Public Opinion Quarterly, Vol. 75, No. 4, Winter 2011, pp. 709–747

COMPARING THE ACCURACY OF RDD TELEPHONE SURVEYS AND INTERNET SURVEYS CONDUCTED WITH PROBABILITY AND NON-PROBABILITY SAMPLES

DAVID S. YEAGER*
JON A. KROSNICK*
LINCHIAT CHANG
HAROLD S. JAVITZ
MATTHEW S. LEVENDUSKY
ALBERTO SIMPSER
RUI WANG

Abstract This study assessed the accuracy of telephone and Internet surveys of probability samples and Internet surveys of non-probability samples of American adults by comparing aggregate survey results against benchmarks. The probability sample surveys were consistently more accurate than the non-probability sample surveys, even after post-stratification with demographics. The non-probability sample survey measurements were much more variable in their accuracy, both across measures within a single survey and across surveys with a single measure. Post-stratification improved the overall accuracy of some of the non-probability sample surveys but decreased the overall accuracy of others.

DAVID S. YEAGER is a PhD candidate at Stanford University, Stanford, CA, USA. JON A. KROSNICK is Frederic O. Glover Professor in Humanities and Social Sciences, Professor of Communication, Political Science, and (by courtesy) Psychology at Stanford University, Stanford, CA, USA, and is University Fellow at Resources for the Future. LINCHIAT CHANG is the owner of LinChiat Chang Consulting, LLC, San Francisco, CA, USA. HAROLD S. JAVITZ is Principal Scientist at SRI International, Inc., Menlo Park, CA, USA. MATTHEW S. LEVENDUSKY is Assistant Professor of Political Science at the University of Pennsylvania, Philadelphia, PA, USA. ALBERTO SIMPSER is Assistant Professor of Political Science at the University of Chicago, Chicago, IL, USA. RUI WANG is a PhD candidate at Stanford University, Stanford, CA, USA. The authors thank Norman Nie, who initially envisioned this project, collaborated in the early design of the data collection, and made the study possible; Douglas Rivers, for collaborating in the study design and data collection; SPSS, Inc., for funding the project; and Gary Langer, Arthur Lupia, Josh Pasek, Yphtach Lelkes, Neil Malhotra, Guarav Sood, and members of the Political Psychology Research Group at Stanford University for helpful suggestions. *Address correspondence to David Yeager, 271 Jordan Hall, Stanford University, Stanford, CA 94305, USA; e-mail: dyeager@stanford.edu; or to Jon Krosnick, 432 McClatchy Hall, 450 Serra Mall, Stanford University, Stanford, CA 94305, USA; e-mail: krosnick@stanford.edu.

doi: 10.1093/poq/nfr020 Advance Access publication October 5, 2011

Higher completion and response rates of the surveys were associated with less accuracy. Accuracy did not appear to change from 2004/2005 to 2009 for any of the methods, and these conclusions are reinforced by data collected in 2008 as well. These results are consistent with the conclusion that non-probability samples yield data that are neither as accurate as nor more accurate than data obtained from probability samples.

Since the 1940s, the gold standard for survey research has been to conduct primarily face-to-face interviews with a random sample of American adults. Surveys such as the Centers for Disease Control and Prevention's National Health Interview Survey (NHIS) and National Health and Nutrition Examination Survey (NHANES) have involved interviews with tens of thousands of randomly selected Americans with 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 survey research today, though it is still done regularly (e.g., by the American National Election Studies and the General Social Survey, and in developing nations). In the 1970s, random digit dialing (RDD) telephone surveys became very popular. And in recent years, Internet surveys have been conducted at increasing rates.

An accumulating body of evidence suggests that this latter shift may have some advantages. In a number of experiments, participants have been randomly assigned to complete a questionnaire either on a computer or via oral administration by an interviewer (e.g., Chang and Krosnick 2010; Link and Mokdad 2004, 2005; Rogers et al. 2005), or the same people have completed a questionnaire in both modes (e.g., Cooley et al. 2000; Ghanem et al. 2005; Metzger et al. 2000; see also Chatt et al. 2003). In general, these studies have found that computer data collection yielded higher concurrent validity, less survey satisficing, less random measurement error, and more reports of socially undesirable attitudes and behaviors than did data collected by interviewers (cf. Bender et al. 2007). Thus, computer administration appears to offer significant measurement advantages.

Many Internet surveys are being done these days with probability samples of the population of interest (e.g., Carbone 2005; Kapteyn, Smith, and van Soest 2007; Lerner et al. 2003; Malhotra and Kuo 2008; Moskalenko and McCauley 2009; Skitka and Bauman 2008). But most commercial companies that collect survey data in the United States 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 non-probability samples of people who were not systematically selected from a population using conventional sampling methods. For some such surveys, banner ads were placed on websites inviting people to sign up to do surveys regularly. In other instances, e-mail invitations were sent to large numbers of people whose e-mail addresses were 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 signed up to join an Internet survey "panel" and were later invited to complete questionnaires. For any given survey, a potential respondent's probability of selection from the panel is usually known, but his or her probability of selection from the general population of interest is not known.

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 the variables of interest in a study, or if they can be fully accounted for by making adjustments before or after data collection (with methods such as quotas, stratified random sampling from the panel, matching [Diamond and Sekhon 2005; Vavreck and Rivers 2008], post-stratification weighting [see, e.g., Dever, Rafferty, and Valliant 2008; see also Battaglia et al. 2009; Gelman 2007; Kish 1992], or propensity score weighting [Lee 2006; Lee and Valliant 2009; Schonlau et al. 2009; Taylor et al. 2001; Terhanian et al. 2001]), then the observed distributions of those variables in a non-probability sample should be identical to the distributions in the population. However, if these conditions do not hold, then survey results from non-probability samples may not be comparable to those that would be obtained from probability samples.

To date, numerous studies have compared probability samples interviewed via telephone to non-probability samples interviewed via the Internet. No studies have uncovered consistently equivalent findings across the two types of surveys, and many have found significant differences in the distributions of answers to demographic and substantive questions (e.g., Baker, Zahs, and Popa 2004; Berrens et al. 2003; Bethell et al. 2004; Braunsberger, Wybenga, and Gates 2007; Chang and Krosnick 2002, 2009; Crete and Stephenson 2008; Elmore-Yalch, Busby, and Britton 2008; Klein, Thomas, and Sutter 2007; Loosveldt and Sonck 2008; Niemi, Portney, and King 2008; Roster et al. 2004; van Ryzin 2008; Schillewaert and Meulemeester 2005; Schonlau et al. 2004; Sparrow 2006; Spijkerman et al. 2009; Taylor, Krane, and Thomas 2005).

In comparisons with RDD telephone surveys and face-to-face probability sample surveys, a few studies have found non-probability sample Internet surveys to yield less accurate measurements in terms of voter registration (Niemi et al. 2008; cf. Berrens et al. 2003), turnout (Chang and Krosnick 2009; Malhotra and Krosnick 2007; Sanders et al. 2007), candidate choice (Malhotra and Krosnick 2007; cf. Sanders et al. 2007), health (Bethell et al. 2004), and demographics (Chang and Krosnick 2009; Crete and Stephenson 2008; Malhotra and Krosnick 2007).[1]

[1] Braunsberger et al. (2007) reported the opposite finding: greater accuracy in a non-probability sample Internet survey than in a telephone survey documenting health insurance. However, they did not state whether the telephone survey involved pure random digit dialing; they said it involved "a random sampling procedure" from a list "purchased from a major provider of such lists" (p. 761). And Braunsberger et al. (2007) did not describe the source of their validation data in detail.

To supplement that literature, the present article assesses the accuracy of measurements on a variety of topics collected during 2004 and 2005 in an RDD telephone survey of American adults, an Internet survey of a national probability sample, and Internet surveys from seven non-probability samples. Accuracy was assessed by comparing the surveys' estimates to benchmarks from official government records or high-quality federal surveys with high response rates.

Three categories of variables were examined: 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 federal surveys with high response rates.

Estimates from each survey were first 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.

To permit comparison of the variability of accuracy within and across the three types of surveys, the surveys commissioned for this article were compared with additional RDD telephone surveys and probability sample Internet surveys commissioned by other organizations at about the same time as the surveys commissioned for this study. We explored whether higher response rates were associated with greater accuracy. Surveys commissioned for this article were compared with surveys conducted in 2009 by some of the same firms, to assess whether accuracy has improved or declined over time. We also analyzed data collected in 2008 by the Advertising Research Foundation (ARF), using a similar design to that employed by our commissioned studies.

Data

SURVEYS COMMISSIONED IN 2004/2005

Nine well-established survey data collection firms each administered the same questionnaire to a sample of American adults in 2004 or 2005. Identical written instructions given to all firms asked them to provide "1,000 completed surveys with a census-representative sample of American adults 18 years and older, residing in the 50 United States."

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.

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 computer 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 sample so that it would match the population in terms of some demographics. For six of the non-probability sample surveys, potential participants were not invited to complete the questionnaire 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. A summary of the data collection methods appears in table 1.

2004 ARCHIVAL SURVEYS

To select six additional RDD telephone surveys, the iPoll database (maintained by the Roper Center at the University of Connecticut) was searched to identify all such surveys of national samples of American adults in that archive conducted during June, July, and early August of 2004 (the period when most of the commissioned 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.[2] These surveys were in the field for an average of 3.8 days (Range: 3–8), and all had a lower AAPOR RR3 than the commissioned telephone survey.

The firm that conducted the commissioned probability sample Internet survey provided a list of all surveys they conducted of the general American adult public during the same period in 2004. From that list of seven surveys, six surveys were randomly selected, and demographic data for those surveys were obtained. These surveys were in the field for an average of 12.7 days (Range: 8–23), and their cumulative response rates were similar to that of the commissioned survey.

[2] Three of these six surveys were conducted by the same firm that conducted the commissioned RDD survey, and the other three were conducted by another firm.

Table 1. Sample Description Information for Nine Commissioned Surveys

Survey	Invitations	Responses	Response/ Completion Rate	Field Dates	Unequal Probability of Invitation?	Quota Used?	Incentives Offered
Probability Samples							
Telephone	2,513	966	35.6%[1]	June–November 2004[2]	N	N	$10 (For nonresponses)
Internet	1,533	1,175	15.3%[3]	June–July 2004	Y	N	Points; free Internet access; sweepstakes
Non-Probability Samples							
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	1,086	8%	August 2004	Y	N	None
6	Unknown	1,112	Unknown	June 2004	N	Y	Internet bill credit; frequent flier miles
7	2,123	1,075	51%	July 2004	Y	Y	$1

[1] AAPOR RR3.

[2] 81 percent of telephone respondents were interviewed within the first 30 days, and 95 percent were interviewed within the first 90 days.

[3] AAPOR CRR1.

2008 ADVERTISING RESEARCH FOUNDATION SURVEYS

In 2008, the Advertising Research Foundation (ARF) conducted a study dubbed "Foundations of Quality" (FoQ), in which 17 firms each conducted a non-probability sample Internet survey using the same questionnaire between May and October of 2008 (total N = about 100,000 respondents). Analysis results have been reported on some public websites, and we analyzed those data to generate statistics comparable to those generated with our commissioned surveys.[3]

2009 SURVEYS

Surveys conducted in 2009 were obtained from three of the firms that provided data for the commissioned surveys, and these surveys were compared to the same firms' 2004/2005 commissioned data. The 2009 probability and non-probability sample Internet surveys were commissioned for other purposes, and the 2009 RDD survey was selected randomly from among a set of surveys that the RDD firm did for other clients during 2009. These were the only three firms from which data collected in 2009 were available for analysis.

Measures

SURVEYS COMMISSIONED IN 2004/2005

Identical questions measuring primary demographics, secondary demographics, and non-demographics were asked in the same long questionnaire that was administered by each survey firm (see appendix 1 for questions and response options).[4] 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, homeownership, and household income.[5] Non-demographics included frequency of smoking cigarettes, consumption of at least 12 drinks of alcohol during their lifetimes, the average number of drinks of alcohol consumed on days when people drank, ratings of quality of health, and possession of a U.S. passport and a driver's license.

[3] http://blog.joelrubinson.net/2009/09/how-do-online-and-rdd-phone-research-compare-latest-findings/; http://regbaker.typepad.com/regs_blog/2009/07/finally-the-real-issue.html.

[4] Other questions were also asked, but trustworthy benchmarks could not be obtained for any of those variables, so they were not used to assess survey accuracy (for an explanation, see the online supplement at http://poq.oxfordjournals.org/).

[5] One of the non-probability sample Internet survey firms used income to stratify its panel when selecting respondents to invite for the survey commissioned from it. Because the other eight surveys did not use income for stratification, and this firm was no more accurate in terms of its income distribution than the other surveys, we treat income as a secondary demographic.

2004 ARCHIVAL SURVEYS

The archival surveys were evaluated using data on nine demographics: sex, age, race, ethnicity, education, region, marital status, number of adults, and income. The six telephone surveys obtained from iPoll measured demographics differently than was done in the commissioned surveys, so the response categories used to compute error for age, number of adults in the household, and income in these analyses were slightly different from those used for the commissioned surveys.

2008 ARF SURVEYS

The questions used to generate publicly available results for the full sample of over 100,000 respondents were homeownership, smoking 100 cigarettes in one's lifetime, current cigarette-smoking status, having a residential telephone, and owning a cell phone. The only benchmark question reported publicly with separate results for each of the 17 non-probability Internet firms (and thus the only question for which a standard deviation across firms could be calculated) assessed current cigarette-smoking status.

2009 SURVEYS

To compare the accuracy of the 2004 and 2009 probability sample Internet surveys, we examined all suitable variables measured in both surveys: sex, age, race, ethnicity, education, region, marital status, income, homeownership, household size, and work status. The same variables were available to compare the 2005 and 2009 non-probability sample Internet surveys. Comparing the 2004 and 2009 RDD telephone surveys was possible with a slightly different set of variables: sex, age, race, ethnicity, education, region, work status, and annual household income.

Benchmarks

The U.S. Department of State provided the number of passports held by American adults.[6] The U.S. Federal Highway Administration provided the number of driver's licenses held by American adults.[7] Large government surveys with response rates of over 70 percent were used to obtain the remaining benchmarks.

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

[7] The total number of driver's licenses held by people age 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.cfm) and was divided by the total population of Americans age 18 and older in 2004 to obtain the percentage.

The primary and secondary demographics benchmarks were taken from the 2004 Annual Social and Economic (ASEC) supplement to the Current Population Survey (response rate = 84 percent), and the 2004 American Community Survey (ACS; response rate = 92 percent).[8] Non-demographic benchmarks for cigarette smoking, alcohol consumption, and quality of health came from the 2004 National Health Interview Survey (NHIS; response rate = 72.4 percent).

Weights for Commissioned Surveys

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. Next, 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 nine surveys.[9] 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), and a cross-tabulation of sex by education (10 groups).[10]

Analysis Method

Survey accuracy was assessed by computing the difference between the proportion of respondents selecting the modal response for each variable in the benchmark data and the proportion of survey respondents giving that answer

[8] Because data from one of the non-probability samples 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.

[9] These weights were constructed following recommendations from the Report from the American National Election Study's (ANES) Committee on Optimal Weighting (DeBell and Krosnick 2009) and other sources (e.g., Battaglia et al. 2009). 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 can then be done using that variable as well. We did not implement this part of the procedure, because we used the secondary demographics as benchmarks to assess accuracy.

[10] In an initial contact letter, all firms were asked for survey weights, but only three firms provided post-stratification weights to eliminate demographic discrepancies between the U.S. population and the samples of respondents: 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. The firm that conducted non-probability sample Internet survey 1 provided weights that included a propensity score adjustment in addition to a post-stratification adjustment. In the analyses reported here, we chose to use the weights that yielded the most accurate survey results, which were the weights we constructed, and so no results are reported using the weights provided by the firms.

(the response categories are listed in column 1 of table 3).[11] We assessed the statistical significance of the difference between the benchmark data's measurement of the modal category and each survey's estimate of that category using standard errors from the benchmark surveys and the target surveys. Because the driver's license and passport benchmarks were not obtained from surveys, they were treated as measured without error. For each survey, the average absolute error was computed across all three categories (primary demographics, secondary demographics, and non-demographics), and we tested the statistical 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.

All of the above analyses were first conducted with no weights on the Internet survey data and weights to adjust for unequal probability of selection on the telephone survey data. The analyses were then repeated using post-stratification weights and examining only the secondary demographics and non-demographics. Finally, t-tests were computed to assess whether the change in accuracy due to weighting was statistically significant using bootstrapped standard errors. For more details on methods and analyses, see the supplementary data online at http://poq.oxfordjournals.org/.[12] For more information on past findings generated with the present study's commissioned surveys, see appendix 2.

Results

ACCURACY OF THE 2004/2005 COMMISSIONED SURVEYS

Primary Demographics

Without post-stratification. Without post-stratification, the probability samples provided the most accurate estimates of the primary demographics. The telephone survey and the probability sample Internet survey had average absolute errors of 3.29 and 1.96 percentage points, respectively, which were significantly different

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

[12] The online supplement provides descriptions of the firms' methods for collecting data; descriptions of the archival surveys; other questions included in the questionnaire that could not be used to assess accuracy; benchmark sources and calculations; missing data management techniques; descriptions of the data used and analyses conducted to assess the accuracy of surveys from 2009; a description of the weighting algorithm and the program that implemented it; a description and copy of the bootstrapping procedure used for statistical testing; 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; targets used to build post-stratification weights; and confidence intervals for the commissioned telephone survey's weighted measurement of benchmarked variables not used to create weights.

(i.e., p < .05) from each other (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.

Accuracy Across All Benchmarks

Without post-stratification. Without post-stratification, the telephone survey and the probability sample Internet survey were not significantly different from each other in terms of average accuracy for all 19 benchmarks (average absolute errors of 3.53 and 3.33 percentage points, respectively) and were both significantly more accurate than all of the non-probability sample Internet surveys (which ranged from 4.88 to 9.96 percentage points; 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 (2.90 percentage points), and slightly (but not significantly) less accurate for the probability sample Internet survey (average absolute error 3.40 percentage points). The telephone and probability sample Internet surveys were significantly or marginally significantly (i.e., p < .05, or, in one comparison, p < .10) more accurate than all 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).[13]

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, another way to characterize a survey's accuracy is the error of the variable 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 percentage points for the

[13] 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.

Table 2. Overall Accuracy Metrics for Commissioned Probability and Non-Probability Sample Surveys, Without Post-Stratification and with Post-Stratification

	Probability Sample Surveys		Non-Probability Sample Internet Surveys						
Evaluative Criteria	Telephone	Internet	1	2	3	4	5	6	7
Average percentage point absolute error									
Primary demographics									
Without post-stratification	3.29	1.96	4.08[b]	5.02[ab]	6.44[ab]	6.39[ab]	5.33[ab]	4.72[ab]	12.00[ab]
All benchmarks									
Without post-stratification	3.53	3.33	4.88[ab]	5.55[ab]	6.17[ab]	5.29[ab]	4.98[ab]	5.17[ab]	9.90[ab]
Secondary and non-demographics									
Without post-stratification	3.64	3.96	5.25[ab]	5.79[ab]	6.05[ab]	4.79[a]	4.81[a]	5.38[ab]	8.93[ab]
With post-stratification	2.90	3.40	4.53[ab]	5.22[ab]	4.53[a]	5.51[ab]	5.17[ab]	5.08[ab]	6.61[ab]
Rank: Average absolute error									
Primary demographics									
Without post-stratification	2	1	3	5	8	7	6	4	9
All benchmarks									
Without post-stratification	2	1	3	7	8	6	4	5	9
Secondary and non-demographics									
Without post-stratification	1	2	5	7	8	3	4	6	9
With post-stratification	1	2	3	7	4	8	6	5	9
Largest percentage point absolute error									
All benchmarks									
Without post-stratification	11.71	9.59	17.99	13.23	15.55	15.25	13.70	15.97	35.54

Continued

Table 2. *Continued*

	Probability Sample Surveys		Non-Probability Sample Internet Surveys						
Evaluative Criteria	Telephone	Internet	1	2	3	4	5	6	7
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	9.00	8.42	14.56	11.90	12.09	15.14	12.95	9.98	17.83
% Significant differences from benchmark									
All benchmarks									
Without post-stratification	47%	63%	58%	68%	79%	58%	63%	58%	89%
Secondary and non-demographics									
Without post-stratification	46%	69%	69%	77%	77%	54%	62%	62%	85%
With post-stratification	31%	46%	69%	69%	69%	77%	62%	77%	77%

[a] Significantly different from the telephone sample survey at $p < .05$.

[b] Significantly different from the probability sample Internet survey at $p < .05$.

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

		Probability Sample Surveys		Non-Probability Sample Internet Surveys						
Benchmark comparison	Value	Telephone	Internet	1	2	3	4	5	6	7
Primary demographic										
Female	51.68%									
Without post-stratification										
Estimate		55.40%*	50.57%	53.80%	49.82%	48.73%*	50.73%	52.26%	49.61%	55.45%*
Percentage point error		3.72	−1.11	2.12	−1.86	−2.95	−0.95	0.58	−2.07	3.77
With post-stratification										
Estimate		51.63	51.50	51.27	51.66	51.31	52.46	52.32	51.71	50.98
Percentage point error		−0.05	−0.18	−0.41	−0.02	−0.37	0.78	0.64	0.03	−0.70
Age 38–47	20.83									
Without Post-stratification										
Estimate		20.88	22.10	19.78	19.54	21.26	20.55	19.45	20.80	15.54*
Percentage point error		0.05	1.27	−1.05	−1.29	0.43	−0.28	−1.38	−0.03	−5.29
With post-stratification										
Estimate		21.55	20.73	22.34	19.89	20.36	22.89	20.88	19.49	20.84
Percentage point error		0.72	−0.10	1.51	−0.94	−0.47	2.06	0.05	−1.34	0.01
White only	82.02									
Without post-stratification										
Estimate		79.15	79.12*	84.10*	86.22*	89.62*	88.73*	87.11*	82.19	46.48*
Percentage point error		−2.87	−2.90	2.08	4.20	7.60	6.71	5.09	0.17	−35.54
With post-stratification										
Estimate		82.02	82.02	82.02	82.01	82.03	82.02	82.03	82.02	81.77
Percentage point error		0.00	0.00	0.00	−0.01*	0.01	0.00	0.01	0.00	−0.25

Continued

Table 3. *Continued*

		Probability Sample Surveys		Non-Probability Sample Internet Surveys						
Benchmark comparison	Value	Telephone	Internet	1	2	3	4	5	6	7
Non-Hispanic	87.62									
Without post-stratification										
Estimate		94.61*	90.86*	88.64	95.09*	96.65*	96.64*	94.78*	93.44*	90.19*
Percentage point error		6.99	3.24	1.02	7.47	9.03	9.02	7.16	5.82	2.57
With post-stratification										
Estimate		87.62	87.62	87.62	87.63	87.62	87.63	87.63	87.62	87.70
Percentage point error		0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.08
High school degree	31.75									
Without post-stratification										
Estimate		27.36*	31.41	13.76*	18.52*	16.20*	16.50*	19.03*	20.45*	16.60*
Percentage point error		−4.39	−0.34	−17.99	−13.23	−15.55	−15.25	−12.72	−11.30	−15.15
With post-stratification										
Estimate		31.75	31.71	34.65*	31.79	34.75*	33.98	32.81	31.75	31.80
Percentage point error		0.00	−0.04	2.90	0.04	3.00	2.23	1.06	0.00	0.05
South	35.92									
Without post-stratification										
Estimate		34.22	38.82*	36.13	38.01	39.03*	29.80*	40.99*	44.85*	26.25*
Percentage point error		−1.70	2.90	0.21	2.09	3.11	−6.12	5.07	8.93	−9.67
With post-stratification										
Estimate		35.92	35.92	35.92	35.92	35.91	35.90	35.95	35.92	35.94
Percentage point error		0.00	0.00	0.00	0.00	−0.01	−0.02	0.03	0.00	0.02

Continued

Table 3. *Continued*

		Probability Sample Surveys		Non-Probability Sample Internet Surveys						
Benchmark comparison	Value	Telephone	Internet	1	2	3	4	5	6	7
Secondary demographic										
Married	56.50									
Without post-stratification										
Estimate		61.87*	59.82*	58.77*	59.93*	61.49*	56.82	58.15	53.71	45.54*
Percentage point error		5.37	3.32	2.27	3.43	4.99	0.32	1.65	−2.79	−10.96
With post-stratification										
Estimate		58.55	57.11	55.49	57.64	56.33	51.33*	48.81*	52.64*	51.99*
Percentage point error		2.05	0.61	−1.01	1.14	−0.17	−5.17	−7.69	−3.86	−4.51
2 people in household	33.84									
Without post-stratification										
Estimate		34.19	37.46*	41.50*	36.52	39.98*	40.55*	38.25*	35.25	23.96*
Percentage point error		0.35	3.62	7.66	2.68	6.14	6.71	4.41	1.41	−9.88
With post-stratification										
Estimate		32.24	34.56	38.38*	34.72*	37.98*	36.44	31.85	32.90	27.72*
Percentage point error		−1.60	0.72	4.54	0.88	4.14	2.60	−1.99	−0.94	−6.12
Worked last week	60.80									
Without post-stratification										
Estimate		60.58	61.69	63.12*	53.59*	67.05*	61.18	55.76*	60.00	63.29
Percentage point error		−0.22	0.89	2.32	−7.21	6.25	0.38	−5.04	−0.80	2.49
With post-stratification										
Estimate		58.50	62.46	62.17	48.90*	60.24	60.12	51.39*	57.07*	58.44
Percentage point error		−2.30	1.66	1.37	−11.90	−0.56	−0.68	−9.41	−3.73	−2.36
3 bedrooms	43.38									
Without post-stratification										
Estimate		44.43	45.88	43.56	46.14	45.05	45.18	41.71	41.25	36.87*
Percentage point error		1.05	2.50	0.18	2.76	1.67	1.80	−1.67	−2.13	−6.51

Continued

Table 3. *Continued*

		Probability Sample Surveys		Non-Probability Sample Internet Surveys						
Benchmark comparison	Value	Telephone	Internet	1	2	3	4	5	6	7
With post-stratification										
Estimate		44.82	44.91	43.52	48.25*	47.95*	45.19	39.69*	42.11	41.00
Percentage point error		1.44	1.53	0.14	4.87	4.57	1.81	−3.69	−1.27	−2.38
2 vehicles	41.46									
Without post-stratification										
Estimate		41.09	45.53*	43.73	44.41*	46.93*	41.82	42.03	40.18	41.50
Percentage point error		−0.37	4.07	2.27	2.95	5.47	0.36	0.57	−1.28	0.04
With post-stratification										
Estimate		40.94	45.41*	43.62	42.12	44.80*	37.88*	39.17	38.35*	45.66*
Percentage point error		−0.52	3.95	2.16	0.66	3.34	−3.58	−2.29	−3.11	4.20
Owns home	72.50									
Without post-stratification										
Estimate		78.75*	71.72	72.68	68.66*	71.71	71.18	69.32*	64.83*	52.46*
Percentage point error		6.25	−0.78	0.18	−3.84	−0.79	−1.32	−3.18	−7.67	−20.04
With post-stratification										
Estimate		76.20*	69.51*	67.23*	66.62*	72.44	67.24*	66.69*	62.84*	61.26*
Percentage point error		3.70	−2.99	−5.27	−5.88	−0.06	−5.26	−5.81	−9.66	−11.24
HH income $50K –59.9K	15.11									
Without post-stratification										
Estimate		14.05*	23.26	21.57*	23.00*	18.44*	19.96*	19.63*	19.52*	19.28*
Percentage point error		−1.06	8.15	6.46	7.89	3.33	4.85	4.52	4.41	4.17
With post-stratification										
Estimate		14.57	22.18*	21.47*	22.50*	17.95*	20.18*	22.01*	19.38*	16.07
Percentage point error		−0.54	7.07	6.36	7.39	2.84	5.07	6.90	4.27	0.96

Continued

Table 3. *Continued*

		Probability Sample Surveys		Non-Probability Sample Internet Surveys						
Benchmark comparison	Value	Telephone	Internet	1	2	3	4	5	6	7
Non-demographic										
Non-smoker	78.25									
Without post-stratification										
Estimate		76.63	74.91*	76.26	68.85*	70.40*	73.73*	68.76*	70.55*	65.82*
Percentage point error		−1.62	−3.34	−1.99	−9.40	−7.85	−4.52	−9.49*	−7.70	−12.43
With post-stratification										
Estimate		75.69	74.00*	72.44*	68.25*	69.85*	69.39*	70.18*	70.18*	60.42*
Percentage point error		−2.56	−4.25	−5.81	−10.00	−8.40	−8.86	−8.07	−8.07	−17.83
Had 12 drinks in lifetime	77.45									
Without post-stratification										
Estimate		84.54*	87.04*	92.09*	89.57*	90.48*	92.25*	91.15*	88.89*	81.55*
Percentage point error		7.09	9.59	14.64	12.12	13.03	14.80	13.70	11.44	4.10
With post-stratification										
Estimate		84.60*	85.87*	92.01*	87.87*	89.54*	92.59*	90.40*	87.43*	90.63*
Percentage point error		7.15	8.42	14.56	10.42	12.09	15.14	12.95	9.98	13.18
Has 1 drink on average	37.67									
Without post-stratification										
Estimate		43.46*	42.98*	40.22*	37.74	38.73	38.77	40.18	33.75*	22.07*
Percentage point error		5.79	5.31	2.55	0.07	1.06	1.10	2.51	−3.92	−15.60
With post-stratification										
Estimate		39.47	40.33	34.60*	39.54	38.19	32.49*	37.24	32.14*	32.69*
Percentage point error		1.80	2.66	−3.07	1.87	0.52	−5.18	−0.43	−5.53	−4.98
Health ‘‘very good’’	31.43									
Without post-stratification										
Estimate		33.64	37.91*	38.73*	40.45*	40.39*	38.91*	34.12	36.42*	42.93*

Continued

Table 3. *Continued*

		Probability Sample Surveys		Non-Probability Sample Internet Surveys						
Benchmark comparison	Value	Telephone	Internet	1	2	3	4	5	6	7
Percentage point error		1.71	5.98	6.80	8.52	8.46	6.98	2.19	4.49	11.00
With post-stratification										
Estimate		33.30	38.93*	37.06*	39.05*	40.27*	39.58*	33.53	33.86	36.90*
Percentage point error		1.37	7.00	5.13	7.12	8.34	7.65	1.60	1.93	4.97
Does not have a passport	78.50									
Without post-stratification										
Estimate		66.79*	75.19*	63.82*	70.21*	65.99*	65.36*	68.87*	62.53*	63.30*
Percentage point error		−11.71	−3.31	−14.68	−8.29	−12.51	−13.14	−9.63	−15.97	−15.20
With post-stratification										
Estimate		69.50*	76.28	72.14*	75.19*	68.68*	71.82*	72.59*	69.09*	69.33*
Percentage point error		−9.00	−2.22	−6.36	−3.31	−9.82	−6.68	−5.91	−9.41	−9.17
Has a driver’s license	89.00									
Without post-stratification										
Estimate		93.77*	89.63	95.27*	95.10*	96.08*	95.00*	93.02*	94.97*	92.66*
Percentage point error		4.77	0.63	6.27	6.10	7.08	6.00	4.02	5.97	3.66
With post-stratification										
Estimate		92.66*	87.91	92.10*	91.40*	93.06*	92.97*	88.54	93.22*	93.00*
Percentage point error		3.66	−1.09	3.10	2.40	4.06	3.97	−0.46	4.22	4.00

Note.—All errors are deviations from the benchmark.
*$p < .05$.

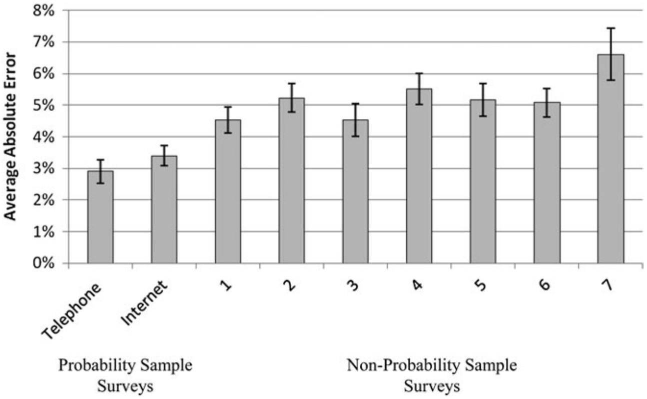

Figure 1. Average Percentage Point Absolute Errors for Commissioned Probability and Non-Probability Sample Surveys across Thirteen Secondary Demographics and Non-Demographics, with Post-Stratification. NOTE.—Error bars represent + or − 1 standard error.

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 35.54 percentage points. With post-stratification, the same was true—the probability samples had smaller "largest absolute errors", of 9.00 and 8.42 percentage points, in contrast to the non-probability sample Internet surveys' "largest absolute errors," which ranged from 9.98 to 17.83 percentage points.

Number of significant differences from benchmarks. The same conclusion was supported by 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 survey's estimates were significantly different from the fewest benchmarks (47 percent). The probability sample Internet survey's estimates were significantly different from somewhat more benchmarks (63 percent), and the non-probability sample Internet surveys were significantly different from about the same percent or more of the benchmarks (range: 58 to 89 percent). With post-stratification, however, the probability samples were more obviously superior: their estimates were significantly different from 31 and 46 percent of the benchmarks, respectively, whereas the non-probability sample Internet surveys were significantly different from between 62 and 77 percent of the benchmarks.

Superiority of Some Non-Probability Samples?

The average accuracies examined thus far may seem to suggest that some non-probability samples were more accurate than others, but these differences were almost never statistically significant.[14] 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 19 benchmarks ranged from −.65 to .70 and averaged .27. Similarly, the correlation between average absolute error for each survey and absolute error for each of the 19 benchmarks ranged from −.94 to .92 and averaged .37. These two average correlations were similar when post-stratification was done (.23 and .27, respectively). Thus, these results challenge the conclusion that some of the non-probability sample Internet surveys were consistently more accurate than the rest.

COMPARISONS WITH 2004/2005 ARCHIVAL SURVEYS

Because the RDD telephone survey and probability sample Internet survey commissioned for this article were overseen closely by university researchers, and the telephone survey involved a very long field period and extensive efforts to maximize the response rate, one might imagine that these surveys overstate the accuracy of most surveys done with these methods at that time. In fact, however, the surveys commissioned for this article were almost always less accurate than the archival surveys we examined (commissioned telephone survey: average absolute error 3.74 percentage points vs. archival telephone surveys: 3.40, 3.18, 3.55, 3.40, 3.46, 3.79 percentage points; commissioned probability sample Internet survey: 2.67 percentage points vs. archival probability sample Internet surveys: 1.71; 2.10; 1.43; 1.94; 1.62; 1.45 percentage points).

Consistency of Absolute Errors in the 2004/2005 Surveys.

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 (i.e., the commissioned survey plus the six archival surveys) was 3.51 percentage points (for the primary demographics and some secondary demographics), with a standard deviation of 0.23 percentage points. The corresponding average absolute error and standard deviation were 1.84 and 0.44 percentage points for the seven probability sample Internet surveys, respectively. In contrast, for the seven non-probability sample Internet surveys, these figures were 5.60 and 2.20 percentage points, respectively. Thus, the standard deviation for the non-probability sample surveys' average error was nearly ten

[14] 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.

times larger than the telephone surveys' standard deviation and five times larger than the probability sample Internet surveys' standard deviation.

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 sample surveys, 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 percentage points for the seven probability sample telephone surveys, 1.31 percentage points for the seven probability sample Internet surveys, and 5.19 percentage points for the seven non-probability sample surveys. Thus, it is easier to predict a probability sample survey's accuracy on one benchmark knowing its accuracy on another benchmark than to predict a non-probability sample survey's accuracy on one benchmark knowing its accuracy on another benchmark.

The consistency of the accuracy of the probability sample surveys is illustrated in figure 2. In terms of nine demographic variables, the range of absolute errors was much narrower for the seven RDD telephone surveys (shown on the left of figure 2) and the seven probability sample Internet surveys (shown in the middle) than for the seven non-probability sample Internet surveys (shown on the right). Thus, it is difficult to anticipate whether a non-probability sample Internet survey will be somewhat different from a population benchmark or substantially different from it, whereas the probability sample surveys were consistently only minimally different.

RELATION OF COMPLETION RATES TO ACCURACY

Remarkably, higher response rates were associated with lower accuracy of the surveys. Among the seven RDD telephone surveys, response rates were

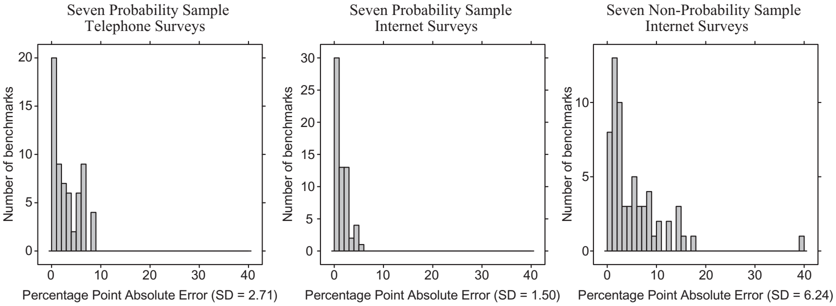

Figure 2. Histograms Showing the Variability in Absolute Errors for Comparisons to Nine Benchmarks from Seven Commissioned and Archival Probability Sample Telephone Surveys, Seven Commissioned and Archival Probability Sample Internet Surveys, and Seven Commissioned Non-Probability Sample Internet Surveys.

positively correlated with the size of each survey's average absolute error without post-stratification ($r = .47$). Completion rates were also positively correlated with the size of average absolute error among the seven probability sample Internet surveys ($r = .47$) and among the non-probability sample Internet surveys ($r = .61$). Thus, higher completion rates and response rates were coincident with less accuracy, not more.

COMPARISON WITH THE 2008 ARF SURVEYS

In the ARF's study, the post-stratification weighted average absolute error across the benchmark questions not used for weighting was nearly identical to that found in the present study's commissioned 2004/2005 non-probability sample Internet surveys (5.2 vs. 5.2 percentage points). For the same current cigarette-smoking benchmark, the standard deviation of absolute errors across multiple non-probability sample surveys was nearly identical in the ARF study and in the present study's commissioned 2004/2005 surveys as well (3.74 vs. 3.45 percentage points), as was the largest absolute error among non-probability sample surveys (12 vs. 12). Thus, when considering only the limited set of publicly available results from the ARF's 2008 FoQ study, there was remarkable correspondence to the present study's results.

COMPARING ERRORS IN THE 2004/2005 AND 2009 SURVEYS

In no instance was a firm's 2009 average error significantly different from its 2004/2005 average error. T-tests comparing the average absolute error for each firm's surveys in 2004/2005 and 2009 using bootstrapped standard errors failed to reach significance (telephone: $\Delta = -0.23$ percentage points, $t = 0.38$, $p > .10$; probability sample Internet $\Delta = -0.09$ percentage points, $t = 0.18$, $p > .10$; non-probability sample Internet $\Delta = -0.81$ percentage points, $t = 1.39$, $p > .10$). Moreover, in 2009, as in 2004/2005, the probability sample Internet survey was significantly more accurate than the non-probability sample Internet sample survey (difference in average absolute error = 2.84 percentage points, $t[1100] = 4.76$, $p < .05$).

Discussion

This investigation supports the following conclusions:

(1) The probability sample telephone and Internet surveys commissioned for this study were more 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 = 2.90 and 3.40 percentage points, respectively, for the telephone and probability sample Internet surveys).

(2) The non-probability sample Internet surveys were always less accurate, on average, than the probability sample surveys (average absolute errors for secondary demographics and non-demographics = 5.23 percentage points) and were less consistent in their accuracy. Thus, the accuracy of any one measure in a non-probability sample survey was 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, on average, challenging the hypothesis that optimizing methods of conducting non-probability sample Internet surveys can maximize their accuracy.
(5) Completion rates and response rates of the surveys were negatively correlated with their accuracy, challenging the notion that higher completion rates and response rates are indications of higher accuracy.
(6) The accuracy of probability and non-probability sample surveys in 2004/2005 and 2009 is about the same.

Conclusion (1) is useful because probability sample surveys routinely come under attack, being accused of inaccuracy due to low response rates and a shrinking landline telephone sampling frame (e.g., Crampton 2007; Kellner 2004, 2007; Zogby 2007, 2009; see also Ferrell and Peterson 2010). It is rarely possible to evaluate the credibility of such assertions, because benchmarks to assess accuracy are rarely available. Therefore, this investigation's administration of measures suitable for comparison to benchmarks made this sort of evaluation 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 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.

RESONANCE WITH PREVIOUS RESEARCH

The evidence reported here 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. Those investigators found numerous statistically significant and sizable differences between the surveys in terms of demographics and non-demographics.

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 substantial differences between probability and non-probability sample surveys. However, because these studies did not compare the estimates to trusted benchmarks, it is impossible to tell which data collection method yielded more accurate results. The present study suggests that in general, RDD telephone surveys are likely to be more accurate than non-probability sample Internet surveys. This conclusion is supported by similar studies that compared estimates from an RDD telephone survey and a non-probability sample Internet survey to benchmarks and found greater error in the latter (e.g., Bethell et al. 2004; Chang and Krosnick 2009; Crete and Stephenson 2008; Niemi et al. 2008; van Ryzin 2008).[15]

The present study's finding of substantial variability in results across non-probability sample surveys (in 2004/2005 and in 2008) resonates with other studies documenting the same result (e.g., Baim et al. 2009; Elmore-Yalch et al. 2008; Vonk, Ossenbruggen, and Willems 2006). All of this reinforces confidence in the conclusions supported by the present research.

EFFECTS OF SURVEY WEIGHTS ON ACCURACY

Advocates of non-probability sample surveys sometimes assert that inadequacies in panel recruitment and survey participation can be corrected by suitable adjustments implemented when inviting panelists to complete the survey or after data collection. And some firms that sell such data sometimes say that they have developed effective, proprietary methods to do so. The evidence reported here challenges those assertions in various ways: (1) the non-probability sample surveys differed in how they used quotas, adjusted the probability of invitation from the panel, and provided incentives, yet none emerged as superior to the others; (2) the sizes of errors observed within and across such surveys were not consistent; (3) the weights that included a propensity score adjustment did not reduce the errors in that non-probability sample Internet survey; and (4) post-stratification of non-probability samples did not consistently improve accuracy, whereas post-stratification did increase the accuracy of probability sample surveys. This suggests that weights may not always be effective for removing the biases in non-probability sample surveys, although they are effective at reducing error in probability sample surveys.

[15] Other research has compared non-probability sample Internet surveys to probability sample surveys conducted face-to-face and has found similar results (e.g., Newman et al. 2002; Smith 2003; Smith and Dennis 2005). For instance, Loosveldt and Sonck (2008) found numerous sizable differences between a non-probability sample Internet survey's measurements and those of a probability sample face-to-face survey in Belgium, as did Faas and Schoen (2006) with data from Germany. Again, these investigators did not provide evidence on which survey was the more accurate. Malhotra and Krosnick (2007) compared probability sample face-to-face surveys with non-probability sample Internet surveys and showed that the face-to-face surveys' results were more accurate.

The logic of such weighting hinges on the assumption that the members of underrepresented 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 underrepresented subgroups do not resemble non-interviewed members of such groups as closely as occurs with probability sampling. For example, if young, African-American, highly educated males were underrepresented 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. This may be the reason why weighting up the participating members of this group increased error rather than decreasing it.

Resonating with this logic, many researchers (Couper et al. 2007; Dever et al. 2008; Duffey et al. 2005, Lensvelt-Mulders, Lugtig, and Hubregtse 2009; Loosveldt and Sonck 2008; Schonlau et al. 2009) have shown that although weighting and propensity score adjustments can sometimes significantly reduce error in non-probability sample surveys, large discrepancies from the population of interest often remain even when the data are weighted. We look forward to seeing the results of future research seeking to refine and improve these methods.

THE SPECIAL CASE OF PRE-ELECTION POLLS

Pre-election polls are perhaps the most visible context in which probability sample and non-probability sample surveys compete and can be evaluated. A number of publications document excellent accuracy of non-probability sample Internet surveys (with some notable exceptions), some instances of better accuracy than probability sample surveys, and some instances of lower accuracy (Harris Interactive 2004, 2008; Stirton and Robertson 2005; Taylor et al. 2001; Twyman 2008; Vavreck and Rivers 2008). However, to produce these numbers, analysts must make numerous decisions about how to identify likely voters, how to handle respondents who decline to answer vote choice questions, how to weight data, how to order candidate names on questionnaires, and more. And these decisions can be shaped partly by the results of numerous surveys measuring the same preferences that were publicized previously during a campaign. So differences or similarities between polls in terms of accuracy may reflect differences in analysts' procedures rather than differences in the inherent accuracy of the data collection methods. Thus, it is difficult to know how to relate the present findings to the evidence on candidate preferences during campaigns.

COMPLETION RATES AND RESPONSE RATES

The evidence reported here that higher completion rates and response rates were not associated with more accuracy is consistent with the growing body of work supporting the same conclusion (e.g., Curtin, Presser and Singer 2000; Holbrook, Krosnick, and Pfent 2007; Keeter et al. 2000; Keeter et al. 2006; Merkle and Edelman 2002). The evidence of negative relations of completion rates and response rates with accuracy challenges the claims that probability sampling is undermined by low response rates and that efforts devoted to maximizing completion rates and response rates necessarily maximize the accuracy of surveys.

LIMITATIONS

The present study has significant 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. Just as the evidence reported here shows that random sampling of respondents yields more generalizable results, random sampling of measures would permit an increase in confidence when generalizing research 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 it may not be possible for the community of scholars to agree on what constitutes the population of measures, in which case more investigations such as the present one can be conducted with convenience samples of measures to provide a basis for further confidence in general conclusions.

Second, the present study focused on commissioned 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, and some highly visible firms were not included. Although the results from the seven data collection firms in 2004 closely mirrored those obtained from 17 data collection firms in the 2008 ARF study, 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 data collection firms for participation.

HOW ACCURATE WERE THE BENCHMARKS?

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. Because those surveys had extremely high response rates (so non-response bias was extremely small) and involved very large samples of respondents, there is reason to have confidence in them.

Nevertheless, the use of human interviewers in the benchmark surveys and in the telephone surveys we examined might have created an illusion of similarity between their results. Human interviewers sometimes induce a bias toward giving socially desirable answers, especially in telephone interviews, and this bias is less present in answers to self-administered questionnaires (Chang and Krosnick 2009; Harmon et al. 2009; Holbrook, Green, and Krosnick 2003; Holbrook and Krosnick 2010; Turner et al. 2005; Turner et al. 2009; Villarroel et al. 2006). Therefore, for measures tinged with social desirability implications, the benchmark comparisons might overstate the accuracy of the telephone survey data.

Such bias seems unlikely to have contaminated the CPS or ACS's measures of demographics such as sex or number of bedrooms in the household or, of course, the government's statistics on passports or driver's licenses. But one might imagine that reports of cigarette smoking, alcohol consumption, and health quality could be distorted by social desirability pressures.

In fact, however, such pressures do not appear to bias adults' reports of smoking and drinking. Numerous studies have used the "bogus pipeline technique" to test for social desirability bias in such reports, and meta-analyses of these studies concluded that the bogus pipeline technique did not increase adults' reports of these behaviors (Aguinis, Pierce, and Quigley 1993, 1995). An additional study analyzed eight years of data from the NHANES, a large, federal face-to-face survey, and found remarkable correspondence between self-reports of nicotine consumption and blood test results (Yeager and Krosnick 2010).[16] In that study, less than 1 percent of adults said they did not recently consume nicotine yet had blood tests that suggested otherwise.

Furthermore, if social desirability pressures distorted self-reports of smoking and drinking, and if these pressures operated more in interviewer-administered surveys than in self-administered surveys, we should have seen reports of less smoking and drinking in the RDD telephone survey commissioned for this article than in the probability sample Internet survey we commissioned. Yet no significant differences between the two were found in reports of smoking or drinking.

Some investigators have speculated that reports of health quality may be subject to social desirability response bias (e.g., Adams et al. 2005; McHorney, Kosinski, and Ware 1994; Siemiatycki 1979), and several studies have found reports of higher health quality in interviewer-administered surveys than in self-administered surveys (e.g., Baker et al. 2004; Bethell et al. 2004; Hochstim 1967; McHorney et al. 1994; Schonlau et al. 2004; Siemiatycki 1979). Likewise, the present study found that telephone survey respondents reported

[16] Klein et al. (2007) reported results that they said indicated substantial bias in smoking self-reports in interviewer-administered surveys, but see Yeager and Krosnick (2010) for a reanalysis of that data, or see http://blogs.abcnews.com/thenumbers/2009/12/survey-accuracy-revisiting-the-benchmarks-.html.

significantly higher health quality than did the probability sample Internet survey respondents.

Nonetheless, when the data from the present study were reanalyzed dropping the health quality benchmark, the same conclusions were reached about the relative accuracy of the nine commissioned surveys. So, the conclusions of this research do not seem likely to have been distorted by social desirability response bias in the benchmark surveys.[17]

Conclusion

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, yielded quite accurate results. In contrast, non-probability samples were not as accurate and were sometimes strikingly inaccurate, regardless of their completion rates. 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 five 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 intended mostly to assess whether two variables were related to each other along the lines that theory anticipated. The continued use of non-probability samples seems quite reasonable if one's goal is not to document the strength of an association in a population but rather to reject the null hypothesis that two variables are completely unrelated to each other throughout the population. Yet if a researcher's goal is to document the distribution of a variable in a population accurately, non-probability sample surveys appear to be considerably less suited to that goal than probability sample surveys.

[17] Another potential source of bias in the telephone survey is acquiescence response bias, which is the tendency to say "yes" in response to yes/no questions, regardless of their content. Some previous research suggests that this bias is more prevalent in telephone surveys than in computer-based surveys (e.g., Chang and Krosnick 2010). Consistent with this finding, the present study found that the telephone survey overestimated possession of passports and driver's licenses, more so than did the probability sample Internet survey (ps $<$ .05). So, some of the overestimation in the telephone survey data may have been due to acquiescence and not due to sample composition.

Appendix 1

MEASURES IN THE 2004/2005 COMMISSIONED SURVEYS

Identical questions measuring primary demographics, secondary demographics, and non-demographics were asked in the same order in a survey that lasted about 30 minutes on average by telephone.

Primary demographics.

Sex: [Only asked of Internet respondents. For the telephone survey, interviewers coded respondent sex.] "Are you male or female?" (Internet response options: Male, Female; Categories used for analysis: Male, Female)

Age: "In what year were you born?" To calculate age, open-ended responses were subtracted from 2004 (the year in which the survey was conducted) [programming restricted answers to range from 0 to 1986]; Categories used for analysis: 18–27, 28–37, 38–47, 48–57, 58–67, 68+.

Ethnicity: "Are you Spanish, Hispanic, or Latino?" (Telephone and Internet response options and categories used for analysis: Yes, No)

Race: Telephone: "Please tell me which of the following races you consider yourself to be: White; Black or African American; American Indian or Alaska Native; Asian; Native Hawaiian, or Other Pacific Islander." Internet: "Which of the following races do you consider yourself to be?" (Categories used for analysis: White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Other)

Education: "What is the highest level of school you have completed or the highest degree you have received?" (Telephone: Interviewer coded responses into the categories listed below;

Internet response options and categories used for analysis: Less Than 1st Grade, 1st Grade, 2nd Grade, 3rd Grade, 4th Grade, 5th Grade, 6th Grade, 7th Grade, 8th Grade, 9th Grade, 10th Grade, 11th Grade, 12th Grade with No Diploma, High School Diploma or an Equivalent, Such as a GED, Some College But No Degree, Associate Degree from an Occupational/Vocational Program, Associate Degree from an Academic Program; Bachelor's Degree, such as B.A., B.S., or A.B.; Master's Degree, such as M.A., M.S., Masters in Engineering, Masters in Education, or Masters in Social Work; Professional School Degree, such as M.D., D.D.S., or D.V.M.; Doctorate Degree, such as Ph.D., Ed.D.)

Region: "In what state do you live?" (*Telephone:* Interviewers recorded open-ended responses. *Internet response options:* All 50 states and the District of Columbia; *Categories used for analysis:* Northeast, Midwest, South, West, using Census region classifications)

Secondary demographics.

Marital Status: "Are you now married, widowed, divorced, separated, or never married?" (Categories used for analysis: Married, Widowed, Divorced, Separated, Never Married)

People in Household: "Including yourself, how many adults, 18 years old or older, live in your household? Do not include college students who are living away at college, persons stationed away from here in the armed forces, or persons away in institutions." "How many people age 17 or younger live in your household?" The two open-ended answers were summed (Categories used for analysis: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15+).

Work Status: "Last week, did you do ANY work for either pay or profit?" (Categories used for analysis: Yes, No)

Number of Bedrooms: "How many bedrooms are in your house, apartment, or mobile home? That is, how many bedrooms would you list if your house, apartment, or mobile home were on the market for sale or rent?" (Telephone: Interviewers recorded open-ended responses; Internet response options: No Bedrooms, 1 Bedroom, 2 Bedrooms, 3 Bedrooms, 4 Bedrooms, 5 or More Bedrooms; Categories used for analysis: 0, 1, 2, 3, 4, 5+)

Number of Vehicles: "How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of your household?" *(Categories used for analysis*: None, 1, 2, 3, 4, 5, 6+)

Owning a Home: "Are your living quarters. . .owned or being bought by a household member, rented for cash, or occupied without payment of cash rent?" (Response options and categories used for analysis: Owned, Rented, Occupied without Payment of Cash Rent)

Household Income: "Thinking about your total household income from all sources, including your job, how much was your total household income in 2003 before taxes?" A respondent who refused to answer was asked "Was it $35,000 or more?" and, if so, he or she was asked "Was it $50,000 or more?," "Was it $60,000 or more?," "Was it $75,000 or more?," "Was it $100,000 or more?," "Was it $150,000 or more?," "Was it $200,000 or more?," or "Was it $250,000 or more?," in that order, until the respondent said "no" to one of those questions. If the respondent said his or her income was less than $35,000, he or she was asked “Was it $10,000 or more?,” “Was it $15,000 or more?,” or “Was it $25,000 or more?,” in that order, until the respondent said “no” to one of those questions (Categories used for analysis: Less Than $10,000, $10,000–14,999, $15,000–24,999, $25,000–34,999, $35,000–49,999, $50,000–59,999, $60,000–74,999, $75,000–99,999, $100,000–149,999, $150,000–199,999, $200,000–249,999, $250,000+).

Non-demographics.
Smoking: “Do you smoke cigarettes every day, some days, or not at all?” (Categories used for analysis: Every Day, Some Days, Not at All)

Drinking in Lifetime: [Asked only of respondents who were 21 years old or older]“In your ENTIRE LIFE, have you had at least 12 drinks of any type of alcoholic beverage?” (Categories used for analysis: Yes, No)

Drinking This Year: [Only asked if a respondent answered “yes” to the previous question] “In the PAST YEAR, on those days that you drank alcoholic beverages, on average, how many drinks did you have?” Respondents who answered “no” to the previous question were coded as missing for this question. Coding respondents who answered “no” to the previous question as “0” instead of as missing did not change the pattern of results reported here. (Categories used for analysis: 0 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)

Quality of Health: “Would you say your health in general is excellent, very good, good, fair or poor?” (Categories used for analysis: Excellent, Very Good, Good, Fair, Poor)

Passport: “Do you personally own a valid U.S. passport?” (Categories used for analysis: Yes, No)

Driver’s License: “Do you personally have a current driver’s license?” (Categories used for analysis: Yes, No)

Appendix 2

SUMMARY OF PREVIOUS FINDINGS GENERATED FROM THE COMMISSIONED SURVEYS AND THEIR RELATION TO THE PRESENT STUDY

Collection of the data analyzed in this article was initiated by Norman Nie, following the model of an earlier project (Chang and Krosnick 2002, 2009). The general outline of the new data collection was designed by Jon Krosnick, Norman Nie, and Douglas Rivers; LinChiat Chiang assembled the questionnaire; coordination of the data collection firms was done by Chang, Krosnick, and Rivers.

Previous reports of findings generated using these data have been authored by Krosnick and Rivers (2005a, 2005b), Krosnick, Rivers, Simpser, Levendusky, and Chang (2005), and Graham (2005, 2007). Presentations at the annual conference of the 2005 American Association for Public Opinion Research by Krosnick and Rivers (2005a, 2005b) included results generated by a team of researchers at Stanford University. Some of those results were included in a paper being drafted at that time by Krosnick et al. (2005), which was never completed or released. Errors were discovered in the calculations, and the analyses were never redone by that team. The present article follows from that manuscript.

Analyses of the same data conducted by staff at Knowledge Networks were described in a publication released by Knowledge Networks (Graham 2005) and were presented by Graham (2007) at a conference sponsored by the Advertising Research Foundation.

In concept and form, tables in the present article resemble those presented by Krosnick and Rivers (2005a), Krosnick et al. (2005), and Graham (2005, 2007), make the same distinction between primary and secondary demographics, and compute error similarly.

To assess the accuracy of the various surveys, Krosnick and Rivers (2005a) and Krosnick et al. (2005) compared measurements of 21 variables with benchmarks thought to be accurate, and Graham (2005, 2007) compared 14 of those same variables with benchmarks and made comparisons with benchmarks for six additional variables.

The present article revisits the same datasets, but there is no overlap of the results reported here with the results reported by Krosnick and Rivers (2005a), Krosnick et al. (2005), or Graham (2005, 2007), because the present article reports results of analyses not conducted previously, and the prior papers reported results of analyses not reported here.

Among the present article's results that were not reported previously are: (1) new comparative measures of accuracy, including the rank order of the average absolute errors across data collection firms, the consistency of those ranks across variables, comparative analysis of the largest absolute error for each firm, the consistency of those largest absolute errors across measures, and the number of significant differences from benchmarks for each firm; (2) accuracy results for four new variables; (3) accuracy of six other telephone surveys conducted in 2004; (4) accuracy of six other probability sample Internet surveys conducted in 2004; (5) accuracy of three other surveys conducted in 2009; (6) accuracy of 17 other surveys conducted in 2008; (7) comparisons of two new sets of weights, one capped and one uncapped; (8) associations of response/completion rates with accuracy; (9) tests of statistical significance of differences in accuracy between surveys; and (10) statistical significance tests assessing whether post-stratification improved the accuracy of the surveys.

Seventeen of the variables that Krosnick and Rivers (2005a) and Krosnick et al. (2005) examined and seven of the variables examined by Graham (2005) are not examined here, because we believe no defensible benchmarks are available for these variables (for an explanation, see the online supplement accompanying the present article). Moreover, Krosnick and Rivers (2005a), Krosnick et al. (2005), and Graham (2005) used benchmark values from surveys conducted in 2003 or earlier or official government records calculated in 2003, whereas all surveys described in this article were conducted in 2004 or later.

None of the present article's results match those reported by Krosnick and Rivers (2005a), Krosnick et al. (2005), or Graham (2005, 2007), because different reference categories are used here, different benchmark values from 2004 are used here to correct errors in the earlier work, variables are grouped differently to yield summary statistics, no results are reported in this article using weights provided by the data collection companies, and computational errors in the earlier reported analyses have been corrected.

Supplementary Data

Supplementary data are freely available online at http://poq.oxfordjournals.org/.

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* 161:389–98.

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* 23:352–73.

———. 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* 16:515–27.

Baim, Julian, Michal Galin, Martin R. Frankel, Risa Becker, and Joe Agresti. 2009. "Sample Surveys Based on Internet Panels: 8 Years of Learning." Paper presented at the Worldwide Readership Symposium, Valencia, Spain.

Baker, Reg, Dan Zahs, and George Popa. 2004. "Health Surveys in the 21st Century: Telephone vs. Web." Paper presented at the Eighth Conference on Health Survey Research Methods, Peachtree City, GA.

Battaglia, Michael P., David Izrael, David C. Hoaglin, and Martin R. Frankel. 2009. "Practical Considerations in Raking Survey Data." *Survey Practice,* (June). http://surveypractice.org/2009/06/29/raking-survey-data/.

Bender, Bruce G., Susan J. Bartlett, Cynthia S. Rand, Charles Turner, Frederick S. Wamboldt, and Lening Zhang. 2007. "Impact of Reporting Mode on Accuracy of Child and Parent Report of Adherence with Asthma Controller Medication." *Pediatrics* 120:e471–77.

Berrens, Robert P., Alok K. Bohara, Hank Jenkins-Smith, Carol Silva, and David L. Weimer. 2003. "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples." *Political Analysis* 11:1–22.

Bethell, Christina, John Fiorillo, David Lansky, Michael Hendryx, and James Knickman. 2004. "Online Consumer Surveys as a Methodology for Assessing the Quality of the United States Health Care System." *Journal of Medical Internet Research* 6(1):e2.

Braunsberger, Karin, Hans Wybenga, and Roger Gates. 2007. "A Comparison of Reliability Between Telephone and Web-Based Surveys." *Journal of Business Research* 60:758–64.

Carbone, Enrica. 2005. "Demographics and Behavior." *Experimental Economics* 8:217–32.

Chang, LinChiat, and Jon A. Krosnick. 2002. "RDD Telephone vs. Internet Survey Methodology for Studying American Presidential Elections: Comparing Sample Representativeness and Response Quality." Paper presented at the American Political Science Association Annual Meeting, Boston.

———. 2009. "National Surveys via RDD Telephone Interviewing versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73:641–78.

———. 2010. "Comparing Oral Interviewing with Self-Administered Computerized Questionnaires: An Experiment." *Public Opinion Quarterly* 74:154–67.

Chatt, Cindy, Mike Dennis, Rick Li, and Paul Pulliam. 2003. "Data Collection Mode Effects Controlling for Sample Origins in a Panel Survey: Telephone versus Internet." Paper presented at the Annual Meeting of the Midwest Chapter of the American Association for Public Opinion Research, Chicago.

Cooley, Philip C., Heather G. Miller, James N. Gribble, and Charles F. Turner. 2000. "Automating Telephone Surveys: Using T-ACASI to Obtain Data on Sensitive Topics." *Computers in Human Behavior* 16:1–11.

Couper, Mick P., Arie Kapteyn, Matthias Schonlau, and Joachim Winter. 2007. "Noncoverage and Nonresponse in an Internet Survey." *Social Science Research* 36:131–48.

Crampton, Thomas. 2007. "About Online Surveys, Traditional Pollsters Are: (C) Somewhat Disappointed." *New York Times,* May 31. http://www.nytimes.com/2007/05/31/business/media/31adco.html?pagewanted=print.

Crete, Jean, and Laura Stephenson. 2008. "Internet and Telephone Survey Methodology: An Evaluation of Mode Effects." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64:413–28.

DeBell, Matthew, and Jon A. Krosnick. 2009. "Weighting Plan for the American National Election Studies." *American National Election Studies Technical Report,* Ann Arbor, MI.

Dever, Jill A., Ann Rafferty, and Richard Valliant. 2008. "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?" *Survey Research Methods* 2:47–62.

Diamond, Alexis, and Jasjeet Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." Working Paper, University of California at Berkeley. http://sekhon.berkeley.edu/papers/GenMatch.pdf.

Duffy, Bobby, Kate Smith, George Terhanian, and John Bremer. 2005. "Comparing Data from Online and Face-to-Face Surveys." *International Journal of Market Research* 47:615–39.

Elmore-Yalch, Rebecca, Jeffrey Busby, and Cynthia Britton. 2008. "Know Thy Customer? Know Thy Research! Comparison of Web-Based and Telephone Responses to a Public Service

Customer Satisfaction Survey." Paper presented at the Transportation Research Board 2008 Annual Meeting, Washington, D.C.

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* 22:177–90.

Ferrell, Dan, and James C. Peterson. 2010. "The Growth of Internet Research Methods and the Reluctant Sociologist." *Sociological Inquiry* 80:114–25.

Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22:153–64.

Ghanem, Khalil, Heidi E. Hutton, Jonathan M. Zenilman, Rebecca Zimba, and Emily J. Erbelding. 2005. "Audio Computer Assisted Self-Interview and Face-to-Face Interview Modes in Assessing Response Bias Among STD Clinic Patients." *Sexually Transmitted Infections* 81:421–25.

Graham, Patricia. 2005. "The Decision-Maker's Guide to Online Research." http://www.knowledgenetworks.com/dmg/dmg_09010.html.

———. 2007. "Using Known Benchmarks to Inform the Accuracy of Online Research." Presentation at the Advertising Research Foundation Online Research Quality Meeting, New York. http://s3.amazonaws.com/thearf-org-aux-assets/downloads/cnc/orqc/09-10-07_ORQC_Graham.pdf.

Harmon, Thomas, Charles F. Turner, Susan M. Rogers, Elizabeth Eggleston, Anthony M. Roman, Maria A. Villarroel, James R. Chromy, Laxminarayana Ganapathi, and Sheping Li. 2009. "Impact of T-ACASI on Survey Measurements of Subjective Phenomena." *Public Opinion Quarterly* 73:255–80.

Harris Interactive. 2004. "Final Pre-Election Harris Polls: Still Too Close to Call But Kerry Makes Modest Gains." http://www.prnewswire.com/news-releases/final-pre-election-harris-polls-still--too-close-to-call-but-internet-poll-results-suggest-a-kerry-victory-75137942.html.

———. 2008. "Election Results Further Validate Efficacy of Harris Interactive's Online Methodology." *Business Wire*, November 6. http://ir.harrisinteractive.com/releasedetail.cfm?ReleaseID=396524.

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.

Hochstim, Joseph R. 1967. "A Critical Comparison of Three Strategies of Collecting Data from Households." *Journal of the American Statistical Association* 62:976–89.

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, and Jon A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique." *Public Opinion Quarterly* 74:37–67.

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*, eds. James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith de Leeuw, Lilli Japec, Paul J. Lavrakas Michael W. Link, and Roberta L. Sangster. New York: Wiley-Interscience.

Kapteyn, Arie, James P. Smith, and Arthur van Soest. 2007. "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands." *American Economic Review* 97:461–73.

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* 70:759–79.

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* 64:125–48.

Kellner, Peter. 2004. "Can Online Polls Produce Accurate Findings?" *International Journal of Market Research* 46:3–23.

———. 2007. "Down with Random Samples." http://my.yougov.com/commentaries/peter-kellner/down-with-random-samples.aspx.

Kish, L. 1992. "Weighting for Unequal *P*." *Journal of Official Statistics* 8:183–200.

Klein, Jonathan D., Randall K. Thomas, and Erika J. Sutter. 2007. "Self-Reported Smoking in Online Surveys: Prevalence Estimate Validity and Item Format Effects." *Medical Care* 45: 691–95.

Krosnick, Jon A., and Douglas Rivers. 2005a. "Web Survey Methodologies: A Comparison of Survey Accuracy." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Miami Beach, FL.

———. 2005b. "Comparing Major Survey Firms in Terms of Survey Satisficing: Telephone and Internet Data Collection." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Miami Beach, FL.

Krosnick, Jon A., Douglas Rivers, Alberto Simpser, Matthew Levendusky, and LinChiat Chang. 2005. "Web Survey Methodologies: A Comparison and Evaluation." Unpublished manuscript draft, Stanford University, Stanford, CA.

Lee, Sunghee. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22:329–49.

Lee, Sunghee, and Valliant Richard. 2009. "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment." *Sociological Methods Research* 37:319–43.

Lensvelt-Mulders, Getty, Peter Lugtig, and Marianne Hubregtse. 2009. "Separating Selection Bias and Non-Coverage in Internet Panels Using Propensity Matching." *Survey Practice*. (August). http://surveypractice.org/.

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* 14:144–50.

Link, Michael W., and Ali H. Mokdad. 2004. "Are Web and Mail Feasible Options for the Behavioral Risk Factor Surveillance System?" Paper presented at the Eighth Conference on Health Survey Research Methods, Peachtree City, GA.

———. 2005. "Alternative Modes for Health Surveillance Surveys: An Experiment with Web, Mail, and Telephone." *Epidemiology* 16:701–4.

Loosveldt, Geert, and Nathalie Sonck. 2008. "An Evaluation of the Weighting Procedures for an Online Access Panel Survey." *Survey Research Methods* 2:93–105.

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." *Journal of Politics* 70:120–35.

McHorney, Colleen A., Mark Kosinski, and John Ware, Jr. 1994. "Comparisons of the Costs and Quality of Norms for the SF-36 Health Survey Collected by Mail versus Telephone Interview: Results from a National Survey." *Medical Care* 32:551–67.

Merkle, Daniel, and Murray Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." In *Survey Nonresponse*, eds. Robert Groves, Don Dillman, John Eltinge, and Roderick Little, pp. 243–58. New York: Wiley.

Metzger, David S., Beryl Koblin, Charles Turner, Helen Navaline, Francesca Valenti, Sarah Holte, Michael Gross, Amy Sheon, Heather Miller, and Philip Cooley. HIVNET Vaccine Preparedness Study Protocol Team. 2000. "Randomized Controlled Trial of Audio Computer-Assisted Self-Interviewing: Utility and Acceptability in Longitudinal Studies." *American Journal of Epidemiology* 152:99–106.

Moskalenko, Sophia, and Clark McCauley. 2009. "Measuring Political Mobilization: The Distinction Between Activism and Radicalism." *Terrorism and Political Violence* 21:239–60.

Newman, Jessica Clark Don C. Des Jarlais, Charles F. Turner, Jay Gribble, Phillip Cooley, and Denise Paone. 2002. "The Differential Effects of Face-to-Face and Computer Interview Modes." *American Journal of Public Health* 92:294–97.

Niemi, Richard, Kent Portney, and David King. 2008. "Sampling Young Adults: The Effects of Survey Mode and Sampling Method on Inferences About the Political Behavior of College Students." Paper presented at the Annual Meeting of the American Political Science Association, Boston.

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.

Rogers, Susan M., Gordon Willis, Alia Al-Tayyib, Maria A. Villarroel, Charles F. Turner, Lazminarayana Ganapathi, Jonathan Zenilman, and Rosemary Jadack. 2005. "Audio Computer-Assisted Interviewing to Measure HIV Risk Behaviors in a Clinic Population." *Sexually Transmitted Infections* 81:501–7.

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* 46:359–73.

Sanders, David, Harold D. Clarke, Marianne C. Stewart, and Paul Whiteley. 2007. "Does Mode Matter for Modeling Political Choice? Evidence from the 2005 British Election Study." *Political Analysis* 15:257–85.

Schillewaert, Niels, and Pascale Meulemeester. 2005. "Comparing Response Distributions of Offline and Online Data Collection Methods." *International Journal of Market Research* 47:163–78.

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* 21:218–22.

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.

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* 22:128–38.

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* 51:515–30.

Siemiatycki, Jack. 1979. "A Comparison of Mail, Telephone, and Home Interview Strategies for Household Health Surveys." *American Journal of Public Health* 69:238–45.

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* 15:167–79.

Smith, Tom W., and Michael J. Dennis. 2005. "Online vs. In-Person: Experiments with Mode, Format, and Question Wordings." *Public Opinion Pros* (December). http://www.publicopinionpros.norc.org/from_field/2005/dec/smith.asp.

Sparrow, Nick. 2006. "Developing Reliable Online Polls." *International Journal of Market Research* 48:659–80.

Spijkerman, Renske, Ronald Knibbe, Kim Knoops, Dike van de Mheen, and Regina van den Eijnden. 2009. "The Utility of Online Panel Surveys versus Computer-Assisted Interviews in Obtaining Substance-Use Prevalence Estimates in the Netherlands." *Addiction* 104:1641–45.

Stirton, John, and Euan Robertson. 2005. "Assessing the Viability of Online Opinion Polling During the 2004 Federal Election." *Australian Market and Social Research Society*.

Taylor, Humphrey, John Bremer, Cary Overmeyer, Jonathan W. Siegel, and George Terhanian. 2001. "The Record of Internet-Based Opinion Polls in Predicting the Results of 72 Races in the November 2000 U.S. Elections." *International Journal of Market Research* 43:127–36.

Taylor, Humphrey, David Krane, and Randall K. Thomas. 2005. "Best Foot Forward: Social Desirability in Telephone vs. Online Surveys." *Public Opinion Pros* (February). http://www.publicopinionpros.norc.org/from_field/2005/feb/taylor_2.asp.

Terhanian, George, Renee Smith, John Bremer, and Randall K. Thomas. 2001. "Exploiting Analytical Advances: Minimizing the Biases Associated with Internet-Based Surveys of Non-Random Samples." *ARF/ESOMAR: Worldwide Online Measurement* 248:247–72.

Turner, Charles F., Alia Al-Tayyib, Susan M. Rogers, Elizabeth Eggleston, Maria A. Villarroel, Anthony M. Roman, James R. Chromy, and Phillip C. Cooley. 2009. "Improving Epidemiological Surveys of Sexual Behavior Conducted by Telephone." *International Journal of Epidemiology* 38:1118–27.

Turner, Charles F., Maria A. Villarroel, Susan M. Rogers, Elizabeth Eggleston, Laxminarayana Ganapathi, Anthony M. Roman, and Alia Al-Tayyib. 2005. "Reducing Bias in Telephone Survey Estimates of the Prevalence of Drug Use: A Randomized Trial of Telephone Audio-CASI." *Addiction* 100:1432–44.

Twyman, Joe. 2008. "Getting It Right: YouGov and Online Survey Research in Britain." *Journal of Elections, Public Opinion, and Parties* 18:343–54.

van Ryzin, Gregg G. 2008. "Validity of an Online Panel Approach to Citizen Surveys." *Public Performance and Management Review* 32:236–62.

Vavreck, Lynn, and Douglas Rivers. 2008. "The 2006 Cooperative Congressional Election Study." *Journal of Elections, Public Opinion, and Parties* 18:355–66.

Villarroel, Maria A., Charles F. Turner, Elizabeth Eggleston, Alia Al-Tayyib, Susan M. Rogers, Anthony M. Roman, Philip C. Cooley, and Harper Gordek. 2006. "Same-Gender Sex in the USA: Impact of T-ACASI on Prevalence Estimates." *Public Opinion Quarterly* 70:166–96.

Vonk, Ted, Robert Ossenbruggen, and Pieter Willems. 2006. "The Effects of Panel Recruitment and Management on Research Results." In *Panel Research 2006*. Amsterdam: ESOMAR.

Yeager, David S., and Jon A. Krosnick. 2010. "The Validity of Self-Reported Nicotine Product Use in the 2001–2008 National Health and Nutrition Examination Survey." *Medical Care* 48:1128–32.

Zogby, Jonathan. 2007. "The New Polling Revolution: Opinion Researchers Overcome Their Hangups with Online Polling." *Campaigns and Elections* (May):16–19.

———. 2009. "For Interactive Polling, the Future Is Now." *Campaigns and Elections: Politics*. (June). http://politicsmagazine.com/magazine-issues/june-2009/for-interactive-polling-the-future-is-now/.