

Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with  
Probability and Non-Probability Samples

David S. Yeager and Jon A. Krosnick

Stanford University

LinChiat Chang

Kantar Health

Harold S. Javitz

SRI International, Inc.

Matthew S. Levindusky

University of Pennsylvania

Alberto Simpser

University of Chicago

Rui Wang

Stanford University

August, 2009

Jon Krosnick is University Fellow at Resources for the Future. The authors thank Norman Nie for making this research possible, Douglas Rivers for collaborating in the study design and data collection, SPSS, Inc., for funding the project, and Josh Pasek, Yphtach Lelkes, Neil Malhotra and other members of the Political Psychology Research Group at Stanford for helpful suggestions. Some data used here were obtained from the Roper Center for Public Opinion Research at the University of Connecticut. Address correspondence to David Yeager 271 Jordan Hall, Stanford University, Stanford, California 94305 (email: [dyeager@stanford.edu](mailto:dyeager@stanford.edu)) or to Jon Krosnick, 432 McClatchy Hall, 450 Serra Mall, Stanford University, Stanford, California 94305 (email: [krosnick@stanford.edu](mailto:krosnick@stanford.edu)).

# Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples

## Abstract

This study compared the accuracy of surveys of probability samples of American adults with surveys of non-probability samples of people who volunteer to do surveys for money or prizes. Data from one Random Digit Dialing (RDD) telephone survey, one Internet survey of a probability sample recruited by RDD, and seven Internet surveys of non-probability samples were compared against benchmarks to assess accuracy. The probability sample surveys were consistently more accurate than the non-probability sample surveys, even after post-stratification with demographics. With the non-probability sample surveys, post-stratification improved the accuracy of some measures and decreased the accuracy of other measures, and post-stratification improved the overall accuracy of some surveys while decreasing the overall accuracy of others. These results suggest caution before asserting or presuming that non-probability samples yield data that are as accurate or more accurate than data obtained from probability samples.

# Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples

## **Introduction**

Since the 1940s, the gold standard for survey research has been face-to-face interviews with a random sample of American adults. Surveys such as the U.S. Census Bureau's Current Population Survey (CPS) and the Centers for Disease Control and Prevention's National Health Interview Survey (NHIS) have involved interviews with tens of thousands of randomly selected Americans who have been interviewed face-to-face with extremely high response rates. As a result, such surveys are among America's most trusted sources of information about its population.

Outside of government, face-to-face interviewing of probability samples is unusual in the world of survey research today, though it is still done regularly (e.g., by the American National Election Studies and the General Social Survey). Since the 1970s, Random Digit Dialing (RDD) telephone surveys have been very popular, and in recent years, Internet surveys have been conducted at increasing rates. Some of these Internet surveys have been done with probability samples of the population of interest (e.g., Carbone, 2005; Kaptyen, Smith and van Soest, 2007; Lerner et al. 2003; Malhotra and Kuo 2008, Moskalenko and McCauley 2009; Skitka and Bauman 2008), and some evidence suggests that data collection from probability samples via the Internet may yield results that are as accurate as or more accurate than RDD telephone interviews or face-to-face interviewing with area probability samples (e.g., Chang and Krosnick in press; Smith 2003).

For example, in a laboratory experiment, respondents were randomly assigned to answer questions on a computer or orally over an intercom (simulating a telephone), and the former

method yielded higher concurrent validity, less survey satisficing, and less social desirability bias (Chang and Krosnick in press). Similar evidence was produced by a field experiment that administered the same questionnaire to (1) a probability RDD sample interviewed by telephone, and (2) a probability RDD sample of the same population who answered questions via the Internet. Responses collected via telephone interviewing manifested more random measurement error, more survey satisficing, and more social desirability response bias than did data collected via the Internet (Chang and Krosnick in press). Thus, Internet survey administration appears to offer significant measurement advantages.

But most commercial companies that collect survey data in the U.S. via the Internet do not interview probability samples drawn from the population of interest with known probabilities of selection. Instead, most of these companies offer only non-probability samples of people who were not systematically sampled from a population using conventional sampling methods (for an overview, see Postoaca 2006). For some such surveys, banner ads are placed on websites inviting people to volunteer to do surveys in order to earn money or to win prizes. In other instances, email invitations are sent to large numbers of people whose email addresses are sold by commercial vendors or maintained by online organizations because of past purchases of goods or services from them. Some of the people who read these invitations then sign up to join an Internet survey “panel” and are later invited to complete questionnaires. For any given survey, a potential respondent’s probability of selection from the panel is usually known, but their probability of selection from the population of interest is not. We refer to this as “opt-in” sampling.<sup>1</sup>

---

<sup>1</sup> When probability sampling is implemented, each invited individual is selected from the population of interest by a fully documentable method. Therefore, the probability of selection of

In theory, non-probability samples may sometimes yield results that are just as accurate as probability samples. If the factors that determine a population member's presence or absence in the sample are all uncorrelated with variables of interest in a study, then the observed distributions of those variables should be identical to the distributions in the population (within the limits of sampling error). And if the factors that determine a population member's presence or absence in the sample are all uncorrelated with the magnitudes of associations between pairs of variables measured in the study, then the observed associations should also be identical to those in the population (again, within the limits of sampling error). However, if these conditions do not hold, then survey results may not be comparable to those that would be obtained from probability samples.

To date, one study has compared probability samples interviewed by telephone and via the Internet to an opt-in Internet sample. This study found the latter to be less representative of the population in terms of demographics and to over-represent people with high interest in the topic of the survey (Chang and Krosnick in press). However, in that study, the questionnaire was entirely devoted to a single topic (i.e., politics), so participants may have chosen to complete the survey based on their interest in that topic.

---

each participating sample member is known, and the extent of non-response due to non-contact and to refusal are known. Put differently, the researcher knows (1) which people were invited to participate in the survey and did not, and (2) when individuals should have been invited to participate but were not (e.g., due to non-contact). We refer to such samples as "opt-out", because researchers invite particular known individuals to participate, and those individuals can decline to do so. In contrast, "opt-in" non-probability samples are built by methods whereby most sample members were not personally selected to be invited to participate (see Postoaca 2006: 67). Instead, invitations are offered broadly to an unknown group of people (e.g., visitors to a website), and such individuals have to choose to opt into the sample by taking active steps to contact the survey research organization, rather than opting out in response to a personal contact from such an organization.

To explore the generalizability of these findings, the present study collected data on a variety of topics via an RDD telephone survey, an Internet survey of a probability sample, and Internet surveys of seven non-probability opt-in samples of American adults. The estimates from each survey were then compared to benchmarks from official government records or high-quality federal surveys with very high response rates.

Three categories of measurements were compared: primary demographics, secondary demographics, and non-demographics. Primary demographics are those that were used by some of the survey firms to create weights or to define strata used in the process of selecting people to invite to complete the Internet surveys. Thus, explicit steps were taken by the survey firms to enhance the accuracy of these specific measures in the Internet surveys. Secondary demographics were not used to compute weights or to define sampling strata, so no procedures were implemented explicitly to assure their accuracy. Non-demographics included (1) factual matters on which we could obtain accurate benchmarks from official government records, and (2) behaviors that were measured in high-quality federal surveys with very high response rates.

The primary objective of this study was to learn whether non-probability samples of Internet volunteers (“opt-in samples”) yield measurements that are as accurate as those from probability samples. To investigate this question, the estimates from each survey were compared to benchmarks when no post-stratification weights were applied to the data. Next, post-stratification weights were applied, which allowed assessment of the extent to which these weights altered conclusions about the relative accuracy of the probability and non-probability samples’ data. We also explored whether any of the non-probability sample surveys was consistently more accurate than the others and how variation in accuracy across the non-probability samples compared with variability in accuracy across probability samples. We begin

below by describing the methods of data collection used, the methods of analysis implemented, and the study's results (for more details, see the online appendix: [INSERT WEBSITE]).<sup>2</sup>

## **Method**

### RESPONDENTS

Nine survey data collection firms each administered an identical questionnaire to a sample of approximately 1,000 American adults. All of the firms were widely recognized with established track records of carrying out these types of surveys.

The telephone survey involved conventional RDD methods to recruit and interview a probability sample. The probability sample Internet survey was conducted with members of a panel (of about 40,000 American adults) that was recruited via RDD methods. Individuals who wished to join the panel but did not have computers or Internet access at home were given them at no cost. A subset of the panel was selected via stratified random sampling to be invited to complete this survey via the Internet.

For six of the seven non-probability sample Internet surveys, invited individuals were selected via stratified random sampling from panels of millions of volunteers who were not probability samples of any population. The remaining company used "river" sampling, whereby pop-up invitations appeared on the screens of users of a popular Internet service provider. In three of the seven non-probability sample surveys, quotas were used to restrict the participating

---

<sup>2</sup> The online appendix provides descriptions of the firms' methods for collecting data; question wordings and response options; benchmark sources and calculations; missing data management techniques; a description of the weighting algorithm and the program that implemented it; a description and copy of the bootstrapping procedure used for statistical testing; all t-tests comparing the firms' average errors; t-tests assessing whether post-stratification improved accuracy for each survey; the variability of accuracy across the telephone surveys, probability sample Internet surveys, and non-probability sample Internet surveys; results obtained when using weights provided by the firms, when capping weights, and when dropping health status as a benchmark; and targets used to build post-stratification weights.

sample so that it would match the population in terms of some demographics. With all of the non-probability sample panels, members were not invited to complete a survey if they had completed more than a certain number of surveys already that month or if they had recently completed a survey on the same topic recently. A summary of the data collection methods appears in Table 1.

## MEASURES

Identical questions measuring primary demographics, secondary demographics, and non-demographics were asked in the same order within a long questionnaire administered by each survey firm.

*Primary and secondary demographics.* Primary demographics included sex, age, race/ethnicity, education, and region of residence. Secondary demographics included marital status, total number of people living in the household, employment status, number of bedrooms in the home, number of vehicles owned, home ownership, and household income.<sup>3</sup>

*Non-demographics.* Six questions asked respondents about frequency of smoking cigarettes, whether they have ever had 12 drinks of alcohol during their lifetimes, the average number of drinks of alcohol they have on days when they drink, ratings of quality of their health, and possession of a U.S. passport and a driver's license.

## BENCHMARKS

The U.S. Department of State provided the number of passports held by American adults.<sup>4</sup> The U.S. Federal Highway Administration provided the number of driver's licenses held

---

<sup>3</sup> Income was used in stratification during the process of selecting panel members to invite to complete one of the opt-in surveys. Because the other eight surveys did not use income for stratification, we treat income as a secondary demographic.

<sup>4</sup> The total number of U.S. passports held by Americans aged 16 and over as of May of 2005 was

by American adults.<sup>5</sup> Large government surveys with response rates of over 80% were used to obtain the remaining benchmarks. The primary and secondary demographics benchmarks were taken from the 2004 CPS Annual Social and Economic (ASEC) supplement and the 2004 American Community Survey (ACS).<sup>6</sup> Non-demographic benchmarks for cigarette smoking, alcohol consumption, and quality of health came from the 2004 NHIS.

#### WEIGHTING THE TELEPHONE SURVEY SAMPLE FOR UNEQUAL PROBABILITY OF SELECTION

To adjust the telephone survey sample for unequal probabilities of selection, a weight was constructed using the number of non-business landlines that could reach the household and the number of adults living in the household.<sup>7</sup>

---

obtained via personal communication from an official in the U.S. Department of State and was divided by the total population of Americans aged 16 and older in 2005 to obtain a percentage. This was the only benchmark on passports available from the U.S. Department of State and does not match the surveys in two regards: all but one of the surveys were conducted during 2004, whereas the State Department information is from 2005, and the surveys collected data only from individuals age 18 and older.

<sup>5</sup> The total number of driver's licenses held by persons aged 18 and older in the United States in 2004 was obtained from the U.S. Federal Highway Administration's website (<http://www.fhwa.dot.gov/policy/ohpi/hss/hsspubs.htm>) and was divided by the total population of Americans aged 18 and older in 2004 to obtain a percentage.

<sup>6</sup> Because one of the non-probability samples' data were collected in early 2005, we tested whether that sample's accuracy appeared to be better when compared to benchmarks collected in 2005, and it did not. We therefore compared all surveys to benchmarks collected in 2004.

<sup>7</sup> Members of the probability sample Internet survey firm's panel were randomly selected to be invited to complete our survey, with unequal probabilities. These probabilities were directly proportional to the number of adults living in the panel member's household and were inversely proportional to the number of non-business telephone landlines that could reach the panel member's household at the time of recruitment. During recruitment, telephone numbers were over-sampled if a mailing address could be obtained for them (to allow mailing an advance letter before recruitment) and if they were located in areas served by MSN-TV or were located in areas with high densities of racial minorities, so these households had proportionally lower probabilities of selection to participate in our survey. Some panelists were recruited during a pilot phase that took place in only a few metropolitan areas, and these panelists had proportionally lower probabilities of selection than other panelists recruited later. Panelists who

## POST-STRATIFICATION

*Weights we constructed.* Because some survey companies did not provide post-stratification weights or did not explain how their post-stratification weights were computed, we constructed a set of weights using the same method for all survey firms (cf. the Report from the American National Election Study's [ANES] Committee on Optimal Weighting [DeBell and Krosnick 2009]; Battaglia et al. 2009). These weights maximized the match of the survey sample with the 2004 CPS ASEC supplement via raking using the following variables: race (3 groups), ethnicity (2 groups), census region (4 groups) a cross-tabulation of sex by age (12 groups), a cross-tabulation of sex by education (10 groups).<sup>8</sup>

*Weight computation.* A custom raking program was written in Stata (StataCorp 2007) to create a set of weights for each survey.<sup>9</sup> For the telephone survey, this process began with the weight to adjust for unequal probability of selection. For the other surveys, the process began with a weight of 1 for all cases. Then, in each of fifty iterations, the raking program modified the weights to move the sample closer to the true proportion within each cell.<sup>10</sup> The program included a relaxation parameter to avoid over-correction. After the program adjusted the weights

---

had completed more than the allowed number of surveys in a month were assigned a selection probability of zero. The probability of selection was also adjusted to eliminate discrepancies between the full panel and the population in terms of sex, race, age, education, and Census region (as gauged by comparison with the CPS). Therefore, no additional weighting was needed to correct for unequal probabilities of selection during the recruitment phase of building the panel.

<sup>8</sup> The ANES procedure suggests inspecting the marginal distributions for secondary demographics, such as marital status and number of people in the household, after weighting on primary demographics. If a discrepancy larger than 5 percentage points appears for a variable, weighting could be done using that variable as well. We did not implement this procedure because we used the secondary demographics as benchmarks to assess accuracy.

<sup>9</sup> A similar pattern of results was obtained when weights were created using the "survey" package in R (Lumley 2004).

<sup>10</sup> The weights changed only minimally after 50 iterations when more were allowed.

for each demographic group, the weights were adjusted so they had a mean of 1 before the next iteration was implemented.

The weights were capped at 5, as suggested by the ANES's Committee on Optimal Weighting (DeBell and Krosnick 2009) and others (Izrael, Battaglia, and Frankel 2009), in order to prevent any one respondent from having undue influence. The telephone survey sample's uncapped weights ranged from .03 to 8.30, and the probability sample Internet survey's uncapped weights ranged from .20 to 5.3. Leaving the weights uncapped allowed the probability samples to perfectly match the primary demographic benchmarks and had no noticeable effect on their accuracy for the other benchmarks.

The uncapped weights for the non-probability sample surveys ranged from .003 to 70, with an average highest weight across the samples of 30. For these samples, capping the weights at 5 made them slightly less accurate on the primary demographics, and some of them notably more accurate in terms of the secondary demographics and non-demographics. Because capping the post-stratification weights improved the accuracy of some non-probability samples in terms of secondary demographics and non-demographics and did not decrease the accuracy of any of those samples, weights were capped at each weighting iteration.

The procedure for handling missing information on the primary demographics is described in the online appendix. For all practical purposes, if a respondent failed to provide an answer to a primary demographic question, the weighting algorithm weighted that case using the questions for which the respondent did provide data.

*Weights provided by the survey firms.* Post-stratification weights to eliminate demographic discrepancies between the U.S. adult population and the samples of panel members who completed our surveys were provided by the firms for three surveys: the probability sample

telephone survey, the probability sample Internet survey, and one of the non-probability sample internet surveys. The probability sample Internet survey's weights adjusted for unequal probability of invitation. We found that the post-stratification weights provided by the survey firms never yielded as much accuracy as did the weights we constructed, so no results are reported using their weights.

## **Analysis**

### **ACCURACY OF THE FULL SAMPLES**

*Accuracy measure.* We assessed the accuracy of the various surveys by computing the deviations between the survey and the benchmark in terms of the proportion of respondents selecting the modal response for each variable in the benchmark data source (these categories are listed in column 1 of Table 3).<sup>11</sup>

The statistical significance of the difference between zero and each accuracy measure in each survey was gauged using a z-test of proportions. For each survey, the average absolute error was computed across all three categories (primary, secondary and non-demographic), and we tested the significance of the differences between pairs of surveys by bootstrapping standard errors for each survey's average absolute error and computing t-tests to compare these averages.

The above analyses were first conducted with no weights on the Internet survey data and unequal probability of selection weights on the telephone survey data. Then the analyses were repeated with post-stratification weights on all samples using only the secondary demographics and the non-demographics.<sup>12</sup> Finally, standard errors were bootstrapped for the difference

---

<sup>11</sup> When we computed the average absolute error across all response categories for each variable, we reached the same conclusions as are reported in the text about relative survey accuracy.

<sup>12</sup> When the full sample post-stratification weights were used in the bootstrap procedure, each bootstrap sample deviated from the secondary and non-demographic benchmarks partially

between the un-weighted and weighted average absolute errors, and t-tests were computed to assess whether the change in accuracy due to weighting was statistically significant.

#### VARIATION ACROSS SURVEYS

The present study's design allowed assessment of the variability in results across non-probability sample Internet surveys. But because only one telephone survey and one probability sample Internet survey were conducted, it was not possible to compute such variation across implementations of those methods. Therefore, supplementary analyses were conducted with additional data. Specifically, primary and secondary demographic data were obtained for six additional pre-existing RDD telephone surveys and six additional pre-existing probability sample Internet surveys, and they were compared to the benchmarks. This permitted assessment of the consistency of results across these two types of surveys, to complement the same sort of assessment that was possible with data from the main study's non-probability sample Internet surveys.

To select the telephone surveys, the iPoll database (maintained by the Roper Center at the University of Connecticut) was searched to identify all RDD telephone surveys of national samples of American adults conducted during June, July, and mid-August of 2004 (the period when most of our surveys were in the field). Surveys that asked respondents to report the number of telephone lines and the number of adults in the household were eligible for selection, so that weights to correct for unequal probability of selection could be calculated. Of the seven eligible surveys, six were randomly selected, and all provided data on nine demographics: sex,

---

because the weights were not built for that sub-sample, which introduced bias in estimates of the sample's deviation from the population. Therefore, a new set of weights was calculated to force each bootstrap sample to match the population in terms of primary demographics, and then benchmark accuracy was calculated for that bootstrap sample.

age, race, ethnicity, education, region, marital status, number of adults, and income.<sup>13</sup>

The firm that conducted our main study's probability sample Internet survey provided a list of all surveys they conducted during the same period in 2004. From that list of seven surveys, six surveys were randomly selected, and demographic data for those six surveys were provided to us.

## **Results**

### **PRIMARY DEMOGRAPHICS**

*Without post-stratification.* The probability samples provided the most accurate estimates for the primary demographics when not post-stratified. The telephone survey and the probability sample Internet survey had average absolute errors of 3.43 percentage points (hereafter marked as %) and 2.47%, respectively, which were not significantly different from one another (see row 1 of Table 2). All of the non-probability sample Internet surveys were significantly less accurate than the probability sample Internet survey in terms of primary demographics, and all but one of the non-probability sample Internet surveys were significantly less accurate than the telephone survey (see row 1 of Table 2).

*With post-stratification.* After post-stratification, all of the samples closely matched the primary demographic benchmarks (see rows 5, 10, 15, 20, 25 and 30 of Table 3). This suggests that the weighting procedure had the intended effect on the variables used to create the weights.

---

<sup>13</sup> The six iPoll telephone surveys measured demographics differently than was done in the present study's questionnaire, so the response categories used to compute error for age, number of people in the household, and income in these analyses were age 30-44, two adults in the household, and income between \$50,000 and \$75,000, respectively. Three of these six surveys were conducted by the same firm that conducted the main study's RDD survey, and the other three were conducted by another firm.

## ACCURACY ACROSS ALL BENCHMARKS

*Without post-stratification.* Without post-stratification, the telephone survey and the probability sample Internet survey were not significantly different from one another in terms of average accuracy for all 19 benchmarks (average absolute errors of 3.58% and 3.49%, respectively) and were both significantly more accurate than all of the non-probability sample Internet surveys (which ranged from 4.90% to 10.16%; see row 2 of Table 2).

*With post-stratification.* After post-stratification, accuracy for the secondary demographics and non-demographics was best for the telephone survey (2.92%), and slightly (but not significantly) less accurate (average absolute error 3.37%) for the probability sample Internet survey. The telephone and probability sample Internet survey were significantly or marginally significantly more accurate than each of the non-probability sample Internet surveys (see row 4 of Table 2; see also Figure 1).

As expected, post-stratification significantly increased the average accuracy of both probability sample surveys (compare rows 3 and 4 of Table 2). Post-stratification significantly increased accuracy for only two of the seven non-probability sample surveys, increased accuracy marginally significantly for a third, and *decreased* accuracy marginally significantly for a fourth. Post-stratification had no significant or marginally significant impact on accuracy for the remaining three non-probability sample surveys (compare rows 3 and 4 of Table 2).<sup>14</sup>

## OTHER ACCURACY METRICS

*Largest absolute error.* Other accuracy metrics also suggested that the probability sample surveys were more accurate than the non-probability sample surveys. For example,

---

<sup>14</sup> For two of these surveys, post-stratification yielded non-significant improvements in accuracy, and for the third, post-stratification yielded a non-significant decrease in accuracy.

another way to characterize a survey's accuracy is the error of the benchmark on which that survey was least accurate, which we call the survey's "largest absolute error." With no post-stratification, the probability sample surveys had the smallest "largest absolute errors" (11.71% for the telephone and 9.59% for the Internet; see row 9 of Table 2), and the non-probability sample Internet surveys all had larger "largest absolute errors," ranging from 13.23% to 40.48%. With post-stratification, the same was true – the probability samples had largest absolute errors of 8.94% and 8.42%, in contrast to the non-probability sample Internet surveys' largest errors ranging from 12.05% to 17.47%.

*Number of significant differences from benchmarks.* This same conclusion was supported by examining the percent of benchmarks from which each survey's estimates were significantly different ( $p < .05$ ; see rows 12 and 14 of Table 2). Without post-stratification, the telephone surveys' estimates were significantly different from the fewest benchmarks (53%). The probability sample Internet survey's estimates were significantly different from somewhat more benchmarks (63%), and the non-probability sample Internet surveys were significantly different from the same percent or more of the benchmarks (range: 58% to 89%). With post-stratification, however, the probability samples were more obviously superior: their estimates were significantly different from 31% and 38% of the benchmarks, respectively, whereas the non-probability sample Internet surveys were significantly different from between 69% and 77% of the benchmarks.

#### SUPERIORITY OF SOME NON-PROBABILITY SAMPLES?

The average accuracies examined thus far suggest that some non-probability samples were more accurate than others, but these differences were almost never statistically

significant.<sup>15</sup> Furthermore, it is essentially impossible to predict a non-probability sample survey's accuracy on one benchmark using its overall accuracy on all of the benchmarks.

Without post-stratification, the correlation between overall rank order of the surveys in terms of absolute error and the absolute error for each of the nineteen benchmarks ranged from  $-.69$  to  $.72$  and averaged  $.25$ . Similarly, the correlation between average absolute error for each survey and absolute error for each of the nineteen benchmarks ranged from  $-.96$  to  $.95$  and averaged  $.35$ . These two average correlations were similar when post-stratification was done ( $.21$  and  $.29$ , respectively). Thus, these results challenge the claim that some of the non-probability sample Internet surveys were consistently more accurate than others.

#### RELATION OF COMPLETION RATES TO ACCURACY

Although response rates have often been used to assess a survey's likely accuracy, an accumulating group of studies shows that higher response rates do not predict notably increased accuracy in surveys of probability samples (e.g., Curtin, Presser, and Singer 2005; Holbrook, Krosnick, and Pfent 2007; Keeter et al. 2000; Keeter et al. 2006; Merkle and Edelman 2002). The same conclusion is supported here in a comparison of the six non-probability sample surveys that have completion rates. Without post-stratification, the completion rate for a non-probability sample survey was correlated *negatively* with the survey's rank order in terms of its average absolute error ( $r = -.48$ ). The negative correlation was even stronger after post-stratification ( $r = -.60$ ). After removing non-probability sample survey 7, which was an outlier in terms of accuracy, the correlation after post-stratification was much weaker ( $-.22$ ). This disconfirms the

---

<sup>15</sup> Of 21 possible t-tests comparing pairs of non-probability sample Internet surveys' average absolute errors to one another (after post-stratification), only 3 were statistically significant ( $p < .05$ ) - slightly more than would be expected by chance alone. All three of those significant t-statistics indicated that non-probability sample Internet survey 7 was significantly less accurate than others of the non-probability sample Internet surveys.

claim that a higher completion rate is an indication of more accuracy in a non-probability sample survey.

#### CONSISTENCY OF ABSOLUTE ERROR RATES ACROSS SURVEYS

In addition to the probability sample surveys being more accurate than the non-probability sample surveys, the former were also more consistent in their accuracy. Without post-stratification, the average absolute error for the seven probability sample telephone surveys was 3.55% (for the primary demographics and some secondary demographics), with a standard deviation of 0.25%. The same average absolute error and standard deviation were 1.89% and 0.55% for the seven probability sample Internet surveys. In contrast, for the seven non-probability sample Internet surveys, these figures were 5.99% and 2.26%, respectively. Thus, the standard deviation for the non-probability sample surveys' average error was nine times larger than the telephone samples' standard deviation and four times larger than the probability sample Internet surveys' standard deviation. Overall, the probability samples were quite consistent with one another in terms of accuracy, whereas accuracy varied much more across non-probability sample surveys. So it is difficult to anticipate whether a non-probability sample Internet survey will be somewhat less accurate than a comparable probability sample surveys or substantially less accurate.

#### CONSISTENCY OF ABSOLUTE ERROR RATES WITHIN SURVEYS

Not only were the probability sample surveys more consistent in their average errors across surveys than were the non-probability samples, but the former were also more consistently accurate across benchmarks within a survey. Without post-stratification, the average standard deviation (across nine demographics) of the absolute error averaged 2.76% for the seven probability sample telephone surveys, 1.36% for the seven probability sample Internet surveys,

and 5.58% for the seven non-probability samples. Thus, it is easier to predict a probability sample's accuracy on one benchmark knowing its accuracy on another benchmark in a survey than to predict a non-probability sample's accuracy on one benchmark knowing its accuracy on another benchmark in the survey.

## **Discussion**

This detailed investigation leads to the following conclusions:

- (1) Probability sample surveys done by telephone or the Internet were consistently highly accurate across a set of demographics and non-demographics, especially after post-stratification with primary demographics (average absolute errors of secondary demographics and non-demographics = 2.93 and 3.36 percentage points, respectively, for telephone and Internet).
- (2) Non-probability sample surveys done via the Internet were always less accurate, on average, than probability sample surveys (average absolute errors of secondary demographics and non-demographics = 5.28 percentage points) and were less consistent in their level of accuracy. Thus, the accuracy of any one measure in a non-probability sample survey is of limited value for inferring the accuracy of other measures in such surveys.
- (3) Post-stratification with demographics sometimes improved the accuracy of non-probability sample surveys and sometimes reduced their accuracy, so this method cannot be relied upon to repair deficiencies in such samples.
- (4) Although one of the non-probability sample surveys was strikingly and unusually inaccurate, the rest were roughly equivalently inaccurate, challenging the claim that optimizing methods of conducting non-probability sample Internet

surveys can maximize their accuracy.

(5) Completion rates of non-probability sample surveys were slightly negatively correlated with their accuracy, challenging the notion that a higher completion rate is an indication of higher accuracy.

(6) The response rates for the main study's probability sample surveys were 35.6% (for the telephone surveys) and 15.3% (for the Internet surveys), yet these surveys were quite accurate. This finding is consistent with the idea that if a probability sample is drawn and substantial efforts are made to interview as many respondents as possible, even large departures of response rates from those achievable by the most effective methods do not necessarily portend low levels of accuracy.

(7) Survey accuracy improved more using the weights computed using recommendations from the American National Election Study's Committee on Optimal Weighting (DeBell and Krosnick 2009) than using the weights provided by the survey firms, suggesting that the former may merit wider use.

Conclusion (1) is useful because probability sample surveys routinely come under attack, being accused of inaccuracy (e.g., Kellner 2004). Such assertions may sometimes be motivated because a survey's result is not in keeping with what an individual or organization wants to believe is true of the population and may sometimes be motivated by organizations that wish to unseat probability sampling methodology with an alternative approach to social measurement. It is rarely possible to evaluate the credibility of such assertions, because benchmarks to assess accuracy are rarely available. Therefore, this investigation's intentional inclusion of measures

suitable to comparison with benchmarks made this sort of comparison possible and yielded reassuring conclusions about probability sample surveys.

Conclusion (2) should come as no surprise, because no theory provides a rationale whereby samples generated by non-probability methods would necessarily yield accurate results. Because we saw substantial and unpredictable accuracy in a few of the many assessments we made with such surveys, it is possible to cherry-pick such results to claim that non-probability sampling can yield veridical measurements. But a systematic look at a wide array of benchmarks documented that such results are the exception rather than the rule.

The evidence reported here in this regard complements past studies, such as that done by Roster et al. (2004), who compared an RDD telephone survey to an Internet survey of a non-probability sample from the same geographic region. They found numerous statistically different and sizable differences in demographic and non-demographic variables. Schonlau, Asch, and Du (2004, who collected data in California) and Sparrow (2006, who collected data in the United Kingdom) reported similar comparisons that yielded equivalent results. However, from such evidence, it is impossible to tell which data collection method yielded the more accurate results. The present study suggests that in general, telephone survey data are likely to be more accurate than non-probability sample Internet surveys.<sup>16</sup>

---

<sup>16</sup> Loosveldt and Sonck (2008) found numerous sizable differences between the results of a non-probability sample Internet survey and those of a probability sample face-to-face survey in Belgium, as did Faas and Schoen (2006) with data from Germany. And again, these investigators did not provide evidence on which survey was the more accurate. Malhotra and Krosnick's (2007) study comparing a probability sample face-to-face survey with a non-probability sample Internet survey is helpful in this context, because it showed that the face-to-face sample's results were consistently more accurate.

The present study's evidence of the superiority of probability sampling is in keeping with the evidence reported by Chang and Krosnick (in press) but expands upon that evidence in two regards. First, Chang and Krosnick's (in press) data were collected in 2000, when Internet surveying was in its infancy. Although non-probability sample Internet surveys may have improved over the years, the present study suggests that they did not achieve accuracy levels equivalent to those of probability samples as of 2004. Second, Chang and Krosnick (in press) evaluated a survey that was exclusively focused on politics, whereas the present investigation focused on a survey addressing a wide array of different topics. So the conclusions reported by Chang and Krosnick appear to replicate in other types of surveys. Third, Chang and Krosnick (in press) evaluated just one RDD telephone survey, one probability sample Internet survey, and one non-probability sample Internet survey, whereas the present investigation examined many of each and found evidence supporting the same conclusion. Therefore, the present conclusions appear to generalize across survey data collection firms and somewhat different implementations of each method.

The present study found that post-stratification of probability samples consistently increased their accuracy, whereas such weighting did not consistently improve the non-probability samples. The logic of such weighting hinges on the assumption that the members of under-represented groups from whom a researcher has collected data will provide answers mirroring the answers that would have been obtained if more individuals in such groups had been interviewed. So perhaps with non-probability sampling, interviewed members of under-represented subgroups do not resemble non-interviewed members of such a groups as closely as occurs with probability sampling. For example, if young, African-American, highly educated males were under-represented in a non-probability sample Internet survey, the young, African-

American, highly educated males who did participate may not have closely resembled those who did not. As a result, weighting up the participating members of this group may have increased error, rather than decreasing it. Resonating with this logic, Schonlau et al. (2009), Couper et al. (2007), and Loosveldt and Sonck (2008) showed that weighting and matching did not eliminate discrepancies between a population and the subset of that population that had Internet access.

Advocates of non-probability sample surveys sometimes assert that inadequacies in sampling and participation can be corrected by suitable adjustments to be implemented post data collection. And some firms that sell such data sometimes say that they have developed effective methods to do so. The evidence reported here challenges that assertion in two ways: (1) among the non-probability sample surveys, none emerged as superior to the others, and (2) the sizes of errors observed within such surveys were not consistent. Therefore, it appears that currently implemented corrective strategies are not fully effective.

The evidence here that higher completion rates were not associated with more accuracy in the non-probability sample surveys is consistent with the growing body of evidence supporting the same finding in studies of probability sample surveys (e.g., Curtin et al. 2005; Holbrook, Krosnick, and Pfent 2007; Keeter et al. 2000; Keeter et al. 2006; Merkle and Edelman 2002). Those advocating the use of non-probability samples sometimes assert that established methodologies (such as probability sampling) are undermined by inevitably low response rates (e.g., Gosling et al. 2004: 99; Kellner 2004). The evidence reported here showing no relation of completion rates to accuracy across non-probability samples reinforces the notion that such criticisms are misdirected.

The present study does have limitations, some involving no small dose of irony. First, this study examined only a limited set of benchmarks, including demographics and non-

demographics addressing cigarette smoking, alcohol consumption, health quality, and passport and driver's license possession. This list goes beyond the variables that have been examined in past studies of survey accuracy, but it is not a random sample of measures from a universe of all possible measures. The evidence reported here that random sampling yields more generalizable results suggests that random sampling of measures would permit the same increase in confidence when generalizing conclusions. Therefore, it would be useful to conduct investigations in the future with an expanded list of criteria to assess the generalizability of the present study's findings. Perhaps it would be possible to define a population of eligible measures and randomly sample from it. But if not, investigations such as the present one can be conducted with convenience samples of measures to provide a basis for further confidence in general conclusions, even if they will not be definitive.

Second, this study focused on non-probability sample Internet surveys conducted by a set of seven firms, and these firms were not chosen randomly from the population of such firms. Rather, they were selected because of their high visibility in the industry. The evidence documented here using data from seven firms is suggestive about the likely accuracy of data that is collected by other companies. But in the absence of random sampling of companies, we must be cautious about generalizing from these results to all such companies. Ideally, future studies of this sort will involve sufficiently substantial budgets to allow random sampling of survey firms for participation.

Another limitation of the present study is that only one RDD telephone survey and one probability sample Internet survey were commissioned. One might worry that these two surveys are unrepresentative in their accuracy, and analyses of more such surveys would have yielded evidence of lower and highly variable accuracy. In fact, however, the variability in average

absolute errors across seven probability sample telephone surveys and across seven probability sample Internet surveys was quite small, relative to the variability across the non-probability samples. Moreover, both the probability sample telephone survey and the probability sample Internet survey that were the focus of the present benchmark comparisons turned out to be the *least* accurate among the seven probability sample telephone surveys and seven probability sample Internet samples examined in the supplementary analysis. If anything, the present study appears to understate the superiority of probability samples relative to non-probability samples. This suggests that the conclusions of the present study may generalize across probability sample surveys done with best practices methodology (as was true for the studies commissioned for the present study).

All of the analyses reported here presume that the benchmarks used to assess survey accuracy were themselves accurate. Yet most of those benchmarks were obtained from surveys, which no doubt contain some error. However, because those surveys had extremely high response rates (so non-response bias was likely to have been quite small) and primarily involved face-to-face interviewing (which has been shown to yield the most accurate reports; Holbrook, Green, and Krosnick 2003) and very large samples, there is reason to have confidence in them.

One might imagine that the use of human interviewers in the benchmark surveys and in the main study's telephone survey might introduce correlated measurement error that would create an illusion of similarity between them. Specifically, the telephone survey's results might match those of the benchmarks more closely than did the probability sample Internet survey because the benchmark and telephone surveys' data both contained the same sorts of systematic errors. However, we found no such trend in the present data. The telephone survey matched the benchmarks more closely than the probability sample Internet survey about as often (with post-

stratification; 7 out of 13 benchmarks, or 54%) as the probability sample Internet survey outdid the telephone survey in terms of accuracy (6 out of 13 benchmarks, or 46%).

One specific type of bias that might be shared between the surveys involving human interviewers is social desirability response bias. Chang and Krosnick (in press) reported results suggesting that such bias is greater in telephone surveys than in Internet surveys, and Holbrook, Green, and Krosnick (2003) reported evidence that such bias is greater in telephone surveys than in face-to-face surveys. Such bias seems unlikely to have contaminated the CPS measurements of factual matters or reports of having passports or drivers' licenses, and direct tests suggest that adults' reports of cigarette smoking and alcohol consumption are not subject to social desirability response bias (Aguinis, Pierce and Quigley 1993, 1995). But some investigators have speculated that reports of health quality may be subject to such bias, encouraging people to overstate their health (Adams et al. 1999; Adams et al. 2005), though we know of no direct test of this presumption.

Consistent with this logic, the telephone survey respondents in the present study reported better average health than did the probability sample Internet survey respondents (see also Schonlau et al. 2004, for a similar result when comparing an RDD survey to a non-probability sample Internet survey). 21.74% of telephone respondents said their health was excellent, whereas 13.50% of the probability sample Internet survey respondents said so, a significant difference (post-stratification weighted; z-test of proportions,  $p < .001$ ). Likewise, average reported health quality was significantly higher in the telephone survey than in the probability sample Internet survey (Telephone  $M = 2.60$  on a scale from 1 [Poor] to 5 [Excellent]; Internet  $M = 2.44$ ),  $t(2128) = 3.55$ ,  $p < .05$ .

If the same social desirability pressures were operating in the NHIS face-to-face

interviews, this could have caused an illusion of superior accuracy for the present telephone survey. But when we analyzed our data dropping the health quality benchmark, we reached the same conclusions about relative accuracy of the nine surveys we examined. So the conclusions of this research do not seem likely to have been distorted by social desirability response bias in the benchmark surveys.

## **Conclusion**

In a recent book described as an “authoritative work on on-line panels” (Vankatesh 2006), Postoaca (2006) offered a number of interesting observations about opt-in sample surveys, including:

“To many an experienced researcher, random samples are the only correct samples, being the only samples where each member of the sampling frame has an equal probability of being sampled. ... Experience shows, however, that the gap between theory and practice is not insignificant. ... It is well known that offline random samples are prohibitively expensive, while in practice impossible to get online (p. 6).”

“Of course, the quality of the final sample depends on the response rate, but market researchers who operate online using access panels can and do control the response rates without much difficulty (p. 6).”

“In terms of sampling, online market research is still considered risky. The online population is not representative for the overall population. Actually neither is the CATI population representative (p. 6).”

“Online access panel recruitment methods account for much of the differences in quality between online access panels (p. 69).”

Our findings discredit these assertions. Contrary to Postoaca’s (2006) speculations, we

found that (1) it is possible to conduct accurate probability sample Internet surveys and accurate CATI surveys, (2) response rates were not related to survey accuracy, and (3) accuracy did not vary notably across most non-probability sample Internet surveys, despite the use of different sample recruitment and panel maintenance procedures. Speculations such as Postoaca's (2006) have often formed the foundation for claims that non-probability sample Internet surveys yield data that are as accurate or more accurate than those produced by probability sample surveys (see, e.g., Crampton 2007; Zogby 2007), but we find no support for these claims here.

But Postoaca (2006) was quite humble in introducing his book:

“Though it hinges on ‘statistical evidence’, our enterprise will not be one of presenting clear statistical evidence; rather we will use statistical data to support a point of view that is put forward for further negotiation. We are writing this book to start a debate. (p. 3).”

The present paper offers statistical evidence to move this debate forward.

More generally, the present investigation suggests that the foundations of statistical sampling theory are sustained by actual data in practice. Probability samples, even ones without especially high response rates, yield quite accurate results. In contrast, non-probability samples are not as accurate, and are sometimes strikingly inaccurate. Because it is difficult to predict when such inaccuracy will occur, and because probability samples manifested consistently high accuracy, researchers interested in making accurate measurements can continue to rely on probability sampling with confidence.

This is not to say that non-probability samples have no value. They clearly do have value. Indeed, a huge amount of tremendously useful social science has been conducted during the last 5 decades with what are obviously highly unrepresentative samples of participants: college students who are required to participate in studies in order to fulfill course requirements

(e.g., Henry 2008; Sears 1986). However, researchers conducting such studies have usually not set out to document the distributions of variables or the magnitudes of associations between variables in the population (see, e.g., Petty and Cacioppo 1996). Rather, these studies were mostly intended to assess whether two variables were related to one another along the lines that theory anticipated, regardless of the magnitude of that association. The continued use of non-probability samples for such a purpose seems quite reasonable. But if a researcher's goal is to document the frequency distribution of a variable in a population accurately (or at least consistently within some margin of reasonable accuracy), non-probability sample surveys seem less suited to that goal than probability sample surveys.

### References

- Adams, Swann Arp, Charles E. Matthews, Cara B. Ebbeling, Charity G. Moore, Joan E. Cunningham, Jeanette Fulton, and James R. Hebert. 2005. "The Effect of Social Desirability and Social Approval on Self-reports of Physical Activity." *American Journal of Epidemiology* 161: 389-398.
- Adams, Alyce, Stephen Soumerai, Jonathan Lomas, and Dennis Ross-Degnan. 1999. "Evidence of Self-Report Bias in Assessing Adherence to Guidelines." *International Journal of Quality Health Care* 11: 187-192.
- Aguinis, Herman, Charles A. Pierce, and Brian M. Quigley. 1993. "Conditions Under Which a Bogus Pipeline Procedure Enhances the Validity of Self-Reported Cigarette Smoking: A Meta-Analytic Review." *Journal of Applied Social Psychology* 23: 352-373.
- Aguinis, Herman, Charles A. Pierce, and Brian M. Quigley. 1995. "Enhancing the Validity of Self-Reported Alcohol and Marijuana Consumption Using a Bogus Pipeline Procedure: A Meta-Analytic Review." *Basic and Applied Social Psychology* 16: 515-527.
- Battaglia, Michael P., David Izrael, David C. Hoaglin, and Martin R. Frankel. 2009. "Practical Considerations in Raking Survey Data." *Survey Practice: Practical Information for Survey Researcher*. June. Retrieved on August 1, 2009 (<http://surveypractice.org/2009/06/29/raking-survey-data/>).
- Carbone, Enrica. 2005. "Demographics and Behavior." *Experimental Economics* 8:217-232.
- Chang, LinChiat, and Jon A. Krosnick. In press. "National Surveys via RDD Telephone Interviewing vs. the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly*.

- Couper, Mick P., Arie Kapteyn, Matthias Schonlau, and Joachim Winter. 2007. "Noncoverage and Nonresponse in an Internet Survey." *Social Science Research* 36(1): 131-148.
- Crampton, Thomas. 2007. "About Online Surveys, Traditional Pollsters Are: (C) Somewhat Disappointed." *New York Times*, May 31. Retrieved on August 11, 2009 (<http://www.nytimes.com/2007/05/31/business/media/31adco.html?pagewanted=print>).
- Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64:413-428.
- DeBell, Matthew, and Jon A. Krosnick. 2009. "Weighting Plan for the American National Election Studies." *American National Election Studies Technical Report*. Ann Arbor, Michigan.
- Faas, Thorsten, and Harald Schoen. 2006. "Putting a Questionnaire on the Web is Not Enough – A Comparison of Online and Offline Surveys Conducted in the Context of the German Federal Election 2002." *Journal of Official Statistics* 22(2): 177-190.
- Gosling, Samuel D., Simine Vazire, Sanjay Srivastava, and Oliver P. John. 2004. "Should We Trust Web-Based Studies? A Comprehensive Analysis of Six Preconceptions About Internet Questionnaires." *American Psychologist* 59(2): 93-104.
- Henry, Peter J. 2008. "College Sophomores in the Laboratory Redux: Influences of a Narrow Data Base on Social Psychology's View of the Nature of Prejudice." *Psychological Inquiry* 19: 49-71.
- Holbrook, Allyson, Melanie Green, and Jon A. Krosnick. 2003. "Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias." *Public Opinion Quarterly* 67: 79-125.

- Holbrook, Allyson, Jon A. Krosnick, and Alison Pfent. 2007. "The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms." In *Advances in Telephone Survey Methodology*, ed. James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith de Leeuw, Lilli Japiec, Paul J. Lavrakas, Michael W. Link and Roberta L. Sangster. New York: Wiley-Interscience.
- Izrael, David, Michael P. Battaglia, and Martin R. Frankel. 2009. "Extreme Survey Weight Adjustment as a Component of Sample Balancing (a.k.a. Raking)." *Proceedings from the Thirty-Fourth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., Paper 247.
- Kaptyen, Arie, Jim Smith, and Arthur van Soest. 2007. "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands." *American Economic Review* 97: 461-473.
- Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 64:125-148.
- Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70: 759-779.
- Kellner, Peter. 2004. "Can Online Polls Produce Accurate Findings?" *International Journal of Market Research* 46: 3-23.
- Lerner, Jennifer S., Roxana M. Gonzalez, Deborah A. Small, and Baruch Fischhoff. 2003. "Effects of Fear and Anger on Perceived Risks of Terrorism: A National Field Experiment." *Psychological Science* 14: 144-150.

- Loosveldt, Geert, and Nathalie Sonck. 2008. "An Evaluation of the Weighting Procedures for an Online Access Panel Survey." *Survey Research Methods* 2(2): 93-105.
- Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9(8): 1-19.
- Malhotra, Neil, and Jon A. Krosnick. 2007. "The Effect of Survey Mode and Sampling on Inferences About Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples." *Political Analysis* 15: 286-324.
- Malhotra, Neil, and Alexander G. Kuo (2008). "Attributing Blame: The Public's Response to Hurricane Katrina." *The Journal of Politics* 70: 120-135.
- Merkle, Daniel, and Murray Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." Pp 243-258 in Robert Groves, Don Dillman, John Eltinge, and Roderick Little (Eds.), *Survey Nonresponse*. New York: Wiley.
- Moskalenko, Sophia, and Clark McCauley. 2009. "Measuring Political Mobilization: The Distinction Between Activism and Radicalism." *Terrorism and Political Violence* 21: 239-260.
- Petty, Richard E., and John T. Cacioppo. 1996. "Addressing Disturbing and Disturbed Consumer Behavior: Is It Necessary to Change the Way We Conduct Behavioral Science?" *Journal of Marketing Research* 33:1-8.
- Postoaca, Andrei. 2006. *The Anonymous Elect*. Berlin, Germany: Springer.
- Roster, Catherine A., Robert D. Rogers, Gerald Albaum, and Darin Klein. 2004. "A Comparison of Response Characteristics from Web and Telephone Surveys." *International Journal of Market Research* 46: 359-373.

- Schonlau, Matthias, Beth J. Asch, and Can Du. 2003. "Web Surveys as Part of a Mixed Mode Strategy for Populations That Cannot be Contacted by E-mail." *Social Science Computer Review* 21: 218-222.
- Schonlau, Matthias, Kinga Zapert, Lisa P. Simon, Katherine H. Sanstad, Sue M. Marcus, John Adams, Mark Spranca, Hongjun Kan, Rachel Turner, and Sandra H. Berry. 2004. "A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey." *Social Science Computer Review* 22: 128-138.
- Schonlau, Matthias, Arthur van Soest, Arie Kapteyn, and Mick Couper. 2009. "Selection Bias in Web Surveys and the Use of Propensity Scores." *Sociological Methods and Research* 37: 291-318.
- Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51: 515-530.
- Skitka, Linda J., and Christopher W. Bauman. 2008. "Moral Conviction and Political Engagement." *Political Psychology* 29: 29-54.
- Smith, Tom W. 2003. "An Experimental Comparison of Knowledge Networks and the GSS." *International Journal of Public Opinion Research*, 15: 167-179.
- Sparrow, Nigel. 2006. "Developing Reliable Online Polls." *International Journal of Market Research* 48: 659-680.
- StataCorp. 2007. "Stata Statistical Software: Release 10." College Station, TX: StataCorp LP.
- Vankatesh, Alladi. 2006. Endorsement on the back cover of Postoaca, Andrei. 2006. *The Anonymous Elect.* Berlin, Germany: Springer.

Zogby, Jonathan. 2007. "The New Polling Revolution: Opinion Researchers Overcome their Hang-ups with Online Polling." *Campaigns and Elections*, May: 16-19.

**Table 1.** Sample Description Information for Nine Surveys

Survey	Invitations	Responses	Response/ Completion Rate	Field Dates	Unequal Probability of Invitation?	Quota Used?	Incentives Offered
<u>Probability Samples</u>							
Telephone	2,513	966	35.6% <sup>1</sup>	June - November, 2004	N	N	\$10 (for Nonresponses)
Internet	1,533	1,175	15.3% <sup>2</sup>	June - July, 2004	Y	N	Points; Free Internet access; sweepstakes
<u>Non-Probability Samples</u>							
1	11,530	1,841	16%	June, 2004	Y	N	Points; Sweepstakes
2	3,249	1,101	34%	February, 2005	Y	N	Sweepstakes
3	50,000	1,223	2%	June, 2004	Y	Y	Sweepstakes
4	9,921	1,103	11%	June, 2004	Y	N	Sweepstakes
5	14,000	1,323	9%	August, 2004	Y	N	None
6	Unknown	1,137	Unknown	June, 2004	N	Y	Internet bill credit; frequent flier miles
7	2,123	1,129	53%	July, 2004	Y	Y	\$1

<sup>1</sup>AAPOR RR3<sup>2</sup>AAPOR CRR1

**Table 2.** Overall Accuracy Metrics for Probability and Non-Probability Sample Surveys, Without Post-stratification and With Post-Stratification

Evaluative Criteria	Probability Sample Surveys		Non-probability Sample Internet Surveys						
	Telephone	Internet	1	2	3	4	5	6	7
Average absolute error									
Primary demographics									
Without Post-stratification	3.43% <sup>b</sup>	2.47%	4.14% <sup>b</sup>	4.96% <sup>ab</sup>	6.44% <sup>ab</sup>	6.35% <sup>ab</sup>	7.01% <sup>ab</sup>	6.05% <sup>ab</sup>	12.82% <sup>ab</sup>
All benchmarks									
Without Post-stratification	3.58%	3.49%	4.90% <sup>ab</sup>	5.53% <sup>ab</sup>	6.17% <sup>ab</sup>	5.28% <sup>ab</sup>	5.51% <sup>ab</sup>	5.59% <sup>ab</sup>	10.16% <sup>ab</sup>
Secondary and non-demographics									
Without Post-stratification	3.64%	3.96%	5.25% <sup>ab</sup>	5.79% <sup>ab</sup>	6.05% <sup>ab</sup>	4.79% <sup>a</sup>	4.81% <sup>a</sup>	5.38% <sup>ab</sup>	8.93% <sup>ab</sup>
With post-stratification	2.92%	3.37%	4.59% <sup>ab</sup>	5.23% <sup>ab</sup>	4.53% <sup>a</sup>	5.51% <sup>ab</sup>	5.29% <sup>ab</sup>	5.01% <sup>ab</sup>	6.86% <sup>ab</sup>
Rank: Average absolute error									
Primary demographics									
Without Post-stratification	2	1	3	4	7	6	8	5	9
All benchmarks									
Without Post-stratification	2	1	3	6	8	4	5	7	9
Secondary and non-demographics									
Without Post-stratification	1	2	5	7	8	3	4	6	9
With post-stratification	1	2	4	6	3	8	7	5	9
Largest absolute error									
All benchmarks									
Without Post-stratification	11.71%	9.59%	17.99%	13.23%	15.55%	15.25%	15.13%	15.97%	40.48%
Secondary and non-demographics									
Without Post-stratification	11.71%	9.59%	14.68%	12.12%	13.03%	14.80%	13.70%	15.97%	20.04%
With post-stratification	8.94%	8.42%	14.75%	12.05%	12.09%	15.21%	13.14%	9.74%	17.47%
% Significant differences from benchmark									
All benchmarks									
Without Post-stratification	53%	63%	63%	68%	79%	58%	63%	63%	89%
Secondary and non-demographics									
Without Post-stratification	46%	69%	77%	77%	77%	54%	62%	62%	85%
With post-stratification	31%	38%	77%	69%	69%	77%	69%	69%	77%

<sup>a</sup>Significantly different from the telephone sample survey at  $p < .05$

<sup>b</sup>Significantly different from the probability sample Internet survey at  $p < .05$

**Table 3.** Accuracy Benchmark Comparisons for Probability and Non-Probability Sample Surveys: Primary Demographic, Secondary Demographic, and Non-Demographic Benchmarks, Without Post-stratification and With Post-Stratification

Benchmark comparison	Value	Probability Sample Surveys		Non-probability Sample Internet Surveys							
		Phone	Internet	1	2	3	4	5	6	7	
Female	51.68%										
Without Post-stratification											
%		55.40%*	50.57%	53.80%+	49.82%	48.73%*	50.73%	52.26%	49.61%	55.45%*	
Error		3.72	-1.11	2.12	-1.86	-2.95	-0.95	0.58	-2.07	3.77	
With post-stratification											
%		51.67	51.47	51.27	51.67	51.31	52.45	52.26	51.68	51.25	
Error		-0.01	-0.21	-0.41	-0.01	-0.37	0.77	0.58	0.00	-0.43	
Aged 38-47	20.83										
Without Post-stratification											
%		20.88	22.10	19.78	19.54	21.26	20.55	19.45	20.80	15.54*	
Error		0.05	1.27	-1.05	-1.29	0.43	-0.28	-1.38	-0.03	-5.29	
With post-stratification											
%		21.54	20.52	22.35	19.86	20.36	22.84	20.60	19.61	22.55	
Error		0.71	-0.31	1.52	-0.97	-0.47	2.01	-0.23	-1.22	1.72	
White Only	82.02										
Without Post-stratification											
%		78.30*	76.09*	79.58*	85.83*	89.62*	88.49*	66.89*	73.88*	41.54*	
Error		-3.72	-5.93	-2.44	3.81	7.60	6.47	-15.13	-8.14	-40.48	
With post-stratification											
%		82.02	82.02	82.02	82.01	82.03	81.98	81.72	82.02	79.59*	
Error		0.00	0.00	0.00	-0.01*	0.01	-0.04	-0.30	0.00	-2.43	
Non-Hispanic	87.62										
Without Post-stratification											
%		94.61*	90.86*	88.64	95.09*	96.65*	96.64*	94.78*	93.44*	90.19*	
Error		6.99	3.24	1.02	7.47	9.03	9.02	7.16	5.82	2.57	
With post-stratification											
%		87.62	87.62	87.62	87.63	87.62	87.67	87.88	87.62	88.42	
Error		0.00	0.00	0.00	0.01	0.00	0.05	0.26	0.00	0.80	
High school degree	31.75										
Without Post-stratification											
%		27.36*	31.41	13.76*	18.52*	16.20*	16.50*	19.03*	20.45*	16.60*	
Error		-4.39	-0.34	-17.99	-13.23	-15.55	-15.25	-12.72	-11.30	-15.15	
With post-stratification											
%		31.75	31.71	34.65*	31.80	34.75*	34.01	34.24+	31.75	31.84	
Error		0.00	-0.04	2.90	0.05	3.00	2.26	2.49	0.00	0.09	
South	35.92										
Without Post-stratification											
%		34.22	38.82*	36.13	38.01	39.03*	29.80*	40.99*	44.85*	26.25*	
Error		-1.70	2.90	0.21	2.09	3.11	-6.12	5.07	8.93	-9.67	
With post-stratification											
%		35.92	35.92	35.92	35.92	35.91	35.90	35.93	35.92	36.04	
Error		0.00	0.00	0.00	0.00	-0.01	-0.02	0.01	0.00	0.12	

Benchmark comparison	Value	Probability Sample Surveys		Non-probability Sample Internet Surveys						
		Phone	Internet	1	2	3	4	5	6	7
Married	56.50									
Without Post-stratification										
%		61.87*	59.82*	58.77*	59.93*	61.49*	56.82	58.15	53.71+	45.54*
Error		5.37	3.32	2.27	3.43	4.99	0.32	1.65	-2.79	-10.96
With post-stratification										
%		58.74	57.13	55.31	57.63	56.33	51.41*	49.46*	52.81*	52.26*
Error		2.24	0.63	-1.19	1.13	-0.17	-5.09	-7.04	-3.69	-4.24
2 people in household	33.84									
Without Post-stratification										
%		34.19	37.46*	41.50*	36.52+	39.98*	40.55*	38.25*	35.25	23.96*
Error		0.35	3.62	7.66	2.68	6.14	6.71	4.41	1.41	-9.88
With post-stratification										
%		32.43	34.60	38.13*	34.67	37.98*	36.47+	33.37	32.95	26.13*
Error		-1.41	0.76	4.29	0.83	4.14	2.63	-0.47	-0.89	-7.71
Worked last week	60.80									
Without Post-stratification										
%		60.58	61.69	63.12*	53.59*	67.05*	61.18	55.76*	60.00	63.29
Error		-0.22	0.89	2.32	-7.21	6.25	0.38	-5.04	-0.80	2.49
With post-stratification										
%		58.40	62.42	62.25	48.75*	60.24	60.12	53.10*	57.30*	60.04
Error		-2.40	1.62	1.45	-12.05	-0.56	-0.68	-7.70	-3.50	-0.76
3 bedrooms	43.38									
Without Post-stratification										
%		44.43	45.88+	43.56	46.14+	45.05	45.18	41.71	41.25	36.87*
Error		1.05	2.50	0.18	2.76	1.67	1.80	-1.67	-2.13	-6.51
With post-stratification										
%		44.71	45.16	43.57	48.18*	47.95*	45.10	38.33*	42.23	45.36
Error		1.33	1.78	0.19	4.80	4.57	1.72	-5.05	-1.15	1.98
2 vehicles	41.46									
Without Post-stratification										
%		41.09	45.53*	43.73+	44.41*	46.93*	41.82	42.03	40.18	41.50
Error		-0.37	4.07	2.27	2.95	5.47	0.36	0.57	-1.28	0.04
With post-stratification										
%		40.87	45.78*	44.07*	42.15	44.80*	37.93*	39.34	38.89+	44.70*
Error		-0.59	4.32	2.61	0.69	3.34	-3.53	-2.12	-2.57	3.24
Owns home	72.50									
Without Post-stratification										
%		78.75*	71.72	72.68	68.66*	71.71	71.18	69.32*	64.83*	52.46*
Error		6.25	-0.78	0.18	-3.84	-0.79	-1.32	-3.18	-7.67	-20.04
With post-stratification										
%		76.46*	70.12+	67.81*	66.59*	72.44	67.28*	66.67*	62.86*	63.44*
Error		3.96	-2.38	-4.69	-5.91	-0.06	-5.22	-5.83	-9.64	-9.06
HH income \$50K - \$59.9K	15.11									
Without Post-stratification										
%		14.05	23.26*	21.57*	23.00*	18.44*	19.96*	19.63*	19.52*	19.28*
Error		-1.06	8.15	6.46	7.89	3.33	4.85	4.52	4.41	4.17
With post-stratification										
%		14.72	22.14*	21.51*	22.51*	17.95*	20.20*	21.03*	19.60*	16.64
Error		-0.39	7.03	6.40	7.40	2.84	5.09	5.92	4.49	1.53

Benchmark comparison	Value	Probability Sample Surveys		Non-probability Sample Internet Surveys						
		Phone	Internet	1	2	3	4	5	6	7
Non-smoker	78.25									
Without Post-stratification										
%		76.63	74.91*	76.26*	68.85*	70.40*	73.73*	68.76*	70.55*	65.82*
Error		-1.62	-3.34	-1.99	-9.40	-7.85	-4.52	-9.49*	-7.70	-12.43
With post-stratification										
%		75.65+	74.08*	72.77*	68.19*	69.85*	69.35*	68.61*	69.75*	60.78*
Error		-2.60	-4.17	-5.48	-10.06	-8.40	-8.90	-9.64	-8.50	-17.47
Had 12 drinks in lifetime	77.45									
Without Post-stratification										
%		84.54*	87.04*	92.09*	89.57*	90.48*	92.25*	91.15*	88.89*	81.55*
Error		7.09	9.59	14.64	12.12	13.03	14.80	13.70	11.44	4.10
With post-stratification										
%		84.56*	85.87*	92.20*	87.87*	89.54*	92.66*	90.59*	87.19*	89.99*
Error		7.11	8.42	14.75	10.42	12.09	15.21	13.14	9.74	12.54
Has 1 drink on average	37.67									
Without Post-stratification										
%		43.46*	42.98*	40.22*	37.74	38.73	38.77	40.18	33.75*	22.07*
Error		5.79	5.31	2.55	0.07	1.06	1.10	2.51	-3.92	-15.60
With post-stratification										
%		39.62	40.41+	34.42*	39.47	38.19	32.48*	36.85	31.83*	28.97*
Error		1.95	2.74	-3.25	1.80	0.52	-5.19	-0.82	-5.84	-8.70
Health "Very good"	31.43									
Without Post-stratification										
%		33.64	37.91*	38.73*	40.45*	40.39*	38.91*	34.12	36.42*	42.93*
Error		1.71	5.98	6.80	8.52	8.46	6.98	2.19	4.49	11.00
With post-stratification										
%		33.44	38.84*	37.57*	39.09*	40.27*	39.65*	35.20*	33.60	39.53*
Error		1.51	6.91	5.64	7.16	8.34	7.72	3.27	1.67	7.60
Does not have a passport	78.50									
Without Post-stratification										
%		66.79*	75.19*	63.82*	70.21*	65.99*	65.36*	68.87*	62.53*	63.30*
Error		-11.71	-3.31	-14.68	-8.29	-12.51	-13.14	-9.63	-15.97	-15.20
With post-stratification										
%		69.56*	76.43+	72.12*	75.20*	68.68*	71.83*	71.18*	69.17*	66.92*
Error		-8.94	-2.07	-6.38	-3.30	-9.82	-6.67	-7.32	-9.33	-11.58
Has a driver's license	89.00									
Without Post-stratification										
%		93.77*	89.63	95.27*	95.10*	96.08*	95.00*	93.02*	94.97*	92.66*
Error		4.77	0.63	6.27	6.10	7.08	6.00	4.02	5.97	3.66
With post-stratification										
%		92.60*	88.06	92.36*	91.40*	93.06*	92.99*	88.50	93.18*	91.80*
Error		3.60	-0.94	3.36	2.40	4.06	3.99	-0.50	4.18	2.80

Note: All errors are deviations from the benchmark.

\*  $p < .05$ . +  $p < .10$ .

**Figure 1.** Average absolute errors for probability and non-probability sample surveys across thirteen secondary demographics and non-demographics, with post-stratification.

