

# Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research

**By** George Terhanian, Ph.D., Vice President, Research and Methodology  
John Bremer, Research Scientist

**For** The Advertising Research Foundation

**Status** Draft #10 (Not for citation without the permission of the authors)



## INTRODUCTION

During the 1950's, many prominent statisticians dismissed on methodological grounds the findings of the controversial Kinsey Report on Sexual Behavior in the Human Male (Kinsey, Pomeroy, & Martin, 1948). Of particular concern was the researchers' use of non-probability sampling to recruit and interview a sample of white, adult males through whom they attempted to make inferences about all members of this population. The notion that non-probability sampling could produce projectable information offended the sensibilities of many of the report's critics. Prompted by this criticism, the National Research Council asked the American Statistical Association to appoint a blue-ribbon committee to evaluate the Kinsey Report's methodology. The committee included three of the world's most able statisticians: William G. Cochran, John W. Tukey, and Frederick Mosteller. Unlike some of their colleagues from the statistical community, however, Cochran, Tukey, and Mosteller did not categorically dismiss the Kinsey Report's findings. As they asserted: "he did not use a probability sample' is...not a criticism which should end further discussion...(p. 328)."

In fact, Cochran, Mosteller, and Tukey's words are as appropriate today as in 1954—most organizations that currently conduct Internet research do not use probability sampling<sup>1</sup>. Instead, they depend on the cooperation of respondents who have elected to participate in Internet research in a variety of ways, none of which typically involve probability sampling. Cochran, Tukey and Mosteller's advice notwithstanding, many contemporary researchers categorically dismiss the notion that non-prob-

ability-based Internet surveys can produce trustworthy information. Further, these researchers tend to chastise organizations that employ non-probability sampling, reminding them of the failures of the Gallup, Crossley, and Roper organizations<sup>2</sup> to predict Harry S. Truman's victory over Thomas Dewey in the 1948 presidential election (Mitofsky, 1999; Rivers, 2000).

The critics of Internet research do not generally mention, however, that the Gallup, Crossley, and Roper final pre-election forecasts were made on the basis of interviews completed at least two weeks, and at most, two months, prior to the election. Nor do they generally mention that the pre-election polls that employed probability sampling in 1948 were no more accurate than those that used other methods. As Cochran, Tukey and Mosteller (1954) observed, "...the behavior of the few probability samples in the 1948 election does not make it clear that opinion pollers should spend their limited resources on probability samples for the best results (p. 328)."

The aim of this paper is to explain how and why it is possible to produce trustworthy information through Internet research; specifically, through Internet research that relies upon the cooperation of respondents recruited by means other than probability sampling.

## THE IMPORTANCE OF THE LIST

Telephone research begins with a more or less comprehensive list of all residential phone numbers. Through the deployment of a random-digit-dialing procedure, each

1 InterSurvey is the lone exception. They are recruiting a panel of Internet respondents through probability sampling.

2 The Gallup, Crossley, and Roper organizations relied on non-probability sampling to make their forecasts.

household, and each adult individual within the household, has some probability of being contacted. In theory, the respondents eventually completing the survey constitute a representative sample of the entire population. In practice, however, the random-digit-dial procedure does not guarantee that the views of the target population will be accurately represented. Unlike agricultural research in which “plots of ground can[not] excuse themselves from being treated” (Heckman, 1992, p. 215), human beings can easily excuse themselves from participating in telephone research. And they now do so at discomfiting rates. To make matters worse, the bias associated with refusals is likely to increase as time passes, not only because of the negative impact of telemarketing, but also because of technological advances such as answering machines, caller identification, and cellular phones.

Internet research also begins with a list, but there is no equivalent comprehensive list of all email addresses. Even if there were such a list, it would not be possible to draw random samples from the list, as in telephone research. It is unethical and, in states such as Washington, illegal to deliver unwanted, unsolicited email for commercial, marketing or research purposes. As a result, Internet research organizations must depend on lists they build themselves. Harris Interactive, for example, depends on a list of 6.6 million respondents, who have joined the panel through at least twenty-one different sources, all of which are listed at: <http://vr.harrispollonline.com/register/became.asp>.

#### ADJUSTING FOR SELF-SELECTION BIAS AND LEARNING EFFECTS

To produce projectable information, research organizations that rely on lists built by means other than probability sampling must employ thoughtful sampling and weighting approaches to compensate for the differences between the list and the population of interest (e.g., the U.S adult population). Some (Rivers, 2000) argue, however, that no matter how thoughtfully a sample is drawn from such a list or how well the results of a survey are statistically adjusted, one cannot correct for the biases that arise from the difference between the sample and the population of interest. Others (Mitofsky, 1999) argue that all research that depends on the repeated participation of pre-selected respondents (i.e., a panel) is fundamentally flawed on

account of the learning that takes place through participation in multiple surveys, irrespective of how the respondents are initially recruited.

Fifty years ago, we might have made similar arguments. In the interim, however, the field of statistics has witnessed remarkable change, and statisticians have developed sturdy techniques to eliminate or greatly reduce the biases associated with (1) non-probability samples and (2) panels of pre-recruited cooperative respondents who participate in multiple surveys. The techniques of paramount interest are those that approximate the randomization of probability sampling, a method that can be difficult or even impossible to employ.

#### A STURDY SUBSTITUTE FOR PROBABILITY SAMPLING

Cochran, Tukey, and Mosteller recognized the need and potential for a sturdy substitute for full-blown probability sampling in 1954 when they reviewed the Kinsey Report. Rather than “dooming” the report “to the cellar (p. 328),” they made the following recommendation:

“Since it would not have been feasible for KPM to take a large sample on a probability basis, a reasonable probability sample would be, and would have been, a small one and its purpose would be: (1) to act as a check on the large sample, and (2) possibly to serve as a basis for adjusting the results of the large sample (p. 23).”

This is precisely the approach that Harris Interactive has taken.

#### THE HARRIS INTERACTIVE APPROACH

Propensity score adjustment (Rosenbaum & Rubin, 1983) is the sturdy substitute for randomization upon which we rely to adjust for self-selection bias. It is also the technique upon which we rely to reduce or eliminate the potential learning effects associated with participation in multiple Harris Poll Online surveys. We do so in the following manner.

Each month, we run parallel telephone and Internet

surveys wherein we compare a representative sample of the U.S. population with a representative sample of the Harris Poll Online community of 6.6 million respondents. After we complete the interviewing process, we employ logistic regression to develop at least one statistical model that estimates the probability that a respondent participated in the telephone study rather than the Internet study. The probability, known as the “estimated propensity score,” is based on answers to several demographic, behavioral, and attitudinal questions<sup>3</sup>.

Next, in the “propensity score adjustment” step, we group respondents by propensity score within the survey group (telephone or Internet) they represent.

Statistical theory shows us that if the propensity score groupings are carefully and methodically developed, the distribution of characteristics within each Internet grouping will be asymptotically the same as the distribution of characteristics within each corresponding telephone grouping (Rosenbaum & Rubin, 1983). Therefore, by weighting the Internet sample’s propensity group proportions to be the same as the telephone sample’s propensity group proportions, the distribution of characteristics will be asymptotically the same across all propensity groupings within both samples. This procedure produces a result similar to randomization. In other words, the estimated probability of belonging to one group rather than the other will be exactly the same given the variables in the model (as in a randomized controlled experiment).

For subsequent monthly Internet surveys, we estimate each respondent’s propensity score using a model we developed earlier in the month. The approach has allowed us to directly confront, and, in many cases, solve the selection-bias and learning effects problems associated with Internet research.

This comes as no surprise. Propensity score adjustment has also been used to good effect by researchers from diverse sectors<sup>4</sup> to compare smokers with non-smokers (Rosebaum & Rubin, 1984) to estimate the effect of

smoking on mortality rates, high school drop-outs with high school stayers (Rosenbaum, 1986) to estimate the effect of dropping out of high school, and mastectomy versus lumpectomy as a treatment for certain types of breast cancer (USGAO, 1995). In each case, random assignment—the best way to make fair comparisons between two groups—could not be used for ethical or practical reasons.

#### WHAT HARRIS INTERACTIVE EXCLUDES FROM THE MODEL

Rather than taking on faith the accuracy of all telephone responses (and thereby replicating known flaws of telephone research), we instead rely on judgement, theory, and our understanding of the strengths and weaknesses of both telephone and Internet research when developing the propensity score model. In Mosteller’s (1997) words, “the general idea is to let weaknesses from one method of investigation be buttressed by strength from another method, for example, by balancing biases (Chapter 4, p. 116).”

The propensity score models that we use, therefore, exclude questions about the following types of behaviors, the prevalence of which we tend to underestimate through telephone research:

- Traveling
- Dining out
- Cell phone usage, and
- Online shopping and buying

They also exclude questions with method effects. For instance, the question, “Do you believe in God?” tends to produce a significantly higher percentage of the more socially desirable “Yes” response through telephone research than through Internet research, primarily because of the presence of a live interviewer during the telephone interview.

3 Periodically, we are asked by our academic colleagues to identify the questions used in our model. Although we would like to identify these questions, we simply cannot do so. As a publicly owned company, we have obligations to shareholders that preclude us from sharing intellectual property that could conceivably benefit our competitors in the market research industry.

4 The technique of propensity score adjustment has not often used by market researchers or public opinion pollsters despite opportunity and applicability.

## WHY MONTHLY SURVEYS?

The questions we use to estimate each respondent's propensity score may change from month to month for the following reasons:

- The general (telephone) population is changing quickly. According to Harris Poll data, for example, more than 112 million U.S. adults, 18 and older, now access the Internet from home, work, or another location. This represents a gain of ten million over the past six months.
- The Harris Poll Online population is changing (through growth or attrition)—it now numbers 6.6 million, representing a gain of 1.9 million over the past six months.
- As Harris Poll Online respondents participate in more and more surveys, their viewpoints may change. We need to account for the potential learning effects associated with participation in multiple surveys.
- Developing a good propensity score model is hard work. If the variables that we include in our model(s) do not allow us to identify the important differences between the telephone and Internet samples, then adjustment on the basis of the propensity score will be ineffective<sup>5</sup>. For this reason, we continually attempt to refine our model(s).

## ADDITIONAL STEPS TO ENSURE CREDIBLE DATA

We do not rely exclusively on telephone data for validation and calibration purposes. Instead, we attempt to compare responses collected through telephone and Internet research to data known to be true. By combining a parallel survey with information such as revenue numbers from a company like Amazon.com, for example, we are able to improve the accuracy of our weighting process through “triangulation,” another concept that Mosteller (1997) has clearly described:

“In the fourth section of their paper, Boruch and Terhanian discuss work on people who are hard to count and on measurements that are hard to make...With respect to guessing unknown numbers, I have discussed the possibility of trying to estimate the unknown numbers by independently using several different approaches. I call this process triangulation (Chapter 4, p. 117).”

## CRITICISMS OF THE PROPENSITY SCORE ADJUSTMENT

Doug Rivers of InterSurvey has questioned the appropriateness of propensity score adjustment as a means of adjusting for self-selection bias in Internet non-probability samples. He has stated, for instance, that “If the ‘treatments’ are participation and non-participation in a Web survey, then the Web sample contains no non-participants that could be used to estimate the propensity score (Rivers, 2000, p. 38).” With all due respect to Rivers, this statement is not applicable. In fact, we are comparing Internet respondents with telephone respondents. Although we do not know the true probability that our Internet respondents had of participating in the telephone poll, we can estimate this probability through logistic regression. Statistical theory shows us that the estimated propensity score converges to the true propensity score asymptotically. Moreover, simulation results, as well as the properties of the central limit theorem, show us that this happens quickly (Clements, 1997).

Rivers has also stated that parallel surveys are worthless, “since the probability of being in the phone sample is unrelated to the propensity score needed to weight the online sample (p. 38).” This statement is inaccurate. In general, respondents with the same propensity score are those whom we compare. For example, if our model suggests that 20 percent of telephone respondents have a .75 or higher probability of participating in the telephone rather than the Internet survey, we would then weight our Internet sample to ensure that 20 percent of the respondents have a .75 or higher probability of participating in the telephone rather than the Internet survey. Propensity score adjustment, from

<sup>5</sup> To quantify the potential impact associated with the omission of key variables from our model(s), we mount sensitivity analyses (Rosenbaum, 1995) on a routine basis.

this perspective, can be regarded as a powerful matching technique. And the propensity score can be regarded as a weighting factor akin to other weighting factors such as age, race, sex, region, and education level.

Others<sup>6</sup> have questioned how a survey of a portion of a population can possibly represent the entire population when each member of the population does not have some probability of participating in the survey. This problem is not an unfamiliar one to statisticians. Nor is it an insurmountable one.

Consider, for example, a graduate school admissions department that employs statistical modeling to predict how prospective graduate students will fare if admitted. If the graduate school has previously admitted only a few non-white students, then a model that includes race as an independent variable will break down due to sparse data. Fortunately, sparse data need not end the analysis. Through the application of Bayesian statistical methods coupled with a reliance on data from similar graduate schools that have admitted white and non-white students, it is possible to estimate accurately how non-white students will fare in the given school (Rubin, 1980).

Similarly, if the characteristics, beliefs, viewpoints, and behaviors of non-Internet users do not differ appreciably from those of Internet users, then they can be predicted, in theory, through a survey of Internet users coupled with proper statistical adjustment. To test the theory, we have conducted more than two hundred surveys with parallel telephone and Internet components in the past two years. The results of these surveys are indistinguishable in most subject areas.

## CONCLUSION

Using parallel surveys in conjunction with propensity score adjustment is appropriate and vital when the interest lies in producing generalizable information through non-probability sample surveys. The telephone sample constitutes a random sample of the U.S. population, or, in the words of Cochran, Tukey, and Mosteller, “a reasonable probability

sample” that “acts as a check” on the Internet sample, and “as a basis for adjusting the results” of the Internet sample. Through propensity score adjustment, we are able to reduce or eliminate the overt and less-overt differences that distinguish the Internet and telephone samples far more effectively than through demographic weighting alone. And through triangulation, we are able to further improve the accuracy of the information we collect.

Of course, an Internet survey is not the appropriate method for every occasion. If the aim of the research is to understand why people are not online, for example, then it makes no sense to mount an Internet survey. For most types of research, however, an Internet survey is an excellent option.

As we move forward, we will focus our efforts on eliminating or reducing all potential components of error that are associated with Internet surveys of cooperative respondents. Focusing primarily on the elimination or reduction of sampling error has never seemed prudent. Instead, we are guided by the sage advice given by Mosteller and his colleagues in their review of the pre-election polls of 1948:

“If we focus our attention on one of these (sources of error), say sampling, and ignore interviewing, question wording, and other variables, under ideal conditions we might be able to eliminate (the error associated with sampling)...On the other hand, if we had reduced each of the components by 20 percent instead of completely eliminating one, we would have reduced the total (error) nearly twice as much...it is probably necessary to make reductions in error in every part of the operation rather than to try to reduce any particular component to zero” (Mosteller et al., 1949, p. 79).

6 Media pollsters such as Kathy Frankovich and Andrew Kohut have raised this concern in the popular press.

## REFERENCES

- Clements, N.C. (1997). Estimating treatment effects in observational studies: properties of an estimator based on propensity scores. Unpublished doctoral dissertation, University of Chicago.
- Cochran, W.G., Tukey, J.W., and Mosteller, F.M. (1954). Statistical problems of the Kinsey Report. Washington, DC: The American Statistical Association.
- Heckman, J.H. (1992). Randomization and social policy evaluation. In Charles F. Manski & Irwin Garfinkel (Eds.), Evaluating welfare and training programs. Cambridge, MA: Harvard University Press.
- Kinsey, A. C., Pomeroy, W.B., and Martin, C.E. (1948). Sexual behavior in the human male. Philadelphia: W.B. Saunders.
- Mosteller, F. M. (1996). Review of “‘So What?’ The Implications of New Analytic Methods for Designing NCES Surveys.” In Gary Hoachlander, Jeanne E. Griffith, & John H. Ralph (Eds.), From Data to Information: New Directions for the National Center for Education Statistics, NCES 96-901, Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Mitofsky, W.J. (1999). Pollsters.com. Public Perspective, June/July, 24-26.
- Mosteller, F.M., Hyman, H., McCarthy, P.J., Marks, E.S., and Truman, D.B. (1949). The Pre-election polls of 1948. New York. The Social Science Research Council.
- Rubin, D.B. (1980). Using empirical Bayes techniques in the law school validity studies. Journal of the American Statistical Association, 75, 801-827.
- Rivers, D. (2000). Fulfilling the promise of the Web. Quirks Marketing Research Review, February, 34-41.
- Rosenbaum, P. R. (1986). Dropping out of high school in the United States: “An observational study,” Journal of Educational Statistics, 11 (3), 207-224.
- Rosenbaum, P. R. (1995). Observational Studies. New York: Springer-Verlag.
- Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70 (1), 41-55.
- Rosenbaum, P. R., and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association, 79 (387), 516-524.
- U.S. General Accounting Office (1995). Breast conservation versus mastectomy: patient survival in day-to-day medical practice and in randomized studies. Washington, D.C.: USGAO.