HarrisInteractive

2000

Correcting Data from Online Surveys for the Effects of Nonrandom Selection and Nonrandom Assignment

By George Terhanian*, John Bremer, Renee Smith, and Randy Thomas**

For Please do not quote or cite this paper without permission.

White Paper

Introduction

An Internet-based survey is one of many tools that a research organization might use to elicit credible information. An Internet-based survey, however, is certainly not the appropriate tool 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-based survey. If the aim of the research is to understand whether and why Amazon.com customers are satisfied with their online buying experience, however, then it makes complete sense to mount an Internet-based survey. Since Amazon.com routinely collects the e-mail addresses rather than the telephone numbers of its customers, this list of e-mail addresses would constitute a perfect, or near perfect, sampling frame for Internet-based research.

Most Internet-based research, however, is mounted in the absence of a clearly defined sampling frame. Such research offends the sensibilities of many reasonable people who categorically dismiss it on theoretical grounds. In doing so, they often remind those conducting Internet-based research of the failures of the Gallup, Crossley, and Roper organizations to forecast accurately the 1948 presidential election based on nonrandom samples of data.

The categorical dismissal of inferences drawn from Internet-based research, in our opinion, is shortsighted. Social scientists from many disciplines (e.g., Achen 1986; Brehm 1993, 2000; Heckman 1976, 1979; Rosenbaum and Rubin 1983, 1984) have developed sturdy statistical techniques that eliminate or greatly reduce the biases associated with nonrandom selection and nonrandom assignment. At Harris Interactive, we contend that Internet-based research can also be used in the absence of a perfect, or near perfect, sampling frame as long as researchers take the nonrandom aspects of their research design into account. The purpose of this paper is to explain how we correct for the effects of nonrandomness on our survey results. In doing so, we also make clear why it is possible, at times, to produce credible, trustworthy information through Internet-based survey research.

Nonrandom Selection

Nonrandom selection, and hence nonrandom samples, can arise when the decisions of individuals to engage in certain behaviors are correlated with the outcomes those behaviors produce (e.g., Achen 1986). In Internet-based research, there are at least three important decisions individuals must make before their survey responses can be observed.

First, individuals must decide whether they will become part of the US online population. For many Americans, this "decision" is a function of the costs of a computer and Internet access. Second, individuals who are members of the US online population must decide whether to register for and join the Harris Poll Online. Third, individual members of the Harris Poll Online (hereafter HPOL) must decide whether to respond to an invitation to complete a survey.

Because members of the HPOL provide us with basic demographic and "webographic" information when they register, we do have some information about those who do

* George Terhanian is Senior Vice President of the Research and Methodology Division at Harris Interactive.

** John Bremer, Renee Smith, and Randy Thomas are research scientists in the Research and Methodology Division at Harris Interactive. Bremer is also a Ph.D. candidate in political science at the University of Rochester.

not respond. Since we mail our survey invitations to random samples of our 6.5 million HPOL members, we could potentially eliminate the effects of survey nonresponse by using the data for responders and nonresponders along with well-established statistical techniques for censored random variables (e.g., Achen 1986; Heckman 1976, 1979).[1]

The decision to respond to one of our Internet surveys, however, is preceded by two other important decisions. Adjustments for survey nonresponse, therefore, can reduce, but not eliminate, the effects of nonrandom selection in data from Internet-based research. Because you must first be online to join the HPOL and because we send survey invitations only to those who have joined, our Internet-based samples do not contain information about those who are not yet online or about those who are online but have joined the HPOL. In statistical terms, the data we gather via the Internet are truncated,[2] and hence, do not necessarily represent the characteristics of the US general population.[3]

Corrections for Nonrandom Selection

At Harris Interactive, we begin our efforts to tackle the effects of truncation on our observed data by conducting parallel telephone surveys in which respondents are asked the same set of questions posed to our HPOL members. Using the telephone, we are able to obtain survey responses both from individuals who are not yet online and from individuals who are online but are not members of the HPOL.

We then merge the data from the telephone and online respondents and use logistic regression to estimate the probability that a respondent answers our survey online.[4] We could, at this point, use the estimates from the "selection" equation to compute a hazard rate that could be used as a covariate in a second-stage model (e.g., Achen 1986; Heckman 1976, 1979). Because our pooled telephone and online data set has complete data for both selection and outcomes, estimators for sample selection bias will be inefficient.

The approximately 1,200 telephone respondents in our pooled data set are contacted using RDD, which in theory produces a representative sample of the US general population. (The degree to which RDD produces a representative sample, in practice, is debatable.[5]) From a pragmatic standpoint, most critics of Internet-based research argue that survey responses collected via telephone polls based on RDD (or via other media with respondents recruited by telephone and RDD), are representative of the US general population.

For instance, Rivers (2000, 39) argues that "the first step in creating a valid panel for consumer research is to recruit households using random digit dialing." The result, Rivers (2000, 40) says is that, "the sample is representative of the entire population because it uses valid probability sampling techniques, and does not exclude households because they lack computers or Internet access." At Harris Interactive, we agree that corrections for nonrandom selection can be based on probability samples produced by RDD or by other randomization methods.

The Harris Interactive Approach

The approach we take is neither new nor novel. Rather, it is akin to one that Cochrane, Tukey, and Mosteller (1954) described in their review of the controversial *Kinsey Report on Sexual Behavior in the Human Male* (1948), a report that depended on nonrandom samples. Nearly 50 years ago, they (1954, 23) wrote:

[1] Manski (1995, 21) defines the selection problem associated with survey nonresponse as "the problem of identifying conditional probability distributions from random sample data in which the realizations of the conditioning variables are always observed but the realizations of outcomes are censored."

[2] Truncation occurs when an observed variable contains information about a range of values above or below a certain number in the corresponding untruncated random variable (Greene 2000).

[3] Of course, sample selection need not result in bias. When the unobserved causes of selection are uncorrelated with the unobserved factors affecting the out comes of interest, no bias arises (Achen 1986; Brehm 1993).

[4] Brehm (2000) has recently urged political scientists to correct for nonresponse bias in ANES studies by obtaining additional information sufficient to turn a truncated sample into a censored sample. As we discuss below, we take this suggestion one step further.

[5] In practice, almost all surveys suffer from some degree of nonresponse bias. For a detailed discussion of the problems associated with "phantom respondents," see Brehm (1993).

> 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.[6]

Although the technique of "propensity score adjustment" (Rosenbaum & Rubin, 1983) did not yet exist, we suspect that it could have been used quite effectively to eliminate or reduce the self-selection bias in the KPM sample.

Calibration Through Propensity Score Adjustment

Propensity score techniques were initially developed to solve the problem of estimating treatment effects in non-randomized studies. The application of propensity score methods to adjust data obtained from Internet-based research requires researchers to re-conceptualize the problems associated with self selection.

Rather than thinking about the sources of nonrandomness in data from an online survey, we can instead consider the sources of nonrandomness within our pooled data. Once attention shifts to the merged data set, it becomes easier to understand how the nonrandom assignment to treatment (online survey) and control (telephone survey) groups is now equivalent to the problems from nonrandom selection discussed above.

In the pooled telephone and online data, we observe data on assignment and outcomes for all of the key subgroups - those who are offline, those who are online but not members of the HPOL, *and online respondents who are members of the HPOL*. As a result, we now have information that will allow us to estimate the probability of completing the online survey (the treatment) for all types of respondents. The probability estimates obtained from this model are called propensity scores.

Propensity Score Theory

Propensity scores summarize the effect of a set of covariates on the probability of receiving a treatment, and provide researchers with information that can be used to matched respondents from the treatment and control groups. Formally, let Z be a binary variable that denotes treatment status. Z=1 implies that the treatment has been received and Z=0 implies that the treatment has not been received. In the framework used by Harris Interactive, the treatment is participation in the Internet survey and the control is participation in an RDD telephone based survey.

In a properly conducted randomized trial, the treatment assignment, Z, and the response (r_1, r_0) are conditionally independent given a set of covariates X. Although this condition is not generally known to hold in non-randomized studies, Rosenbaum and Rubin (1984) show that it can be induced. They refer to situations in which conditional independence is induced as cases of strongly ignorable treatment assignment.

Generally, treatment assignment can be said to be strongly ignorable if

$$(r_1,r_0)\perp Z \mid X,\ 0<P(Z=1 \mid X)<1$$

In a randomized experiment, the fact that treatment assignment can be ignored allows researchers to be able to compare matched sets of individuals who differ systematically only to the extent that the treatment caused an effect. Propensity score theory is based on the idea that the desirable properties generated by random selection (i.e., ability to ignore treatment assignment) can also be generated by the appropriate selection of covariates.[7]

If the covariates are correctly chosen, then sub-classification of treatment and control groups based on propensity scores will yield sub-classes in which the distribution of characteristics in each sub-classification are approximately equal across treatment and control groups. Conditional on X, all remaining differences are assumed to random.

[6] It is interesting to note that Cochrane, Tukey, and Mosteller did not categorically dismiss evidence produced through a nonrandom sample despite the fact that Mosteller was the lead author of *The Pre-Election Polls of 1948*, published five years earlier.

[7] If treatment assignment is strongly ignorable given X, then it is also strongly ignorable given the propensity score that summarizes those covariates (Rubin and Rosenbaum 1983). This result allows the use of a single covariate as a means of inducing strong ignorability instead of a multi-dimensional set of variables, thus reducing the dimensionality and complexity of the problem.

For the desirable properties of properly adjusted propensity matched data to hold, the data must be appropriately sub-classified into homogeneous groups based on their propensity score. Cochran (1968) shows that the most efficient sub-classification occurs when the data are sub-classified into five distinct groups. Cochran's result holds for propensity score adjustment as well as for other metric methods of sample stratification.[8]

Once data have been sub-classified by propensity score, analysts can make direct comparisons across sub-classes of similar individuals instead of across the data sets themselves. The groups need not be of similar size but must consist of similarly matched members. Cochran (1968) and Rubin and Rosenbaum (1985) both demonstrate that sub-classification into an appropriate number of groupings based on an appropriate set of covariates or on the propensity score, $\lambda(x)$, reduces 90% of the selection bias associated with the non-random aspects of the experiment. This finding is with respect to the outcomes as well as the treatment assignment. (Outcomes in this framework are the responses to the survey questions posed by Harris Interactive.) Furthermore, Rosenbaum and Rubin (1983) show that if treatment assignment is strongly ignorable, samples are large, and subclasses are homogenous in the propensity score then direct adjustment will result in unbiased estimates of the outcomes.

To show that the distributions of characteristics are equal across treatment assignment, within strata, we proceed by introducing some additional notation and information about propensity scores. Let s indicate a particular sub-classified group or strata. In addition, assume that within a particular strata s, the units are homogenous or nearly homogenous with respect to the propensity score. If this is not the case, then the sub-classification algorithm has failed and treatment assignment is not strongly ignorable. In terms of the proposition below, we will assume homogeneity in the propensity score, although near homogeneity, leads to the same result. In addition, we note that for every value of the propensity score, $\lambda(X)=\Gamma$, by definition $P(Z=1|\lambda(X)=\Gamma) = \Gamma$.

These definitions allow us to propose the following proposition made in a similar manner in Rubin and Rosenbaum (1983) and Rosenbaum (1995).

Proposition 1: If $\lambda(x_S)=\Gamma$ then $P(X=x_S|\lambda(X)=\Gamma, Z=1)= P(X= x_S|\lambda(X)=\Gamma, Z=0)$

Proof: $\lambda(x_S)=\Gamma$ implies that by Bayes' Theorem

$$P(X=x_s|\lambda(X)=\Gamma,Z=1)= \frac{P(Z=1|\lambda(X)=\Gamma,X=x_s)*P(X=x_s|\lambda(X)=\Gamma)}{P(Z=1|\lambda(X)=\Gamma)} (1)$$

Because of the homogeneity of the strata, we have that

$$P(Z=1|\lambda(X)=\Gamma,X=x_s)=P(Z=1|X=x_s)=\lambda(x_s)=\Gamma \ (2)$$

In addition, we know that

$$P(Z=1|\lambda(X)=\Gamma)=\Gamma \ (3).$$

(2) and (3) taken together reduce (1) to the following equation:

$$P(X=x_s|\lambda(X)=\Gamma,Z=1)=P(X=x_s|\lambda(X)=\Gamma) \ (4)$$

In the exact same manner, it can be shown that

$$P(X=x_s|\lambda(X)=\Gamma,Z=0)=P(X=x_s|\lambda(X)=\Gamma) \ (5)$$

(4) and (5) taken together prove the proposition.

The result shows that the distribution of the observed characteristics of the treatment and control groups properly stratified are equal within sub-classes across treatment status. Hence, observed differences, conditional on the propensity score, are non-systematic and due to chance. Proper stratification will eliminate 90% of the bias in estimates of treatment effects on outcomes.

Propensity Score Adjustment in Practice

At Harris Interactive, we exploit the results of propensity score theory as we weight data collected from online

[8] For instance, if X is distributed multivariate normal, the propensity score is equivalent to the discriminant function. Because propensity score theory is not based on parametric assumptions, it is more general. Rosenbaum and Rubin (1984) show that the propensity score is the coarsest function of X that will produce balanced sub-classes; that is, the propensity score is the function that needs the smallest number of sub-classes to produce balance.

surveys.[9] The weighting algorithm used by Harris Interactive is as follows:

1. Appropriate covariates are identified such that the condition of strongly ignorable treatment assignment is met either exactly or approximately. These covariates, which include demographic, behavioral, attitudinal, and topic-specific variables, are included in both the telephone and online version of the survey.
2. Respondents are randomly chosen from the Harris Interactive database for invitation to the online version of the survey, and they are chosen via conventional RDD methodologies for contact in the telephone version.
3. Data is collected using both methods and merged.
4. The appropriate propensity score model is estimated using logistic regression, and respondents from the phone and online surveys are sub-classified based on their propensity scores. It must be noted at this point, that there must be sufficient overlap in each stratum between the online respondents and the telephone respondents. If there is not, then the condition of strong ignorability is not met, and the weighting procedure will fail. This is a consideration when constructing a propensity model. Unlike other statistical procedures, propensity models can fit too well resulting in little or no overlaps between treatment and control groups.
5. The data are rim weighted with the propensity stratification as one factor, and traditional demographic variables as others.[10]

By assuming that the large RDD telephone samples produce data relatively free of bias,[11] the online data can be weighted so that the percentages of individuals in each subclass are the same across treatment assignment. Propensity score adjustments, are used in addition to the typical demographic weighting used by many survey houses to yield results that are projectable to the general US population.

As proposition one shows, conditioning on observed covariates produces sub-classes that are approximately equal across treatment assignment for homogenous strata. Unobserved confounding factors, however, are not explicitly controlled for using propensity score adjustment. Choice of model covariates is therefore highly important. Covariates must be chosen not only to reduce the bias from observed characteristics, but also as proxies for unobserved factors that might affect both treatment assignment and outcomes.

Our propensity score models are not designed to replicate known flaws of phone research. Therefore, it does not include questions that elicit information on the following types of behaviors, which we tend to underestimate through phone research:

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

Instead, we attempt to balance the biases of the two methods. 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."

Why Monthly Surveys?

The questions that 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 (e.g., increasing Internet usage)—according to Harris Poll data, more than 56% of all adults, 18 and older, in the United States now access the Internet from home, work, or another location.

[9] It must be mentioned, at this point, that the propensity score used in the practical application of the weighting algorithm Harris Interactive uses is not known exactly as is assumed in the theoretical proof. It is estimated using logistic regression. This is the case in generally all practical uses of propensity score matching. Clements (1997) shows not only that the estimated propensity score converges asymptotically to the real propensity score but also that this convergence is extremely quick. This fact leads to theoretical results that are the same for both estimated and known propensity scores. (See Clements (1997) for further discussion.)

[10] The rim weighting algorithm is based on least squares corrections of the weighting factors. (See Deming and Stephan, 1940)

[11] As mentioned previously, because there are a few areas in which there are known biases in RDD telephone surveys. As a result, we do not include covariates that examine those areas in the propensity model. We do not include in the model questions about travel, dining out, or cell phone use, questions that elicit a socially desirable response, or questions whose responses are dependent on the method used to present them during the survey.

- The Harris Poll Online population may be changing (through growth or attrition)—it now numbers 6.2 million.

- To account for potential learning effects through participation in multiple surveys (e.g., the act of participating in multiple surveys may change respondents' viewpoints).

Empirical Evidence: Does Propensity Score Adjustment Work?

To evaluate the degree to which propensity score adjustment works, we can compare the propensity weighted results from our Internet-based surveys to those from telephone or other surveys. In this section, we discuss a number of such comparisons.

Table 1 reports results from our June 1999 Harris telephone poll (HPTEL) and from the parallel Harris Poll Online (HPOL). To see the differences that propensity weighting makes, we report three sets of HPOL results. HPOL-U refers to the unweighted or raw HPOL results; HPOL-D refers to HPOL results that have been weighted **only** to demographic targets (age, sex, region, race, education, and income); HPOL-P refers to HPOL results that have been weighted using the same demographic targets and propensity score sub-class targets. Using the HPTEL results as a benchmark, we can evaluate the degree to which propensity weighted online data is representative of the US general population.

Because HPOL members volunteer to take our surveys, they have already shown themselves to be more likely to participate in surveys than other members of the online population. Such behavior could produce biased estimates of political or consumer participation. As the top half of Table 1 shows, the unweighted percentages of people who report they have participated in various types of political activities are dramatically higher than are those obtained from the HPTEL. Demographic weighting, moreover, fails to remove all of the differences between the percentages for online and telephone respondents. The propensity weighted data, however, produce results that are very similar to their telephone counterparts. Across the seven participation items, the mean average deviation between the HPTEL and propensity weighted HPOL results is only 2.1 percentage points. The results in the bottom half of Table 1 illustrate the effects of propensity score weighting on attitudinal measures. A similar pattern holds for these items such that the mean average deviation for the propensity weighted data is only 2.2 percentage points across the eight items.

Table 1 June 1999 HPTEL/HPOL Comparisons

	HPTEL	HPOL-U	HPOL-D	HPOL-P
Political Participation				
Contributed Money to Party or Campaign	20%	29%	21%	22%
Called, Written or Visited Elected Official	32%	61%	49%	38%
Written a Letter to Newspaper, Magazine, TV Station	16%	28%	23%	18%
Called into a Talk Show to Express Opinion	10%	15%	12%	11%
Attended Mtg. Where Political/Elected Official Spoke	37%	48%	37%	35%
Worked on Political Campaign	10%	16%	12%	11%
Display Campaign Items	35%	44%	37%	36%
Political Opinion				
Presidential Approval	55%	48%	52%	56%
Kosovo - Informed	94%	99%	98%	98%
News Contributes to Violence	39%	35%	39%	40%
Video Games Contribute to Violence	47%	40%	43%	45%
Television Contributes to Violence	58%	48%	51%	54%
Movies Contribute to Violence	57%	51%	53%	57%
Lack of Supervision Contributes to Violence	90%	93%	91%	92%
Easy Availability of Guns Contributes to Violence	65%	52%	57%	60%

Table 2 (see below) reports the results of propensity weighting on attitudinal measures (mean average deviation = 2.5 percentage points across 6 items) and voting choice items from October, 1999. The data from these trial heats show that both demographic weighting alone, and propensity weighting result in mean absolute deviations of 1.5 percentage points, although the pattern of deviations varies under the two weighting schemes.

Our final comparison of HPTEL and propensity weighted HPOL data appears in Table 3 (see below), which shows results from June, 2000. For the attitudinal items, the mean average deviation of propensity weighted HPOL data is 1.8 percentage points. As before, propensity weighting brings many of the percentages into alignment with the telephone results. Interestingly, the mean average deviation for the trial heat results is higher for the propensity weighted data than it is for the HPOL data weighted only by demographics.

Clearly using the HPTEL results as a benchmark, does not always allow us to determine whether propensity weighting is more effective than demographic weighting alone. To evaluate this matter further, we can also compare our propensity weighted results to the results obtained by other survey organizations.

Empirically, there are two other ways of judging the accuracy of the data beyond looking at how the weighted Internet data compares to the weighted telephone data. One additional way to judge the data is, by comparing our results to the results of surveys conducted independently by other organizations. The second way to judge the accuracy of the data is, to compare the results to known population values.

Table 2 October 1999 HPTEL/HPOL Comparisons

	HPTEL	HPOL-U	HPOL-D	HPOL-P
Public Opinion				
Presidential Approval	57%	51%	55%	54%
Congressional Democratic Approval	42%	32%	38%	38%
Congressional Republican Approval	33%	31%	29%	31%
Trent Lott Approval	29%	29%	28%	29%
Dennis Hastert Approval	25%	24%	25%	26%
VP Gore Approval	42%	33%	35%	37%
Treat Heats				
Presidential Elections (Gore vs. Bush) - Bush	53%	58%	56%	55%
Presidential Elections (Gore vs. Dole) - Dole	37%	36%	39%	38%
Presidential Elections (Bush vs. Bradley) - Bush	54%	55%	52%	55%
Presidential Elections (Bush vs. Dole) - Dole	39%	38%	40%	37%

Table 3 June 2000 HPTEL/HPOL Comparisons

	HPTEL	HPOL-U	HPOL-D	HPOL-P
Public Opinion				
Presidential Approval	56%	53%	56%	60%
Congressional Democratic Approval	38%	34%	36%	36%
Congressional Republican Approval	35%	34%	33%	33%
Trent Lott Approval	27%	27%	27%	27%
Dennis Hastert Approval	25%	26%	25%	25%
VP Gore Approval	41%	35%	36%	38%
Trial Heats				
Vote for Bush	54%	54%	52%	49%
Vote for Gore	46%	42%	43%	46%

Fortunately, the over abundance of political polls allows us to compare our data with that of other organizations asking the same or similar questions about election preferences. In Table 4, we present a comparison of the results of general election match-ups between the nominees of the two parties. Results are presented for three organizations – Harris Interactive, ABC News, and the Gallup Organization. Because of the variation inherent in the results of political polls that are conducted on the telephone due to different methodologies concerning who a likely voter is, these two comparison organizations were selected because of the similarity of their voter selection methods to those of Harris Interactive.[12]

Table 4 Comaprison of Polling Data Across Organizations, 2000

	Harris Interactive	ABC	Gallup
January			
Bush	51%	51%	53%
Gore	42%	41%	42%
February			
Bush	49%	49%	50%
Gore	46%	45%	45%
April			
Bush	48%	46%	50%
Gore	47%	47%	41%
May			
Bush	49%	49%	49%
Gore	45%	44%	44%
June			
Bush	49%	49%	48%
Gore	46%	45%	44%

Average difference 0.006 0.016

Note: Harris Interactive did not have a March Election Poll.

Over the last five months that Harris Interactive and ABC News have been running polls at similar times, the average candidate difference is 0.6 percentage points. In that same period, the average candidate difference between Harris Interactive and the Gallup Organization is 1.6 percentage points. This value is inflated because of a single month (April) in which Gallup had Gore at 6 percentage points less than either Harris Interactive or ABC News. Excluding that result, the average candidate difference is even less than the 1.6 percentage points reported on the table. Clearly, when method effects that are not related to the survey medium are excluded, the outcomes reported by Harris Interactive are similar to the outcomes reported by other organizations using different mediums but asking similar questions.

A second comparison to the results of elections can also be made. We will be forecasting the 2000 general elections in their entirety using the propensity score matching weight-

Table 5 Comparison of Harris Interactive Election Forecasts with Actual Election Results

	NY	Actual	Difference	GA	Actual	Difference	OH	Actual	Difference	CA	Actual	Difference
Bush	50.0%	51.0%	1.0%	66.0%	67.0%	1.0%	51.0%	58.0%	7.0%	28.0%	28.0%	0.0%
Keyes	2.0%	4.0%	2.0%	9.0%	4.0%	5.0%	4.0%	4.0%	0.0%	4.0%	2.0%	2.0%
McCain	46.0%	43.0%	3.0%	25.0%	28.0%	3.0%	43.0%	37.0%	6.0%	24.0%	23.0%	1.0%
Al Gore	65.0%	65.0%	0.0%	80.0%	84.0%	4.0%	76.0%	73.0%	3.0%	33.0%	35.0%	2.0%
Bradley	33.0%	35.0%	2.0%	13.0%	16.0%	3.0%	21.0%	25.0%	4.0%	9.0%	9.0%	0.0%

Canidate Error 2.5%

Spread Error 4.0%

[12] In fact, Louis Harris and Associates, the predecessor of Harris Interactive, was the main polling organization for ABC News for a number of years. The methodologies used to select likely voters are nearly identical.

ing methodology described here. The first set of elections for which the weighting methodology was fully used was a trial run in the primaries of March, 2000.[13] The results are in Table 5. We were able to correctly name the winner in each contest we attempted. Furthermore, our candidate error dropped from 3.8 percentage points in 1998 to 2.5 percentage points in the 2000 primaries. In addition, our average spread error went from 5.6 percentage points in 1998 to 4.0 percentage points in the 2000 primaries. These reductions in the two main types of error associated with election forecasting are significant given that in 1998, Harris Interactive had smaller errors associated with their forecasts than most other polling organizations. Even more significant is the precision in which we were able to forecast Georgia, a state that caused Harris Interactive problems in 1998.

Summary

The Harris Interactive approach to propensity score adjustment of data from online surveys has been shown to reduce and/or eliminate biases due to nonrandom selection and nonrandom assignment. The intuition undergirding this success is illustrated in Figures 1a through 1c and 2a through 2c for data from June, 1999 and May, 2000, respectively.

These figures show the estimated probability of being a telephone respondent for those who responded by telephone, and for those who responded online. Figures 1a and 2a show the distributions of these probabilities before either the telephone or online data have been weighted. Figures 1b and 2b show the distributions of these probabilities after the telephone and online data have each been weighted to meet demographic targets. Figures 1c and 2c show a comparison of the distribution for the demographically weighted telephone data with the distribution for the propensity weighted online data.

As these figures illustrate, propensity weighting results in probability distributions that are equivalent for both telephone and online respondents. Determining which of the two groups a respondent is in now appears to be the result of a coin flip. That is, any differences in the propensity to be a telephone respondent are now due to random rather than systematic factors.

Generalizing from data obtained via online surveys to the US population is not impossible. Provided that treatment assignment conditional on covariates is strongly ignorable, we can make fair comparisons and produce generalizable results. We agree wholeheartedly with Rubin, who tells us “If a nonrandomized study is carefully controlled, the investigator can reach conclusions similar to those he would reach in a similar [randomized] experiment” (1974, 700).

As we move forward, we will focus our efforts on eliminating or reducing all components of error that are associated with Internet-based surveys of cooperative respondents. Focusing primarily on the elimination or reduction of sampling error has never seemed prudent. Instead, we have always been guided by the sage advice given by Mosteller et al. (1949, 79) in their review of the pre-election polls of 1948:

> If we focus our attention on one of these, 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.

Some of our colleagues argue that we have forgotten the past, notably, the mistakes made by polling organizations that depended on non-probability samples in the 1948 pre-election polls. We instead contend, that our use of propensity score adjustment shows that we are indeed keenly aware of the mistakes of the past and have successfully moved beyond them.

[13] It must be mentioned that Harris Interactive did quite well in forecasting the elections of 1998 using demographic weighting. We correctly forecast 95% of the elections that we attempted to estimate, missing on only Georgia. In addition, our spread error and candidate error for the final polls were smaller than the average error for telephone polls.

Appendix

Figure 1 Density Plot of Telephone Respondent Probability: June HPOL/HPTEL Unweighted Data

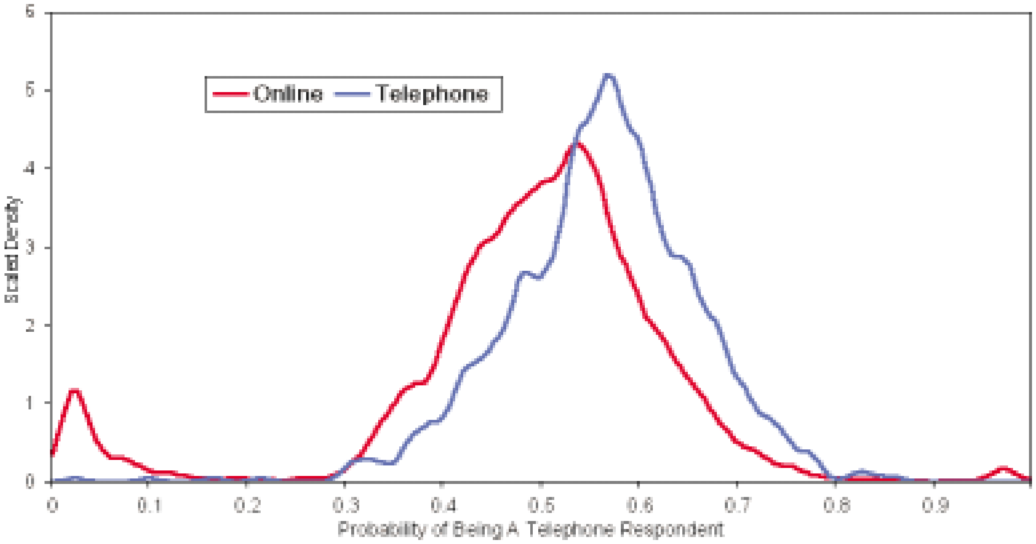


Figure 2 Density Plot of Telephone Respondent Probability: June HPOL/HPTEL Rim Weighted Data

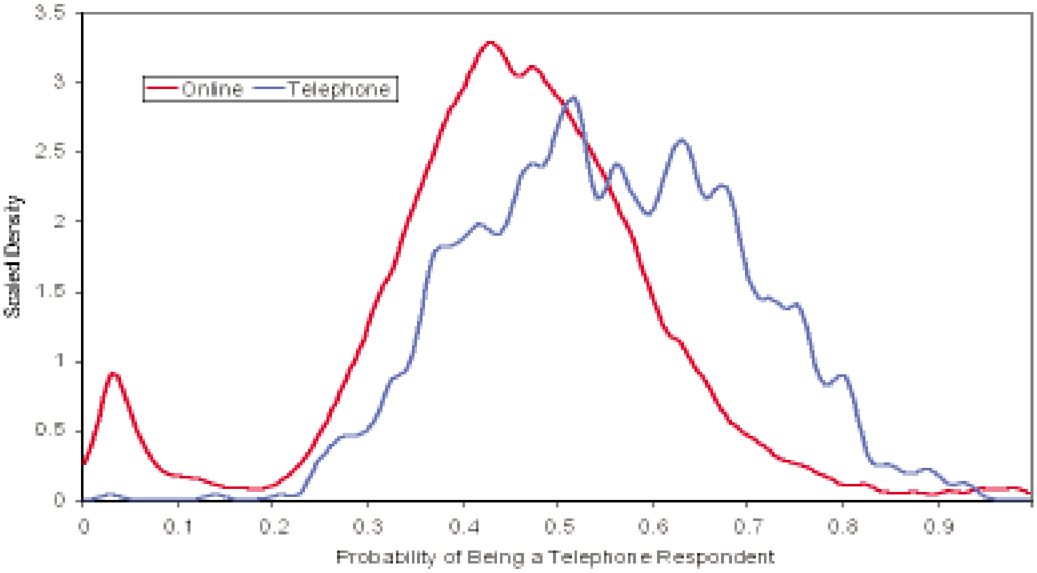


Figure 3 Density Plot of Telephone Respondent Probability: June HPOL/HPTEL Propensity Data

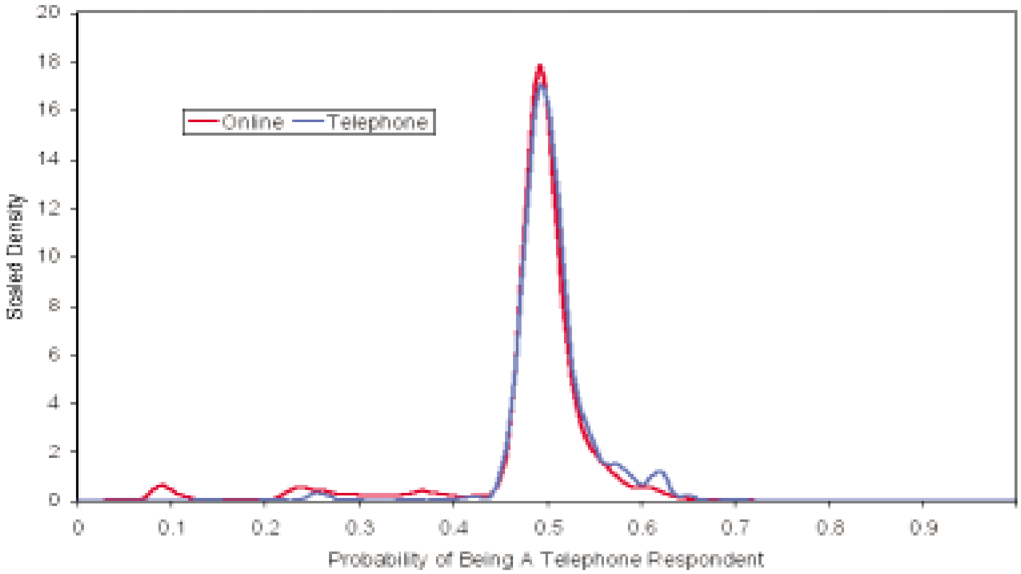


Figure 4 Density Plot of Telephone Respondent Probability: May HPOL/HPTEL Unweighted Data

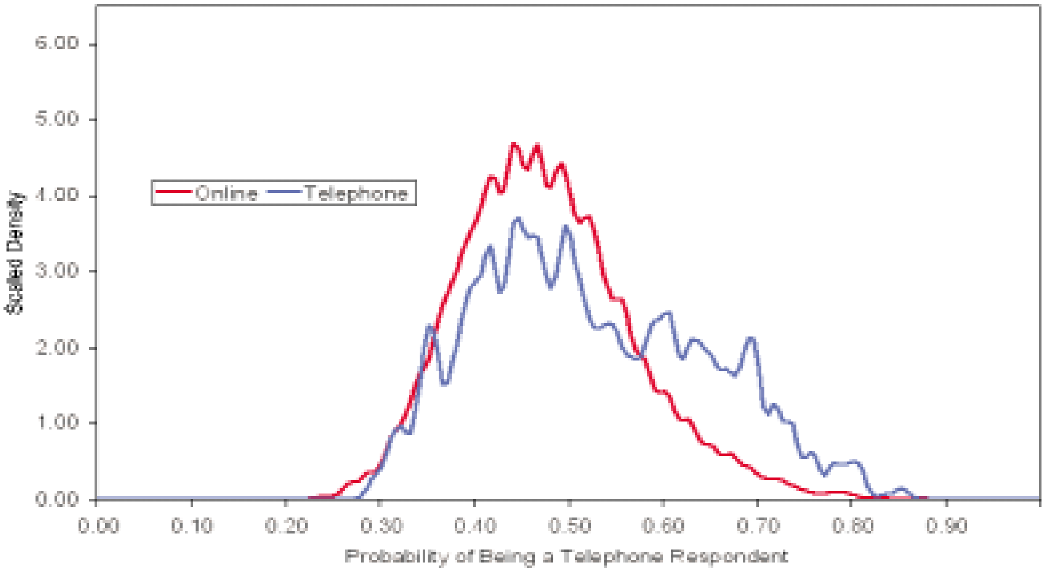


Figure 5 Density Plot of Telephone Respondent Probability: May HPOL/HPTEL Data Weighed by Demographics

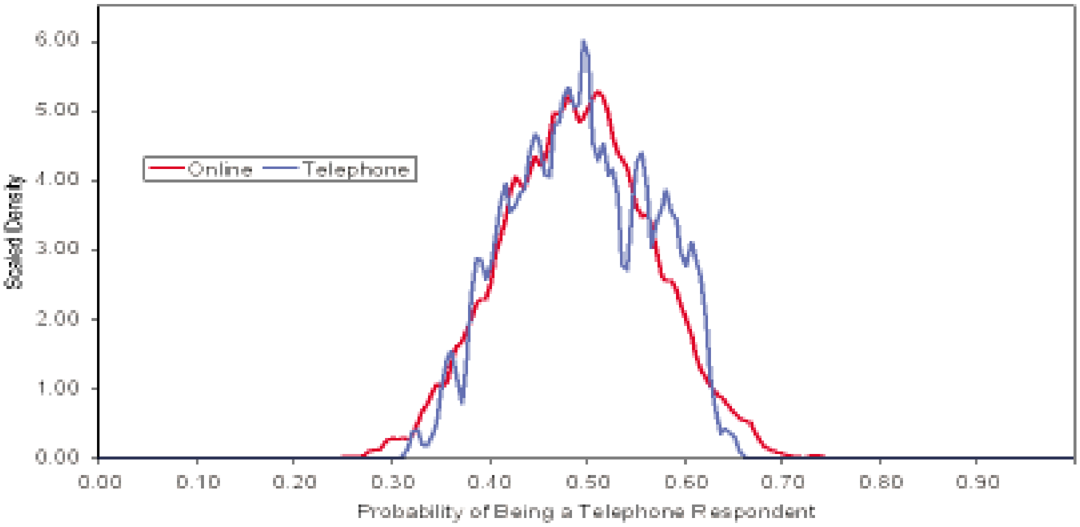
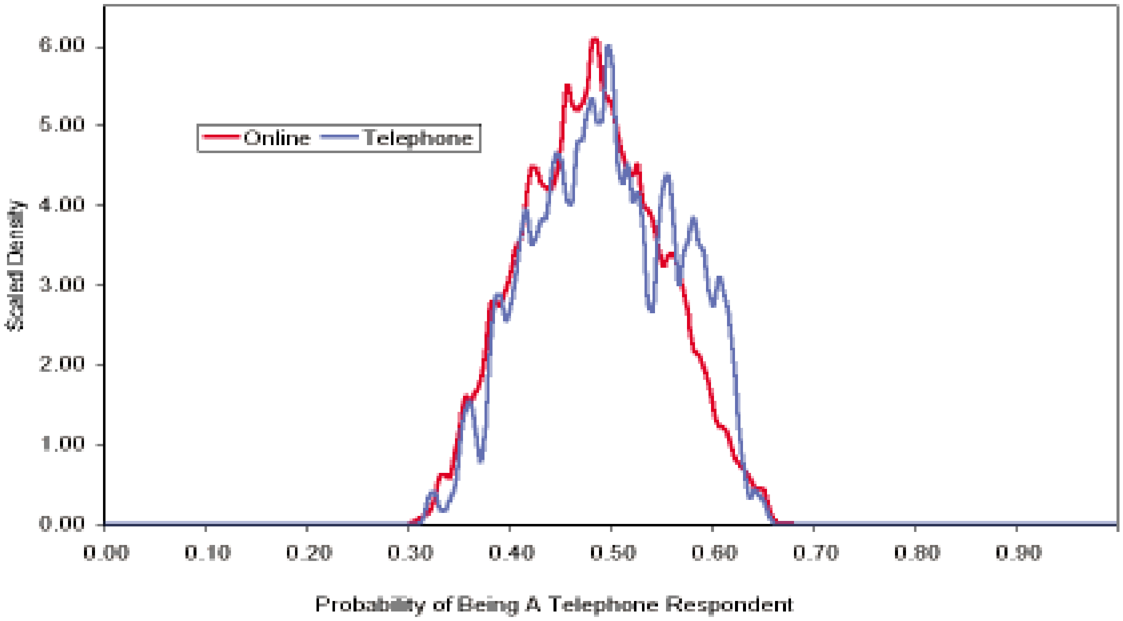


Figure 6 Density Plot of Telephone Respondent Probability: May HPOL/HPTEL Propensity Weighted Data



References

Achen, Christopher H. 1986. The Statistical Analysis of Quasi-Experiments. Berkeley, CA: The University of California Press.

Boruch, R.F., & Terhanian, G. 1996. “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.

Boruch, R.F., & Terhanian, G. 1998. “Controlled Experiments and Survey-Based Approaches to Productivity Research: Cross Design Synthesis” in H. Walberg & A. Reynolds (Eds.) Advances in Educational Productivity, Greenwich, CT: Jai Press.

Brehm, John 1993. The Phantom Respondents: Opinion Surveys and Political Representation. Ann Arbor, MI: University of Michigan Press.

Brehm, John 2000. “Alternative Corrections for Sample Truncation: Applications to the 1988, 1990, and 1992 Senate Election Studies” Political Analysis. 8: 183-200.

Clements, Nancy C. 1997. “Estimating Treatment Effects in Observational Studies: Properties of an Estimator Based on Propensity Scores” Ph.D. dissertation. University of Chicago.

Cochran, W.G. 1968. “The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies” Biometrics. 24: 295-313.

Deming, W.E. and F.F. Stephan 1940. “On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known” Annals of Mathematical Statistics. 11: 427-444.

Greene, William H. 2000. Econometric Analysis. 4th ed. Upper Saddle River, NJ: Prentice Hall.

Heckman, James J. 1976. “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models” Annals of Economic and Social Measurement. 5(4): 475-492.

Heckman, James J. 1979. “Sample Selection Bias as a Specification Error” Econometrica. 47: 153-161.

Manski, Charles F. 1995. Identification Problems in the Social Sciences. Cambridge: Harvard University Press.

Rivers, Douglas 2000. “Fulfilling the Promise of the Web: Research Approach Aims to Make the Internet Safe for Consumer Research” Quirk’s Marketing Research Review. February: 34-41.

Rosenbaum, P. R., and Rubin, D. B. 1983. “The Central Role of the Propensity Score in Observational Studies for Casual 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: 516-524.

Rosenbaum, P.R. and Rubin, D.B. 1985. “Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score” The American Statistician. 39: 33-38.

Rubin, D.B. 1974. “Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies” Journal of Educational Psychology. 66: 688-701.

Rubin, D.B. 1979. “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies” Journal of the American Statistical Association. 74: 318-328.