A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey

MATTHIAS SCHONLAU
RAND

KINGA ZAPERT
Harris Interactive

LISA PAYNE SIMON
California HealthCare Foundation

KATHERINE SANSTAD
University of California, San Francisco

SUE MARCUS
Mt. Sinai School of Medicine

JOHN ADAMS
MARK SPRANCA
RAND

HONGJUN KAN
Ingenix

RACHEL TURNER
The Wellcome Trust

SANDRA BERRY
RAND

The authors conducted a large-scale survey about health care twice, once as a web and once as a random digit dialing (RDD) phone survey. The web survey used a statistical technique, propensity scoring, to adjust for selection bias. Comparing the weighted responses from both surveys, there were no significant response differences in 8 of 37 questions. Web survey responses were significantly more likely to agree with RDD responses when the question asked about the respondent's personal health (9 times more likely), was a factual question (9 times more likely), and only had two as opposed to multiple

AUTHORS' NOTE: This study was funded by the California HealthCare Foundation. We are very grateful for the Foundation's support. The authors would like to thank George Terhanian and Michael Bosnjak for helpful suggestions on an earlier draft as well as the anonymous referees for very helpful comments.

Social Science Computer Review, Vol. 21 No. X, Season 2003 1-11
DOI: 10.1177/0894439303256551

response categories (17 times more likely). For three questions, significant differences turned insignificant when adjacent categories of multicategory questions were combined. Factual questions tended to also be questions with two rather than multiple response categories. More study is needed to isolate the effects of these two factors more clearly.

Keywords: *web survey; propensity weights; phone survey; Harris Interactive; weighting; poststratification; self-selection*

INTRODUCTION

This article compares the responses from a telephone random digit dialing (RDD) survey with a web survey containing identical questions. The survey concerns health care consumers in California.

In RDD surveys, computer-generated random telephone numbers are called. RDD surveys are commonly used for two reasons: First, the random selection mechanism ensures that the ensuing sample is a probability sample, meaning the probability with which each respondent is sampled is known. The probability mechanism makes valid inference about the entire population possible, beyond the sample at hand. RDD surveys reach both listed and unlisted numbers. Second, response rates tend to be relatively high (Dillman, 1978). A low response rate can jeopardize the validity of the inference because nonrespondents may differ from respondents. However, RDD surveys are inefficient because many of the random phone numbers are either not valid numbers, business numbers, phone booths, not part of the target population, or not valid for other reasons. They are also expensive because of the interviewer time and telephone charges for calls that do not produce interviews.

Similar to the Internet itself, web surveys are a more recent phenomenon. There are several ways of administering a survey over the Internet. One may send an e-mail with the survey and receive the completed survey via e-mail. One may send an e-mail with a pointer to a web page that hosts the survey. A third method is to elicit volunteers by advertising the survey on popular web sites or in news groups. Web-based surveys are usually preferable to e-mail surveys because they are easier to program and to administer.

Importantly, web surveys are less expensive than mail surveys and much less expensive than telephone surveys. However, many web surveys allow the participants to self-select into the sample. Web surveys are subject to selection bias and form a convenience sample rather than a probability sample—and therein lies the main criticism of web surveys. A statistical methodology called "propensity scoring" (Rosenbaum & Rubin, 1983) has been used in biostatistics to adjust for selection bias in observational studies since about 1980. Recently, Harris Interactive applied propensity scoring to adjust for bias stemming from respondent selection into web samples. Statistical theory states that if the selection bias is due to observed characteristics, it is possible to adjust for it. How this works in practice is the question we address in this article.

BACKGROUND

The literature on web surveys is growing rapidly. A comprehensive collection of references, upcoming conferences, and other materials has been put together by a web survey research team in Slovenia led by Vasja Vehovar (`http://www.Websm.org`). Schonlau,

Fricker, and Elliott (2002); Fricker and Schonlau (2002); Schaefer and Dillman (1998); and Couper, Blair, and Triplett (1999) gave an overview over studies using web or e-mail surveys.

Web surveys for general populations are currently not universally accepted because there is a debate about their scientific validity (Buckman, 2000). One requirement for scientific validity in probability samples is sufficient coverage of the target population. Dillman (2000) warned that the current coverage of most target populations is inadequate for e-mail or web surveys. Most published studies, however, report on web surveys conducted in closed populations (individual organizations or universities) where e-mail addresses are readily available (Couper et al., 1999). Therefore, coverage issues are less of a concern.

It is important to know whether web surveys yield the same estimates as phone surveys or mail surveys. The evidence is mixed. Taylor (2000) advocated the use of web-based surveys. He reported on several surveys that were conducted in parallel on the phone and on the web and found the weighted response estimates of phone surveys and web surveys to be within 5 percentage points in most cases. He used the same weighting methodology that is used in this study for the web survey. Varedian and Forsman (2002) compared a propensity-weighted web survey with a RDD survey. Both surveys asked about the use of hygiene products. As we will further explain in the discussion, their method differed somewhat from the Harris Interactive method that we describe in this article. Varedian and Forsman found that none of the weighting schemes applied (poststratification, propensity weighting with two different sets of variables) had a pronounced effect on the estimates stemming from the web survey. The web survey and the RDD responses were similar in one question and less so in two others. Bandilla, Bosnjak, and Altdorfer (2003) compared a web survey and a mail survey on "environmental attitudes and values" in Germany. Both surveys constituted a random sample; the web survey sample stemmed from a prerecruited Internet panel. After poststratification on demographic variables, 67% of the items were significantly different in the two survey modes. The number of significantly different items reduced to 27% when restricting the comparison to highly educated respondents. Vehovar, Batagelj, and Lozar (1999) compared weighted estimates of parallel surveys about electronic commerce. Considerable differences in responses were found between the phone survey respondents and web survey respondents. When comparing responses of web survey respondents with those of phone survey respondents who used the Internet, a lot of differences mostly disappeared. The study also concluded that weighting makes little difference in the estimates but noted that some of the weights were very large (exceeding 100) due to the considerable differences between online, telephone, and mail populations. Flemming and Sonner (1999) compared responses from a RDD telephone survey, a web survey of self-selected volunteers, and a web survey of respondents who were randomly selected for an earlier phone survey and who also agreed to fill out web surveys. They concluded:

> Even after weighting, there were a number of substantial differences between the online poll results and the telephone results. What's more, these differences did not follow any systematic patterns that might be explained by differences between Internet users and the public at large. (p. 8)

Many published studies use very small samples, some too small to be useful for inferential purposes. Schaefer and Dillman (1998) compiled a list of 13 studies using e-mail surveys. All but 2 of the studies reported sample sizes of less than 250 respondents. Some of the problems encountered with large-scale surveys were described in Schonlau, Asch, and Du (2003).

Our study looks at the extent propensity-weighted results differ from RDD results and what type of questions tends to lead to differences in the responses. Our sample is larger than that of many studies and our target population, residents of California, is not closed.

METHOD

RAND conducted an RDD telephone survey in California to collect consumer opinions about health care, health care utilization, and factors influencing decisions about health care choices in late 1999. Harris Interactive conducted the same survey over the web. All questions considered here were identical. To make the two surveys more comparable, the RDD survey was weighted using the same poststratification method as the web survey. We now describe the two surveys.

The Telephone (RDD) Survey

The RDD sample survey targeted adults in California households with telephones. In each call a random adult was selected using the Troldahl-Carter method of respondent selection (Czaja, Blair, & Sebestik, 1982), and a brief screener was administered. The screener was designed to select all respondents with public insurance; all African American, Asian, and other race respondents; and all who refused or were coded as "don't know" in response to race. Of the remaining respondents, we randomly selected one third with no insurance; of those remaining, we randomly selected one third with a medical visit in the previous 12 months; and of those remaining, we randomly selected one fifth of Whites and one half of Hispanic/Latinos.

The interview was conducted in English or Spanish. If a respondent declined to be interviewed, additional attempts were made to reach the household and complete the interview on different days of the week and different times of day. If possible, letters were sent to nonresponding households to encourage participation.

We attempted to call 26,222 phone numbers. Of these, 46% (11,955) were known ineligible phone numbers, the eligibility of 38% (9,877) of the numbers could not be determined with certainty, and 16% (4,390) were known eligible phone numbers. Of the households with unknown eligibility, 3,176 were numbers that rang but were unanswered after at least 10 calls, and 6,701 were households that could not be screened (including answering machines that did not appear to be businesses). Of the known eligible numbers, 93% were successfully interviewed, resulting in 4,089 completed surveys. The RAND RDD survey cost $51 per completed case.

The response rate was estimated as follows:

$$ResponseRate = \frac{Completes}{KnownEligibles + PercentEligibleOfScreened * UnscreenedHouseholds}$$

We estimated the response rate to be 43%. This calculation assumed that the numbers that were never answered were ineligible (pay phones or other unassigned numbers) and that the proportion of unscreened households that was eligible equals the proportion for screened households that were determined to be eligible (76%).

The sampling weight is the inverse of the sampling probability. The sampling probability was constructed from the probability that a phone number was drawn from a given county, from the probability that the respondent was selected based on the screener interview, and the number of phone lines in the household. We oversampled some counties to ensure that the

sample contained at least 900 respondents of the following subpopulations: 900 each of White, African American, and Asian; 900 in each of the three age categories (18 to 64, 64 to 74, and older than 75); 900 below the poverty line; and 900 in each of the categories of uninsured, Medicare recipients, and MediCal recipients. Oversampling resulted in a design effect (Kish, 1965) of 1.03, or only 3% loss of efficiency compared to a random sample.

We then matched race within area code according to the 1990 census data. In areas of high population density there may be several area codes in a single county. This step ensures that a race subpopulation in one area code does not get mixed together with the corresponding race subpopulation in another area code. As usual, the sampling weights were constructed as the inverse of the different selection probabilities that arise from this sampling scheme. The weights were then trimmed at the 95th percentile and renormalized.

The Web Survey

Harris Interactive selected 70,932 e-mail addresses of California residents from its database of people who have volunteered to receive surveys and who had not recently received a Harris survey. Of these, 81.5% were selected at random. The remainder was sampled from subpopulations where oversampling was desired. Specifically, the remaining 18.5% were divided up as follows: 4.3% African Americans, 7.5% Asian or Pacific Islanders, and 6.7% Hispanics. In addition, the respondents selected at random also contained minorities. The aged (65 years and older) and the respondents with low income (annual income less than $15,000) were also oversampled. Of 70,932 persons to whom e-mails were sent, 2% started the survey and did not finish, and 12% completed the survey. Only 234 respondents were not eligible either because they were not 18 years or older or did not reside in California. The final sample size was 8,195 eligible completes. The response rate for the web survey was 12%, where response rate means the number of completed surveys divided by the number of attempts. Because the web survey forms a convenience sample, the response rate is of little relevance. Even if the response rate as defined earlier were 100%, the sample would still be a convenience sample. The cost of the web survey was $10 per case based on the original target of 5,000 completed cases. The web survey easily obtained 8,195 eligible completes, and the extra cases were obtained free of charge.

The e-mail did not contain the survey but pointed at a password-protected web page containing the survey. The web page is only accessible for people with the individual password supplied in the e-mail. The survey was started on Thursday, February 24th, 2000. A second wave with the second half of the e-mails was sent one week later. Potential respondents had until Tuesday, March 21, 2000, to respond, which is when the site was shut down. Each potential respondent received one e-mail reminder.

Because this web survey did not constitute a probability sample, the weights were derived exclusively through poststratification. The poststratification matched the Current Population Survey for California within race for gender, age, race, income, health insurance, and in addition, for a variable derived from the propensity score as explained in the following.

Propensity Scoring for Web Surveys

Propensity scoring is a statistical technique (Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984) for adjusting for selection bias due to observed covariates. Selection bias arises in web surveys because not all members of the target population have Internet access and because respondents are volunteers and are not selected at random.

Propensity scoring has a strong theoretical foundation and is widely accepted within the statistical community (Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984). The propensity score is the conditional probability that a respondent is a web survey respondent rather than the respondent of a reference survey given observed covariates. This conditional probability is usually estimated via logistic regression.

To apply propensity scoring to web surveys, a RDD reference survey is needed. The RDD reference survey is generally short and only contains questions related to propensity scoring. It is used to calibrate the web survey responses (the Rand RDD survey on the other hand contains the full questionnaire). The observed covariates are questions that are asked both in the RDD reference survey and in the web survey. The questions used for propensity scoring measure general attitudes or behavior that are hypothesized to differ between the online and the general population. They are usually called attitudinal, lifestyle, or *webographic* questions. Examples for these questions are "Do you often feel alone? (*strongly agree*/*agree*/*not sure*/*disagree*/*strongly disagree*)" or "On how many separate occasions did you watch news programs on TV during the past 30 days?" A census or well-established surveys cannot be used as a reference survey because a census does not contain the attitudinal questions. Harris Interactive conducts reference surveys about once a month. A single reference survey can be used for many different web surveys as long as each web survey contains the attitudinal questions needed for propensity scoring. Ideally, the reference survey and the web survey have the same target population. In our study, Harris Interactive used their standard reference survey, which covers the entire United States, whereas the target population of web surveys was residents of California. Therefore, we have to make the additional assumption that the California population answers webographic questions just like the U.S. population. This assumption is not verifiable.

Web survey respondents and reference survey respondents with the same propensity score are "balanced" with respect to the attitudinal questions. This means that for respondents with the same propensity score, observed differences in attitudinal questions are due to chance rather than systematic bias. Propensity scoring balances observed covariates. Propensity scoring balances unobserved covariates only to the extent that they are correlated with observed covariates. The assumption that unobserved variables can be ignored with respect to selection bias is called *ignorability*.

In practice, the reference survey and web survey data sets are combined to a single data set. The conditional probability, or the propensity score, is computed by performing a logistic regression on the variables representing attitudinal questions using an indicator variable (web survey/RDD reference survey) as the outcome variable. Respondents for both surveys are sorted into five bins according to the propensity scores. Cochran (1968) showed five bins are sufficient to remove 90% of the removable bias. Weights are assigned such that the web survey's (weighted) proportion of respondents in each bin matches the reference survey's proportion in each bin. This is accomplished by constructing a categorical variable with five levels and treating this variable the same as any other poststratification variable. The propensity weights are computed before the data are poststratified.

Poststratification

We have applied the same poststratification procedure to both surveys. The poststratification ensures that both surveys match the Current Population Survey for California within race for gender, age, income, and health insurance status (yes/no).[1] In addition, the web survey also matches the five-level variable derived from propensity scoring to the corresponding variable in the reference survey.

Analysis Plan

We are interested in: (a) whether the weighted web survey responses are significantly different from the weighted RDD responses, (b) what type of questions/factors make it more likely that responses are similar, and (c) for questions with multiple response categories with significant differences, whether combining adjacent categories removes those differences.

To establish whether results are significantly different, we use *t* tests for questions that have only two response categories and χ^2 tests for multiple category questions.

In examining what type of questions tends to yield the same result, we consider several factors. First, is the question a factual question (Tourangeau, Rips, & Rasinski, 2000)? Second, does the question directly ask about the respondent's current or past personal health? And third, is the question a multicategory question as opposed to questions with only two response choices. (All but one question had two or more response choices; only one required a continuous numerical response.)

A factual question is a question that asks respondents about personal activities or circumstances as opposed to asking a respondent's opinion about a certain issue (Tourangeau et al., 2000). We now give examples that illustrate the attitudinal/factual and personal health/not personal health categories. For example, the commonly used question about a respondent's health status ("In general would you say your health is: *excellent, very good, good, fair, poor*") directly asks about the respondent's personal health, yet the perception about one's health status is attitudinal rather than factual. Questions that are both factual and directly ask about the respondent's personal health include: "Has the doctor ever told you that you had any of the following conditions (heart attack, cancer, diabetes, etc.)?" Questions that are factual but do not ask about the respondent's personal health include: "Have you ever been asked to rate the quality of a doctor hospital or health insurance plan in a survey?" Questions that do not directly ask about the respondent's personal health and are not factual include: "How much difficulty do you have right now in making choices or decisions about health care and medical needs (*no difficulty at all, a little difficulty, moderate difficulty, extreme difficulty*)?

We constructed indicator variables for factual, personal health, and multicategory and categorized each of 37 survey questions accordingly. We then computed the odds ratios for the hypothesized factors and tested whether the odds ratio was significantly different from 1. An odds ratio of 1 means that the question type does not affect whether the difference between web and RDD survey is significant.

Finally, for questions with significant differences and with more than two categories, we also investigated whether collapsing adjacent categories could eliminate the differences.

RESULTS

The web survey had 8,195 respondents; the RDD survey had 4,089 respondents. The RDD response rate was 43%. The web response rate was 11%, whereby response rate means the number of completed surveys divided by the number of attempted interviews.

In 8 of 37 questions the responses were not significantly different at $p = 0.01$. When using the Bonferroni correction[2] to account for multiple testing, the number of questions with insignificant differences increases to 11. The result without the Bonferroni correction is broken down by question category: factual (Table 1), personal health (Table 2), and whether the question had multiple categories (Table 3). The odds ratio for factual questions is 0.11 (significant at $p = 0.02$). The odds ratio for questions about personal health is also 0.11 (significant at $p = 0.02$). The odds ratio for multicategory questions is 17.3 (significant at $p = 0.004$).

TABLE 1
Number of Questions That Were Significant/Insignificant by Whether the Question Was Factual or Not

	Significant at 1%	
Factual Question	*No*	*Yes*
No	2	20
Yes	6	6

TABLE 2
Number of Questions That Were Significant/Insignificant by Whether the Question Asked About Personal Health

	Significant at 1%	
Ask About Personal Health?	*No*	*Yes*
No	2	19
Yes	6	7

TABLE 3
Number of Questions That Were Significant/Insignificant by Whether the Question Had Two or More Answer Categories

	Significant at 1%	
Multicategory Answer	*Yes*	*No*
Two answer choices	6	4
Multiple answer choices	2	22

This means that web survey responses are more likely to agree with RDD responses when the question concerns the respondent's personal health (9 times more likely), is a factual question (9 times more likely), and has only two response categories (17 times more likely). Factual questions tended to be two-category questions, meaning that further study is needed to isolate these two factors from one another more clearly.

When applying the Bonferroni correction, personal health questions and multicategory questions become even more significant ($p < 0.01$) even though the odds ratio of multicategory questions decreases to 11. The remaining p values and odds ratio results do not change much.

Because multicategory questions are more likely to lead to significant differences between web and RDD surveys, we investigated whether any two categories of multicategory questions could be combined such that the differences were no longer statistically significant. This was possible for 3 of the 22 multicategory questions that showed significant differences between the RDD and web survey responses. We will illustrate this with the question about health status (Table 4). The hypothesis that the two modes yield the same responses is rejected at $p < 0.0001$.

TABLE 4
Health Status by Survey Mode (in Percentages)

	Excellent	*Very Good*	*Combined Excellent and Very Good*	*Good*	*Fair*	*Poor*
Random digit dialing	22.7	32.9	55.6	27.3	13.9	3.3
Web	12.8	40.2	53.0	32.5	11.8	2.7
Difference	9.9	–7.3	2.6	–5.2	2.0	0.6

If we combine the categories *excellent* and *very good*, we eliminate most of the differences between the two modes. The hypothesis that the combined categories are the same is no longer rejected ($p = 0.056$). Clearly, combining *excellent* and *very good* is a post hoc idea. However, we believe that distinguishing between *excellent* and *very good* is harder than distinguishing between any two other categories. In this case, these two categories are subject to more measurement error, which is removed by combining the categories.

This was also possible for the question "How much of the time do you feel you have enough of the right kind of information for making choices or decisions about health care and medical needs?" Combining the categories *all of the time* and *most of the time* and leaving the remaining categories *some of the time* and *none of the time* unchanged rendered a highly significant difference ($p < 0.0001$) insignificant ($p = 0.33$).

For the question "How much difficulty do you have right now in making choices or decisions about health care and medical needs," combining the first two categories *no difficulty at all* and *a little difficulty* also rendered significant differences ($p < 0.0001$) insignificant ($p = 0.83$).

DISCUSSION

We conducted the same survey both as a RDD telephone survey and as a web survey. One hopes ideally that the two modes yield the same estimates for the target population. As the previous section shows, this is not generally the case. Differences might arise for many reasons. The fact that collapsing two similar response categories sometimes renders a significant difference insignificant may point to measurement error. Nonproprietary research on how to apply the propensity score methodology to web surveys is still in its infancy. The use of more or different attitudinal questions that feed the propensity scoring may lead to further improvements. There may be a mode effect, namely, respondents may respond differently online than they do on the phone. In our study, differences may also have arisen because of different target populations for the reference survey and the web survey. Finally, the implementation of the propensity scoring method may well have room for improvement.

There is currently no one right way of constructing propensity weights. Harris Interactive constructs a categorical variable from the propensity scores, which along with demographic variables is subsequently used for poststratification. Varedian and Forsman (2002) included the demographic variables in the logistic regression that yields the propensity scores. Their one-step procedure is more convenient than the two-step procedure that Harris Interactive uses, but it does not accomplish poststratification. That is, the weighted estimates do not necessarily match the distribution of the poststratification variables.

Irrespective of reasons that may contribute to differences, we have shown that several factors have large odds ratios and therefore strongly influence how likely the RDD and web surveys will yield the same result. This is remarkable given that this study was not confined to a well-controlled, closed population but instead conducted two large-scale, industrial-strength surveys. We believe that web surveys of general populations are here to stay because they can be conducted more timely (in our study 3.5 weeks vs. 3 months) and are much more cost efficient (in our study $10 vs. $51 per completed case). Our results indicate that propensity scoring for web surveys is a promising technology that warrants and needs more study.

NOTES

1. Poststratification ideally matches the joint distribution of the poststratification variables to a known joint distribution. Because of the curse of dimensionality even with large data sets, it is usually impossible to match the full joint distribution. Instead, most surveys just match the univariate distributions of the poststratification variables. Here we also match selected bivariate distributions.

2. On using the Bonferroni correction: Unlike in most applications where rejecting the null hypothesis is a finding, in our application not rejecting the null hypothesis (i.e., the two surveys are not significantly different) is of particular interest. Therefore, the Bonferroni correction is optimistic with respect to that goal rather than conservative.

REFERENCES

Bandilla, W., Bosnjak, M., & Altdorfer, P. (2003). Survey administration effects? A comparison of web-based and traditional written self-administered surveys using the ISSP environment module. *Social Science Computer Review*, *21*, 235-243.

Buckman, R. (2000, October 23). A matter of opinion. *The Wall Street Journal Online* [Electronic version].

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 295-313.

Couper, M. P., Blair, J., & Triplett, T. (1999). A comparison of mail and e-mail for a survey of employees in U.S. statistical agencies. *Journal of Official Statistics*, *15*, 39-56.

Czaja, R., Blair, J., & Sebestik, J. P. (1982). Respondent selection in a telephone survey: A comparison of three techniques. *Journal of Marketing Research*, *19*, 381-385.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: John Wiley.

Dillman, D. A. (2000). *Mail and Internet surveys*. New York: John Wiley.

Fricker, R., & Schonlau, M. (2002). Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods*, *14*, 347-367.

Flemming, G., & Sonner, M. (1999, May). *Can Internet polling work? Strategies for conducting public opinion surveys online*. Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg Beach, FL.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Rosenbaum, P. R. (2002). *Observational studies*. New York/Berlin: Springer-Verlag.

Rosenbaum, P. R., & Rubin. D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using sub classification on the propensity score. *Journal of the American Statistical Association*, *79*, 516-524.

Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology: Results of an experiment. *Public Opinion Quarterly*, *62*, 378-397.

Schonlau, M., Asch, B. J., & Du, C. (2003). Web surveys as part of a mixed mode strategy for populations that cannot be contacted by e-mail. *Social Science Computer Review*, *21*, 218-222.

Schonlau, M., Fricker, R., & Elliott, M. (2002). *Conducting research surveys via e-mail and the web*. Santa Monica, CA: RAND.

Taylor, H. (2000). Does Internet research "work"? Comparing on-line survey results with telephone surveys. *International Journal of Market Research*, *42*(1), 51-63.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Varedian, M., & Forsman, G. (2002). *Comparing propensity score weighting with other weighting methods: A case study on web data*. Paper presented at the American Association for Public Opinion Research Conference, St. Petersburg Beach, FL.

Vehovar, V., Batagelj, Z., & Lozar, K. (1999, May). *Web surveys: Can the weighting solve the problem?* Paper presented at the American Association for Public Opinion Research Conference, St. Petersburg, FL.

Matthias Schonlau is an associate statistician with the RAND corporation and head of the RAND statistical consulting service. His interests include web surveys, visualization, and data mining. He can be reached at matt@rand.org *or through his web site* www.schonlau.net.

Kinga Zapert is vice president of health policy research with Harris Interactive Inc. She has extensive experience conducting surveys with multiple health care constituencies (public, physicians, benefit managers, legislators, etc.) using a wide range of research methodologies, including web-based surveys. She can be reached at KZapert@harrisinteractive.com.

Lisa Payne Simon is acting director of the California HealthCare Foundation Quality Initiative. Ms. Simon also directs the Quality Initiative's comparative health systems performance measurement activities. She can be reached at lpsimon@chcf.org.

Katherine Haynes Sanstad, MBA, is interim deputy director for research for the San Francisco General Hospital Division of the University of California Department of Obstetrics, Gynecology, and Reproductive Sciences. Her background includes epidemiological and social science research in HIV and STI prevention and work on clinical quality of health care and health insurance markets. She can be reached at KHaynesSanstad@psg.ucsf.edu.

Sue Marcus is assistant professor of psychiatry and biomathematics at Mt. Sinai School of Medicine. Her research interests include methods for evaluating treatment efficacy in randomized and nonrandomized trials. She can be reached at sue.marcus@mssm.edu.

John Adams is a senior statistician with the RAND corporation. His interests include survey methodology and computationally intensive methods. He can be reached at john_adams@rand.org.

Mark Spranca is a behavioral scientist, director of the Center for Healthcare and the Internet, and group manager of behavioral and social sciences with the RAND corporation. His interests include consumer decision support, health care information technology, and consumer activation. He can be reached at spranca@rand.org.

Hongjun Kan is currently working as a researcher at Ingenix, Inc. The work was done while he was at RAND corporation. He is interested in health services research and risk adjustment in particular. He can be reached at hongjun_kan@uhc.com.

Rachel Turner is a policy adviser in the policy unit at the Wellcome Trust in London. She works on program evaluation and policy research and can be reached at r.turner@wellcome.ac.uk.

Sandra H. Berry is a senior behavioral scientist and senior director of the Survey Research Group at RAND. Her interests include health behaviors, experimental designs, and survey methods. She can be reached at berry@rand.org.