republican Primary – Bush’s support	58	52	51
Democratic Primary – Gore’s support	62	62	62
Presidential Election (G&B) – Bush’s support	59	61	58

Note: Telephone data were demographically weighted. The sample size after propensity weighting is slightly smaller because not all respondents had provided the relevant variables. Source: Harris Interactive (September 1999).

In Table 5, the use of propensity weighting did not do much to improve on the demographic weighting. Specifically, the online sample is more likely to vote than the telephone sample. This suggests the need for an additional propensity weight.

**TABLE 4  
Comparison of Telephone and Online Political and Other Survey Data With and Without Propensity Weighting**

	Telephone Survey	ONLINE SURVEY	
		Demo-graphic Weighting	Demo-graphic and Propensity Weighting
<b>Base</b>	<b>1006</b>	<b>12868</b>	<b>12864</b>
	%	%	%
Called, Written or Visited Elected Official	32	48	38
Written a Letter to Newspaper, Magazine, TV Station	16	23	18
Called into a Talk Show to Express Opinion	10	12	11
Attended a Meeting Where Politician or Elected Official Spoke	37	37	35
Worked on Political Campaign	10	12	11
Display Campaign Paraphernalia	35	37	36
Believe News Contributes to Violence	39	39	40
Believe Video Games Contribute to Violence	47	43	45
Believe Television Contributes to Violence	58	51	54
Believe Movies Contribute to Violence	57	53	57
Believe Lack of Supervision Contributes to Violence	90	91	92
Believe Easy Availability of Handguns Contributes to Violence	65	57	60

Note: Telephone data were demographically weighted. The sample size after propensity weighting is slightly smaller because not all respondents had provided the relevant variables. Source: Harris Interactive (June 1999).

**Comparing Parallel Telephone and Online Health Care Surveys**

Tables 6 and 7 show some of the similarities and differences between the telephone data and online data (weighted demographically and with propensity weights) for two health care surveys.

**TABLE 5**  
**Comparison of Telephone and Online Survey Data With or Without Propensity Weighting**

	Telephone Survey	ONLINE		Demographic plus, Propensity Weighted
		Un-weighted	Demographic Weighted	
<b>Base</b>	<b>1009</b>	<b>13946</b>	<b>13946</b>	<b>12543</b>
	%	%	%	%
<b>State of the Nation</b>				
Excellent	9	7	6	7
Pretty good	46	53	50	50
Only fair	31	31	32	33
Poor	13	8	10	10
Not sure/Don't know	1	*	1	1
<b>Voting</b>				
Absolutely certain to vote	61	77	68	68
Claim to have voted in 1996	65	82	72	72

Note: The sample size after propensity weighting is slightly smaller because not all respondents had provided the relevant variables.  
 Source: Harris Interactive (September 1999).

In Table 6, almost all of the replies are extremely close. The one big difference is between the 75% and 57% who were “very satisfied” with their counseling. We believe this is a **method effect**, not a sampling difference because:

- The total numbers who were “very” or “somewhat” satisfied were very close (i.e. the online survey had many more “somewhat” satisfied).
- We have seen other similar examples of respondents to online surveys avoiding the extremes (e.g. “very satisfied” or “very unsatisfied”).

**TABLE 6**  
**Comparison of Parallel Telephone and Online Health Care Data (Propensity Weighted) Replies of Women With a Specific Medical Condition (U.S.A.)**

	Telephone	Online
<b>Base</b>	<b>272</b>	<b>664</b>
	%	%
Have been prescribed treatment	33	33
Filled prescription	97	94
On treatment less than 1 year	22	23
Initiated discussion with doctor	33	28
Discussed side effects with doctor	75	67
Discussed benefits with doctor	87	82
Very satisfied with counseling	75	57
Discontinued treatment	25	24

Source: Harris Interactive (July 1999).

**TABLE 7**  
**Comparison of Parallel Telephone and Online Health Care Data (Propensity Weighted) Demographics Profile of People With Seasonal Allergies**

	Telephone	Online
<b>Base</b>	<b>331</b>	<b>4749</b>
	%	%
<b>Age</b>		
18-24	13	14
25-39	32	32
40-49	23	20
50-64	17	17
65+	14	12
<b>Sex</b>		
Men	39	36
Women	61	63
<b>Race</b>		
White	68	76
Black	13	11
Hispanic	13	7
<b>Education</b>		
High School or Less	53	50
Some College	28	28
College Graduate +	18	20
<b>Income</b>		
\$25K or less	39	37
\$25K - \$50K	30	28
\$50K+	32	29
<b>Health Status</b>		
Positive	80	82
Negative	20	18

Source: Harris Interactive (August 1999).

Table 7 compares the demographic profiles of people with seasonal allergies. Both samples shown here were subsamples of much larger surveys. (The total online samples was of 10,000 patients with chronic medical conditions weighted with both demographic and propensity weights to match the existing, telephone survey-based, data on such people.) The similarities are striking. The only significant difference is on race/ethnicity. Although both surveys were weighted by race/ethnicity, the online subsample with seasonal allergies includes relatively fewer blacks and Hispanics.

**The Future**

Because of the speed of the Internet revolution, new and exciting research applications of online research are appearing every week. We frequently stumble on new ideas, new and better ways of doing things, as well as things nobody

thought of doing previously. Moore’s law (Moore, of course, was the other founder of Intel) is that computing speed and power double every 18 months. Internet traffic is reported to be doubling every six months. Our knowledge and understanding of how to use the Internet to conduct both qualitative and quantitative research is probably growing (from a near zero base!) even faster. If there is one certainty, it is that we ain’t seen nothing yet.

### References:

- 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).
- US General Accounting Office, “Breast Conservation Versus Mastectomy: Patient Survival in Day-to-Day Medical Practice and in Randomized Studies” (Washington, DC: USGAO, 1955).
- Humphrey Taylor, “Reading the Electorate: Internet Polls Were Shown to Work,” *America at the Polls*, 1998 (Storrs, CT: Roper Center for Public Opinion Research).
- George Terhanian and Humphrey Taylor, “Heady Days Are Here Again,” *Public Perspective*, June/July 1999 (Storrs, CT: Roper Center for Public Opinion Research).