Selection Bias in Web Surveys and the Use of Propensity Scores

Matthias Schonlau, Arthur van Soest, Arie Kapteyn and Mick Couper

Sociological Methods Research 2009; 37; 291
DOI: 10.1177/0049124108327128

Sociological Methods & Research
Volume 37 Number 3
February 2009 291-318

10.1177/0049124108327128
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Selection Bias in Web Surveys and the Use of Propensity Scores

Matthias Schonlau
RAND Corporation, Pittsburgh, Pennsylvania
Arthur van Soest
Tilburg University, Netherlands
Arie Kapteyn
RAND Corporation, Santa Monica, California
Mick Couper
University of Michigan, Ann Arbor

Web surveys are a popular survey mode, but the subpopulation with Internet access may not represent the population of interest. The authors investigate whether adjusting using weights or matching on a small set of variables makes the distributions of target variables representative of the population. This application has a rich sampling design; the Internet sample is part of an existing probability sample, the Health and Retirement Study, that is representative of the U.S. population aged 50 and older. For the dichotomous variables investigated, the adjustment helps. On average, the sample means in the Internet access sample differ by 6.5 percent before and 3.7 percent after adjustment. Still, a large number of adjusted estimates remain significantly different from their target estimates based on the complete sample. This casts doubt on the common procedure to use only a few variables to correct for the selectivity of convenience samples.

Keywords: *Web surveys; selection; matching; propensity scores*

Internet interviewing opens up unique, new possibilities for empirical research in the social sciences. It creates opportunities to measure new or complex concepts (e.g., preferences, attitudes, expectations, and subjective probabilities) that are hard to measure with other interview modes and to design better measurement methods for existing standard concepts (e.g., income or wealth). Moreover, all this can be achieved in much shorter time frames than is customary in more traditional survey research.

Usually, empirical researchers in the social sciences have to use data collected by others or, if they want to collect data themselves, face time lags of often several years between the first draft of a questionnaire and the actual delivery of the data. Internet interviewing can reduce this time lag to a couple of weeks. The technology furthermore allows for follow-up data collection, preloading, feedback from respondents, and so forth. Moreover, experiments can be carried out of a similar nature as those in economics and psychology laboratories, but on a much larger scale and with broader samples than the convenience samples of undergraduate students typically used in such experiments (see Birnbaum 2004; Bellemare and Kroeger 2007). This alone changes the opportunities for empirical research in the social sciences dramatically. In addition, Internet interviewing creates new possibilities for quality enhancement and quality control. Last but not least, in comparison to other ways of collecting survey data, Internet interviewing turns out to be very cost-effective, especially if respondents can be contacted via e-mail, which in itself also expands possibilities for empirical research.

Any interview mode affects the probabilities of including respondents in a sample. Telephone surveys are facing increasing difficulties as it becomes harder to reach respondents directly because of, for example, the increased use of voice mail and cell phones (e.g., Oldendick and Link 1994; Link and Oldendick 1999; Berrens et al. 2003; Blumberg, Luke, and Cynamon 2004). Similarly, other modes such as in-person or mail-out surveys have their own well-known drawbacks, including response rates that show a decreasing trend (e.g., see the international overview in De Heer 1999). The same type of problem obviously also applies to Internet interviewing since it requires respondents to have Internet access. In addition, Internet surveys are probably not immune from the response rate trends affecting other modes.

A distinction needs to be made between coverage error, nonresponse error, and random sampling error (Groves 1989). This article focuses on coverage error. Coverage error and nonresponse error may lead to biased estimates, whereas sampling error is due to random variation. Even a simple random sample (SRS) with equal selection probabilities will lead to

Authors' Note: Support for this research comes from Grant R01AG20717 from the National Institute of Aging of the U.S. National Institutes to RAND (Arie Kapteyn, Principal Investigator) and from the University of Michigan's Survey Research Center (Robert J. Willis, Principal Investigator). The authors are grateful to three anonymous referees for constructive comments. Please address correspondence to Matthias Schonlau, RAND, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15215; e-mail: matt@rand.org.

sampling error because only a subset of the population is sampled. Nonresponse error arises when a group does not answer a given question of interest (item nonresponse) or does not participate in the survey at all (unit nonresponse). In this case, for example, the sample mean ignoring the nonrespondents typically will be a biased estimator for the population mean (including the nonrespondents), if respondents differ from nonrespondents on the variable of interest (e.g., Groves and Couper 1998). Coverage error arises when the survey is designed such that a specific part of the target population is not included in the frame. For example, if respondents are selected by randomly dialing telephone numbers (random digit dialing, or RDD), households without a phone will never be selected though they are included in the target population. If, for example, the purpose is measuring average household income and the households without a phone have a lower average income than others, ignoring the group with no phones will lead to an upward bias in the estimate of average household income in the population. Of course, the size of the bias would probably be limited in this example if the group of households without a phone connection is small.

At this point in time, for Internet surveys, coverage problems probably play a larger role than for RDD surveys (Couper et al. 2007). Because Internet use is not yet equally spread among all socioeconomic and demographic groups, the coverage problem is likely to lead to biased estimates on variables related to socioeconomic status (SES). This may be a particular problem if the target population is the elderly population, where Internet access is less widespread than in the population as a whole. One way to address this problem is to provide households without Internet access with the tools to get access. This is, for example, done by Knowledge Networks and the RAND Corporation American Life Panel in the United States and by CentERdata in the Netherlands. While this avoids the coverage error, it is still subject to (sampling and) nonresponse error. For Knowledge Networks, multiple levels of nonresponse lead to overall response rates of substantially less than 30 percent (Huggins and Krotki 2001).

Some Internet surveys have sampling frames that are not subject to coverage error. Internet surveys with a good sampling frame typically arise for closed populations such as companies, universities, or professional associations. In these institutions it is easy to identify e-mail addresses, which can be used to contact potential respondents.

Today there are many Internet-based samples used to conduct surveys of various kinds. Typically, no attempt is made to make these samples cover more than the population of active Internet users. For example, prospective respondents may be recruited by e-mail or by placing banners

on frequently visited Web sites. There are obvious problems with such samples (cf. Couper 2000), which are often ignored (Schonlau, Fricker, and Elliott 2002). Not only are the respondents a selective sample of the population at large, they are the most savvy computer users and may therefore be much quicker to understand and answer Internet interview questions than others. Because they may respond differently, one needs to find a way to validly generalize from such a sample to a broader population.

An important tool to correct for the selection effect in observational studies is weighting on the basis of propensity scores (Rosenbaum and Rubin 1983; Little and Rubin 2002). Harris Interactive uses propensity scoring methodology to reweigh a convenience Web sample based on a monthly random phone sample using various sets of about five "webographic" variables (Taylor 2000). Webographic questions are questions that are thought to best capture the differences between the general population and people able and willing to answer Web surveys. The use of propensity scores for surveys requires two samples: a random sample for calibration and a second sample that is calibrated. An alternative to propensity weighting is matching; for example, see the recently developed matching algorithm of Diamond and Sekhon (2005).

This study investigates the usefulness and validity of propensity scores and matching to correct for the selective nature of the subsample of respondents with Internet access in the Health and Retirement Study (HRS). The HRS is a representative survey of elderly cohorts in the United States (with adjustment weights to correct for unit nonresponse and race-, ethnicity-, and age-stratified sampling). A randomly selected group from the subsample of HRS respondents in 2002 who reported having Internet access was invited to participate in an Internet survey in 2003. This experiment is unique in that a vast amount of information is already available on all respondents from the core HRS, irrespective of whether they were included in the Internet survey or not, which greatly enhances the scope for analyzing selection and mode effects in Internet interviewing. It also helps enormously to study how well propensity scores or matching on the basis of a limited set of HRS variables performs in correcting for selectivity in other variables.

In this article, we do not look at the HRS Internet survey but focus on the selection concerning Internet access and the matching and propensity weighting procedures to correct for this. That is, we only use the data from the core HRS in 2002 to investigate the selective nature of the subsample of respondents who reported having Internet access. Not all of them subsequently participated in the Internet survey because only a random subset was invited to participate and because of unit nonresponse. In the current study, we focus on coverage errors and not on nonresponse errors. We analyze how weighting

and matching on a limited set of variables can be used to obtain balance between the Internet access and non–Internet access subsamples, in the sense that weighted statistics of the variables used for the adjustment for the two subsamples are similar. We then investigate to what extent the adjustment also helps to correct for imbalances in other variables not used in the corrections.

The remainder of this article is organized as follows. The next section provides background information on the core HRS and the HRS Internet survey. In the following section, we discuss propensity scoring and our methods to investigate whether propensity scoring is a useful way to correct for selection effects. We then present empirical results and conclusions.

Data Source

The University of Michigan HRS surveys more than 22,000 Americans aged 50 and older every 2 years. The study paints a portrait of an aging America's physical and mental health, health insurance coverage, use of health care, SES, income, wealth and portfolio choice, labor market position, job characteristics, family networks, and family transfers. It started in 1992 with the 1931-1941 birth cohort (for more information on the first wave, see Juster and Suzman 1995). Other cohorts were added later so that the 1998 sample covered the complete U.S. population aged 51 and older and their spouses. In this study we use the HRS wave of 2002, covering the population aged 55 and older and their spouses. The first wave of HRS was conducted using computer-assisted personal interviews. Follow-up surveys were mainly done by telephone, but respondents older than 80 years and households who had no phone were interviewed in person.

To use a sample to draw inference on the population of interest, the sample design needs to have certain characteristics. An SRS in which each member of the population is drawn with equal probability, independently of other members of the population, makes it possible to apply standard textbook procedures and leads to narrower confidence intervals than a sample with unequal probabilities of selection. The SRS design is rare in the practice of social science surveys, due to, for example, regional stratification and unit nonresponse. It then helps to have information on the stratified design and some characteristics of the unit nonrespondents or to have an external source that can be used to determine the size of population segments characterized by, for example, age, ethnicity, and gender. Such information can be used to construct adjustment weights for all observations in the sample. Under the assumption that unit nonresponse and other potential

sources of bias are not related to the variables of interest conditional on the information incorporated in the adjustment weights, the weights can be used to correct for the bias.

The HRS uses adjustment weights based on an external source, the March samples of the Current Population Survey.[1] Weights are constructed first at a household level, using initial sampling probabilities and the birth years and race/ethnicity of the male and female household members, and then at the respondent level (for details, see Health and Retirement Study 2002). Thus, the only information used in the weights is birth year, gender, race, ethnicity, and marital status. The analysis in the current study is at the household level and uses the household-level sampling weights. It is a maintained assumption that these weights appropriately correct for the nonrandom nature of the core HRS for all our variables of interest. What we focus on is the selective nature of the Internet-access subsample of the core HRS.

Because of the cost-effectiveness and other advantages of Internet interviewing, the University of Michigan and RAND set up a pilot project with the overall goal to explore the feasibility of using the Internet to supplement interviewer-administered data collection in the HRS and to explore a variety of methodological issues related to Web-based measurement. The 2002 wave of the HRS contained 16,698 respondents (excluding respondents with zero respondent-level weight). Of these, 29.7 percent reported having Internet access.

Participation in the HRS Internet survey depends not only on Internet access but also on a sequence of selection steps: Internet access, willingness to participate given access, random selection into the group that gets the letter inviting them to participate, and nonresponse given willingness to participate. Couper et al. (2007) look at the several stages in detail and find that Internet access is clearly the most selective step in terms of demographics and SES variables. We will therefore focus on Internet access, not on whether people with access actually participated in the survey. There are 11,279 households in the HRS 2002. In households with more than one respondent, we choose the financial respondent; in the few cases there was more than one financial respondent, we choose a financial respondent at random. Throughout, we compare estimates based on respondents with Internet access, adjusted estimates based on respondents with Internet access, and unadjusted estimates based on the full HRS 2002 sample.

Nowhere in this study do we use the data of the actual Web survey. This makes it possible to study selection issues without having to account for potential mode effects—the possibility that answers to the same question may differ depending on whether the question is asked by phone, in person,

or over the Internet (cf. Schwarz and Sudman 1992). Mode effects are certainly another important issue in selecting the mode of interviewing but need not be considered for the research questions on correcting for selection addressed in the current article.

Most Web surveys do not have an underlying sampling frame like the core HRS. Usually, a convenience sample rather than a random sample or a probability sample is selected. Drawing inference from convenience samples, including estimates of population frequencies and percentages, is a hard problem that is often neglected (also see Schonlau et al. 2002; Butz and Torrey 2006). Drawing inference from observational studies is common in biostatistics because the randomization required for experiments can be unethical when dealing with human participants or difficult to achieve in practice. Propensity scoring (Rosenbaum and Rubin 1983; Rosenbaum 2002) is commonly used to draw inference from observational data. Propensity scoring has also found its place in the survey literature (Czajka et al. 1992; Battaglia et al. 1995; Duncan and Stasny 2001; Smith et al. 2001; Garren and Chang 2002; Iannacchione 2003). Sobel (1995), Winship and Morgan (1999), and Winship and Sobel (2006) give overviews of causal inference targeted at the social and behavioral sciences.

Harris Interactive, a commercial Web survey company, has adopted this approach for the use of Web surveys. The Harris Interactive approach involves partitioning the estimated propensity score from the sample combining the Web and reference sample into a categorical variable. This and other variables are then reweighed such that after the adjustment the marginal distributions of the variables are the same for the Web survey and the reference survey.

The Harris Interactive approach is described in more detail in Schonlau et al. (2004). Application of propensity scores in the context of Web surveys is also described by Danielsson (2004). The central issue is whether and under what circumstances propensity-adjusted estimates are comparable to those based on random samples. An integral component of the issue is what questions should be asked to capture the difference between the respondents with Internet access and the population of interest. As mentioned, Harris Interactive calls these elusive questions webographic questions, comprising both demographic and lifestyle questions. Other researchers call them ''lifestyle'' or ''attitudinal'' questions.

Jointly with Harris Interactive, Duffy et al. (2005) compare several face-to-face and propensity-adjusted online surveys in the United Kingdom. They hypothesize that differences may arise because of social desirability bias, interviewer effects, mode effects, and differences in how response

scales are used. A popular application for the propensity scoring adjustment is forecasting election results (Duffy et al. 2005). Isaksson and Forsman (2003) study political polls for the 2002 election in Sweden. They find that propensity adjustment for forecasting election results works better than the usual poststratification. Yoshimura (2004) presents an application estimating ownership rates of several types of electronic devices in Japan and shows that adjusting Web-based convenience sample estimates using inverse propensity score weighting greatly reduces the differences compared to probability sample–based benchmark estimates.

Varedian and Forsman (2003) investigate the efficacy of propensity score weighting in the context of a marketing survey about the use of hygiene products and attitudes toward local banks. A phone survey ($N = 347$) and a Web survey ($N = 4,724$) were conducted in a northern European country. Their survey included lifestyle questions that were trying to capture a respondent's "modernity." They use logistic regression on lifestyle questions and demographic questions to capture the selection effect. They conclude that the estimates obtained from Web and RDD phone surveys are different. Furthermore, various weighting schemes did not change the results very much.

Schonlau et al. (2004) compared estimates from an RDD phone survey with propensity-adjusted estimates from a Web survey conducted by Harris Interactive. They found that 8 out of 37 estimates investigated were not significantly different. Estimates from the Web survey were significantly more likely to agree with estimates from the RDD phone survey for factual questions, when the question concerned the respondent's personal health, and when the question contained two as opposed to multiple categories.

For the 2002 wave of the HRS, which did not have the specific webographic questions asked by Harris Interactive, we use demographic questions, health-related questions, and others that were available in the 2002 wave of the HRS.

Propensity Scoring and Matching

Here we follow as much as possible Little and Rubin (2002, chap. 3). Let Y be a (vector of) variable(s) of interest, in our case, a set of specific health and health behavior variables and asset amounts (see Table 3 below), and let X be a set of covariates, in our case, race, gender, age, income, self-assessed general health, and home ownership (cf. Table 2 below). Let I denote the dummy variable indicating whether someone has Internet access ($I = 1$) or not ($I = 0$). The propensity score is defined as $P(I|X)$. To use the propensity

scores in constructing adjustment weights, we need the assumption of conditional independence (CI):

$$Y \text{ and } I \text{ are conditionally independent given } X. \qquad \text{(CI-a)}$$

Since I is a binary variable, this can also be written as

$$Y|X, I=1 \text{ has the same distribution as } Y|X, I=0 \text{ for almost all } X. \qquad \text{(CI-b)}$$

Using Bayes rule, this can be rewritten as

$$P(I=1|X, Y) = P(I=1|X) \text{ for almost all } X \text{ and } Y. \qquad \text{(CI-c)}$$

CI is also known as strong ignorability. Under CI, we can combine the HRS weights with inverse propensity scores for Internet access to construct consistent estimators for parameters of the population distribution of Y, generalizing the standard case in Rosenbaum (2002) or Little and Rubin, where propensity weights are based on an SRS. To illustrate this without introducing too much notation, we assume the parameter of interest is the population mean μ_Y. The HRS weights are a combination of sampling weights and additional adjustments based on a few basic demographic characteristics. These demographic characteristics form a subset Z of the complete vector of conditioning variables X. Thus, the HRS adjustment weight is a function of Z; it will be denoted by $w^{HRS} = w^{HRS}(Z)$. The HRS weights are assumed to be proportional to the inverse of the inclusion probabilities of the HRS sample, implying that they provide sufficient adjustment for the case where Y would be observed for the complete HRS sample, consisting of respondents with and without Internet access. Under this assumption, a weighted mean of all HRS sample observations gives an approximately unbiased estimator of μ_Y (Little and Rubin 2002:46).

Under CI, we can combine the HRS weights with the inverse propensity scores for Internet access, $w^{ps} = w^{ps}(X) = P(I=1|X)^{-1}$, to construct an approximately unbiased estimator based on a subsample of respondents with Internet access, as in Little and Rubin (2002:46, equation 3.4). The combined adjustment weights are given by $w = w(X) = w^{HRS}(Z)w^{ps}(X)$. Note that here Z is contained in X. Thus, under CI and the maintained assumption that the HRS weights are appropriate to make HRS representative of the population of interest (cf. the Data Source section), a consistent estimator for μ_Y will be given by the weighted mean over the subsample of respondents with Internet access:

$$\bar{y}_w = \sum_{i=1}^{N} w_i I_i Y_i / \sum_{i=1}^{N} w_i I_i = \sum_{i=1}^{N_I} w_i Y_i / \sum_{i=1}^{N_I} w_i,$$

where the first summation is over the complete sample and the last is over only the subsample of respondents with Internet access (with N_I observations). In the empirical work, we will compare this estimator with two alternative estimators of μ_Y. The first is the unadjusted estimator using respondents with Internet access only, given by

$$\overline{y}_{unadj} = \sum_{i=1}^{N_I} w_i^{HRS} Y_i / \sum_{i=1}^{N_I} w_i^{HRS},$$

where the summation is over the subsample of respondents with Internet access. This is the estimator that accounts for the stratified nature of the HRS, but it assumes that Internet access is completely random. It does not correct for selective Internet access and will generally be inconsistent under CI.

In this specific case, we are in the fortunate situation of having not only the respondents with Internet access but also the respondents without Internet access. The benchmark estimator for this case uses all the observations and is given by

$$\overline{y}_{full} = \sum_{i=1}^{N} w_i^{HRS} Y_i / \sum_{i=1}^{N} w_i^{HRS},$$

where the summation is over the full HRS sample.

In the usual case of an Internet survey, the latter estimator is not available since Y is not observed for the subsample of respondents without Internet access. The specific design we have makes it possible to compute this estimator and compare it to the other two. Comparing the adjusted estimator $\overline{y}_w$ to the benchmark $\overline{y}_{full}$ gives a test of CI: Under CI, both estimates are consistent estimates of μ_Y and should thus be similar. Similarly, comparing the unadjusted estimator $\overline{y}_{unadj}$ to $\overline{y}_{full}$ makes it possible to test whether Internet participation is unconditionally independent of Y. Of course, unconditional independence could also be tested against CI by comparing $\overline{y}_{unadj}$ to $\overline{y}_w$ so that for this, the specific design is unnecessary. The added value of the design is thus that it makes it possible to test CI.

The assumption of CI is crucial. Note that if this holds for a given set of conditioning variables X, it will also hold for any larger set. This leads to the idea of selecting a minimal set X such that CI holds for a large enough set of Y variables of interest. Once such a set X is found, it is sufficient to have observations of respondents with Internet access on Y and propensity scores based on X. This is where the potential efficiency gain of Internet surveying is situated. We know that Internet access is selective,

but if we can find a relatively small set of conditioning variables X, we can still use a Web survey to draw population inference on Y. All we need is a representative survey measuring X to construct the propensity scores, and the smaller the set of variables needed in X, the larger is the efficiency gain.[2] We have preselected a number of potential variables with information in various domains (health, economic status, family composition).

The standard way of computing propensity scores, $\hat{p} = \hat{P}(I = 1|X)$, is to compute predicted values of the logistic regression of the indicator variable for respondents with Internet access on covariates. The propensity score can be used in several ways. We analyze the same data twice using two different methods. First, we use the inverse propensity scores as weights (Rosenbaum 1987; McCaffrey, Ridgeway, and Morral 2004). The inverse propensity scores are multiplied with the HRS weights to get the ultimate adjustment weights used for balancing, as explained above.

Second, we use the "Genetic Matching" algorithm "GenMatch," by Diamond and Sekhon (2005). This computationally intensive algorithm often achieves better balance, that is, it produces a matching that makes the matched sample closer to the sample it is matched with in terms of the distributions of the matching variables than matching on, for example, inverse propensity scores. Briefly, the algorithm uses the *weighted* Mahalanobis distance to match a respondent with Internet access to one without Internet access. The algorithm iteratively selects a diagonal matrix of weights maximizing the minimum p value observed across a number of balance tests performed on distributions of matched baseline covariates (univariate baseline covariates and, optionally, interactions and quadratic terms). The same respondent with Internet access may be matched multiple times to different respondents without Internet access. Once Mahalanobis weights are determined, the best matching respondents can be determined, and then the degree of balance can be determined. The key task is to choose the weights such that the resulting match improves balance. For dichotomous variables, the balance test consists of a t test. For continuous variables, the balance of the entire distribution can be tested using a weighted version of a bootstrapped Kolmogorov-Smirnov test. In this version, the probability of choosing an observation for the bootstrap is $1/(N_I w^{HRS})$. The optimization of the weights uses an evolutionary algorithm, and this makes the procedure computationally intensive. We use all X variables in Table 1 and the linear propensity score for the matching algorithm. We do not use any interactions or squares. The algorithm uses a starting value for each variable, including the propensity score, in the optimization. Starting values can be automatically generated or can optionally be specified. We found that choosing a

high starting value for the propensity score made the algorithm perform best for our data.

From the Genetic Matching algorithm, we obtain observations of respondents with Internet access that match the respondents without Internet access. The same respondent with Internet access observation may be used in multiple matches. The matching estimator for μ_Y, using observations on X for the complete sample but observations on Y for the subsample of respondents with Internet access only, is now given by

$$\overline{y}_{adj} = \frac{1}{\sum_{i \in W \cup NW} w_i^{HRS}} \left\{ \sum_{i \in W} w_i^{HRS} Y_i + \sum_{i \in NW} w_i^{HRS} Y_{m(i)} \right\},$$

where W (Web) is the sample of respondents with Internet access and NW (no Web) is the sample of respondents without Internet access. $Y_{m(i)}$ is an observation from a respondent with Internet access matched to the ith respondent without Internet access. Note that $\overline{y}_{adj}$ uses the propensity scores only for selecting matched observations, not in the computation of $\overline{y}_{adj}$ itself. Under the null hypothesis of CI, $\overline{y}_{adj}$ is consistent for μ_Y. Comparing $\overline{y}_{adj}$ with the mean $\overline{y}_{full}$ using all observations will give a test of CI.

The Genetic Matching algorithm is useful because it finds balance in an automatic fashion. The drawback to Genetic Matching is that the algorithm can be very time consuming. It is currently implemented only in R and therefore requires programming skills in R.

Propensity scores can be used in different ways. Other options include stratification on the propensity score (Cochran 1968) and matching on the propensity score itself. The former is used by, for example, Harris Interactive (see Schonlau et al. 2004). Although a comparison with their approach would be interesting here in principle, we do not pursue this since we do not have their webographic variables for the non-Internet sample.

Results

Table 1 shows the variables we use as covariates for the adjustment: demographics (race/ethnicity, gender, dummies for several education levels, age), marital status, personal income, an indicator of home ownership, and self-assessed health. Age was transformed into a small number of categorical dummy variables, allowing for nonlinear and nonmonotonic effects. Self-assessed health is a categorical variable but was transformed into dummies for excellent, good, fair, and poor (with very good as the

Table 1
Logistic Regression of Internet Access on Various Covariates

Covariate		Odds Ratio	*p* Value
Race/ethnicity	White		
	African American	0.36	.000
	Other race	0.72	.074
	Hispanic	0.28	.000
Gender	Female		
	Male	0.85	.003
Education	<High school	0.32	.000
	High school		
	Some college	2.17	.000
	≥College	3.32	.000
Age	≤55	1.49	.000
	56-65		
	66-75	0.50	.000
	>75	0.20	.000
Marital status	Married		
	Separated, divorced, widowed	0.61	.000
	Never married	0.57	.000
Self-assessed health	Excellent	1.19	.029
	Very good		
	Good	0.72	.000
	Fair	0.55	.000
	Poor	0.42	.000
Income	Indicator (income = 0)	19.22	.000
	Log 10 income	2.18	.000
Owns a house		1.30	.001

benchmark). Log income is the only variable that is not dichotomous. In addition to log personal income, we also include a dummy variable for whether income equals zero.

To assess whether these covariates affect Internet access, we regress the indicator variable of whether a respondent had Internet access on the covariates. Table 1 shows odds ratios and *p* values for this logistic regression. The *p* values refer to the test that the corresponding regression coefficient is 0, that is, that the odds ratio is equal to 1. All variables but one are significant at the 5 percent level, "Other race" being the exception. The common support, the overlap in the range of the predicted values for respondents with and without Internet access, is good; 99.3 percent of the predicted values are contained within the interval that has common support.

In this age group in 2002, Internet access was still fairly limited (29.7 percent with access). But even within this age group, Internet access falls steeply with age. Non-Hispanic Whites have greater access than Hispanics, African Americans, and other ethnicities. Internet access rises substantially with education level: Someone with less than a high school education has a probability of having Internet access that is only about one fourth of the probability of someone with a high school education and who is identical on other characteristics, and about one tenth of the probability of someone with at least a college degree. Large and significant effects are also found for marital status dummies, with the largest probability of Internet access for married people. The probability of Internet access also rises significantly with income. A strong negative effect of health problems is found. For people who report that they are in fair or poor health, the probability of having Internet access is about half that of otherwise similar people in excellent health. Finally, homeowners are more likely to have Internet access than renters.

One of the strengths of the approach we take is that imbalances in the separate covariates are made explicit. Table 2 displays balance before and after the adjustments using both methods, for all covariates used in constructing propensity scores and in the matching procedure. While we are interested in estimates for the full population, balance refers to differences between respondents with and without Internet access. For both methods, the difference between respondents with and without Internet access is greatly reduced. Both adjustment procedures do what they are designed for—they achieve balance for the set of covariates (X in the Propensity Scoring and Matching section) used in constructing the weights and selecting the matches.

Under the CI assumption, the adjustment should also help to obtain balance for other variables of interest (Y in the Propensity Scoring and Matching section) not used in constructing the weights or selecting the matches. This is studied in Tables 3 and 4. Table 3 gives the estimates of the population means of a number of binary variables indicating physical or mental health problems and health-related limitations in activities of daily living. The first column is based on the full sample ($\overline{y}_{full}$). The second column is based on respondents with Internet access only, not correcting for the selectivity of Web access ($\overline{y}_{unadj}$). The fifth column has the adjusted estimates for respondents with Internet access using the GenMatch algorithm ($\overline{y}_{adj}$). The eighth column gives an adjusted estimate based on the use of propensity scores as weights ($\overline{y}_{w}$). The differences refer to the unadjusted/adjusted estimates minus the full-sample estimates.

The unadjusted estimates of the prevalence rates of health problems based on respondents with Internet access in the second column are up to

Table 2
Balance Before and After the Adjustment

		Unadjusted			Adjusted Using GenMatch Algorithm	Adjusted Using Inverse Propensity Scores as Weights
		R With Internet Access	R Without Internet Access	Difference	Difference	Difference
Propensity score		0.02	−1.84	1.86	0.046	−0.080
Race/ethnicity	White					
	African American	4.4%	12.3%	−7.9%	−1.1%	−0.9%
	Other race	1.7%	1.9%	−0.2%	0.0%	1.1%
	Hispanic	2.4%	7.9%	−5.5%	0.2%	4.0%
Gender	Female					
	Male	41.4%	35.4%	6.1%	2.1%	1.3%
Education	<High school	3.7%	29.0%	−25.3%	−1.8%	0.7%
	High school					
	Some college	29.2%	17.8%	11.4%	−0.1%	0.1%
	≥College	38.3%	12.8%	25.5%	−0.2%	−1.0%
Marital status	Married					
	Separated, divorced, widowed	26.5%	48.9%	−22.4%	−0.6%	0.8%
	Never married	3.8%	4.7%	−0.9%	−0.4%	−0.4%
Age	<55	23.7%	8.7%	15.0%	−1.4%	−0.9%
	55-65					
	65-75	20.3%	28.1%	−7.7%	−1.2%	−2.5%
	>75	8.1%	33.3%	−25.2%	1.7%	0.1%

(continued)

Table 2 (continued)

		Unadjusted			Adjusted Using GenMatch Algorithm	Adjusted Using Inverse Propensity Scores as Weights
		R With Internet Access	R Without Internet Access	Difference	Difference	Difference
Self-assessed health	Excellent	21.0%	8.7%	12.2%	0.0%	0.3%
	Very good					
	Good	27.5%	34.1%	−6.6%	1.6%	−1.0%
	Fair	9.7%	21.8%	−12.1%	−1.4%	−0.2%
	Poor	2.9%	9.9%	−7.0%	−1.3%	1.6%
Income	Indicator (income = 0)	13.6%	13.3%	0.3%	−2.5%	0.8%
	Log 10 income	381.7%	358.3%	0.23%	0.13%	−0.041%
Owns a house	Indicator	89.0%	74.8%	14.2%	2.2%	−2.7%

Note: R = respondent.

Table 3
Differences in Prevalence of Comorbidities, Symptoms of Mental Health Problems, and Limitations in Activities of Daily Living

			Unadjusted			Adjusted Using GenMatch			Adjusted Using Inverse Propensity Scores as Weights		
		Full Sample	R With Internet Access	Diff	*p*	R With Internet Access	Diff	*p*	R With Internet Access	Diff	*p*
Comorbidities	High blood pressure	54.8%	44.0%	−10.8%	.000	57.4%	2.6%	.617	51.7%	−3.1%	.969
	Lung disease	10.2%	7.1%	−3.1%	.000	11.0%	0.8%	.847	11.0%	0.8%	.439
	Heart disease	25.3%	16.0%	−9.3%	.000	23.4%	−1.9%	.891	22.5%	−2.8%	.465
	Stroke	7.5%	3.7%	−3.8%	.000	7.1%	−0.4%	.242	5.8%	−1.7%	.116
	Cancer	14.4%	12.4%	−1.9%	.082	14.5%	0.1%	.416	15.5%	1.2%	.067
	Diabetes	16.9%	11.6%	−5.3%	.000	13.4%	−3.4%	.132	14.2%	−2.7%	.061
	Arthritis	61.9%	48.9%	−13.0%	.000	65.3%	3.4%	.990	58.4%	−3.4%	.588
	Ever had psychiatric problems	16.7%	14.1%	−2.5%	.001	15.5%	−1.2%	.193	16.3%	−0.3%	.789
Mental health	Depressed	18.6%	10.6%	−8.0%	.000	14.6%	−4.0%	.005	15.3%	−3.2%	.196
	Lonely	21.2%	11.7%	−9.5%	.000	20.3%	−0.9%	.031	17.7%	−3.6%	.293
	Happy	89.7%	88.6%	−1.1%	.000	93.5%	3.8%	.014	87.4%	−2.3%	.346
	Sad	22.6%	16.3%	−6.3%	.000	22.4%	−0.2%	.368	20.6%	−2.1%	.494
	Effort	26.0%	13.4%	−12.6%	.000	19.9%	−6.1%	.000	19.2%	−6.8%	.002
	Sleep was restless	30.0%	25.2%	−4.8%	.000	31.4%	1.4%	.632	28.2%	−1.8%	.840
	Enjoys life	95.7%	93.3%	−2.4%	.011	97.2%	1.5%	.068	92.7%	−3.0%	.668
	Could not get going	24.2%	15.8%	−8.4%	.000	21.4%	−2.9%	.199	22.9%	−1.4%	.562

(continued)

Table 3 (continued)

			Unadjusted			Adjusted Using GenMatch			Adjusted Using Inverse Propensity Scores as Weights		
		Full Sample	R With Internet Access	Diff	*p*	R With Internet Access	Diff	*p*	R With Internet Access	Diff	*p*
Activities of daily living	Difficulties with. . .										
	Dressing	8.7%	3.7%	−5.1%	.000	6.0%	−2.7%	.262	7.0%	−1.7%	.378
	Walking across room	6.4%	2.2%	−4.2%	.000	5.9%	−0.5%	.104	3.6%	−2.8%	.011
	Bathing/showering	5.8%	1.4%	−4.4%	.000	3.4%	−2.4%	.000	2.0%	−3.8%	.000
	Eating	2.1%	0.5%	−1.6%	.000	0.5%	−1.6%	.001	0.7%	−1.5%	.000
	Getting in/out of bed	5.6%	2.7%	−2.9%	.000	6.5%	0.9%	.273	4.1%	−1.5%	.260
	Using the toilet	5.4%	1.9%	−3.5%	.000	2.8%	−2.6%	.003	2.6%	−2.8%	.003
	Preparing hot meals	5.1%	1.2%	−3.9%	.000	2.2%	−2.9%	.000	1.9%	−3.2%	.000
	Grocery shopping	8.7%	2.3%	−6.4%	.000	3.7%	−5.0%	.000	3.2%	−5.5%	.000
	Using the phone	2.7%	0.4%	−2.3%	.000	0.3%	−2.4%	.000	0.5%	−2.2%	.000
	Taking medications	2.1%	0.7%	−1.3%	.000	0.6%	−1.5%	.003	0.6%	−1.4%	.001
	Managing money	4.7%	1.3%	−3.4%	.000	1.5%	−3.1%	.000	1.3%	−3.3%	.000
	Walking several blocks	31.2%	14.9%	−16.2%	.000	25.6%	−5.6%	.004	26.7%	−4.5%	.076
	Walking one block	14.4%	5.9%	−8.5%	.000	12.9%	−1.5%	.604	12.4%	−2.0%	.236
	Sitting for 2 hours	20.2%	12.7%	−7.5%	.000	16.0%	−4.1%	.035	17.8%	−2.4%	.122
	Getting up from chair	41.3%	28.9%	−12.3%	.000	39.4%	−1.9%	.113	39.2%	−2.1%	.736
	Climbing several flights of stairs	45.8%	31.2%	−14.6%	.000	41.9%	−3.9%	.129	42.6%	−3.2%	.138
	Climbing one flight of stairs	17.2%	8.5%	−8.7%	.000	14.4%	−2.8%	.051	15.3%	−1.9%	.556

Note: Unadjusted estimates only use Health and Retirement Study weights. R = respondent.

16 percentage points different from the full sample. Except for the onset of cancer, all differences are significant. This shows that some correction is necessary to adjust for the selectivity of Internet access. The prevalence rates suggest that respondents with Internet access have lower prevalence of chronic diseases, fewer symptoms of mental health problems, and fewer limitations in their activities of daily living than the rest of the population aged 50 and older. This is in line with the results in Table 1, where we saw that poor or fair self-assessed health (as well as difficulties with grocery shopping) greatly reduced the probability of Internet access. The adjusted estimates of the means in columns 5 and 8 are typically much closer to those of the complete sample. For almost all variables, the difference is smaller than the unadjusted difference. Still, many of them remain statistically significant. There is no clear difference in this respect between the GenMatch adjusted estimates (column 5) and the inverse propensity score adjusted estimates (column 8). For some variables, better balance is obtained with the former, for others with the latter.[3]

Table 4 is constructed in the same way as Table 3, considering variables on health behavior and asset ownership. It also has two continuous variables: the log amounts held in stocks and in checking and saving accounts. The qualitative conclusions are the same as in Table 3. Unadjusted differences between the Internet access sample and the full sample are almost always significant, and the differences are substantially reduced by the adjustment, although many of them remain statistically significant. For example, ownership of stocks is much more common in the Internet sample than in the non-Internet sample. Part of this can be explained by differences in income and demographics, explaining why the differences are reduced by the adjustments. The remaining difference is smaller but still statistically significant, suggesting that the assumption of CI is not valid for the chosen set of conditioning variables.

Combining the variables in Tables 3 and 4, before the adjustment, the average difference in estimates between respondents with Internet access and the entire sample for the indicator variables is 6.5 percentage points. After the adjustment, the average difference for the indicator variables is 3.7 percentage points for the inverse propensity scores as weights adjustment and 3.4 percentage points for the GenMatch adjustment. This is an average reduction of 2.8 to 3.1 percentage points. About 40 percent of the adjusted estimates in Tables 3 and 4 are still statistically significantly different from the full-sample estimates. The two adjustment methods perform equally well and tend to make similar adjustments.

Table 4
Differences in Health-Related Behavior and Ownership and Amounts of Financial Assets

			Unadjusted			Adjusted Using GenMatch			Adjusted Using Inverse Propensity Scores as Weights		
		Full Sample	R With Internet Access	Diff	*p*	R With Internet Access	Diff	*p*	R With Internet Access	Diff	*p*
Health behavior	R smoke ever	60.8%	60.2%	−0.5%	.007	67.4%	6.6%	.001	61.4%	0.6%	.010
	R smoke now	15.7%	14.0%	−1.7%	.212	14.5%	−1.2%	.199	13.0%	−2.7%	.059
	R ever drinks alcohol	49.0%	64.9%	16.0%	.000	63.5%	14.6%	.000	57.5%	8.5%	.000
	Vigorous physical activity 3+ wk	42.5%	51.3%	8.7%	.000	46.1%	3.6%	.020	44.4%	1.8%	.138
	Hospital stay within 2 years	26.2%	20.0%	−6.1%	.000	24.9%	−1.3%	.826	25.6%	−0.6%	.765
	Doctor visit within 2 years	96.8%	95.7%	−1.1%	.000	99.5%	2.7%	.001	95.6%	−1.2%	.014
Assets	Has checking account	88.3%	95.6%	7.3%	.000	94.3%	6.0%	.000	94.2%	5.9%	.000
	Owns stock	35.0%	50.8%	15.8%	.000	46.1%	11.0%	.000	42.0%	6.9%	.003
	Assets stock (log)	1.62	2.48	0.85	.00	2.22	0.60	.000	2.02	0.40	.000
	Assets checking (log)	3.57	4.08	0.51	.00	3.89	0.32	.000	3.82	0.25	.000

Note: Unadjusted estimates only use Health and Retirement Study weights. The last two variables are continuous; the remaining variables are indicator variables. R = respondent.

We also consider bivariate distributions of pairs of variables and test whether differences are still significant after adjustment using propensity weights. This is done by constructing a new variable with four outcomes from the two binary variables and testing whether the new variable is independent of Internet access. Looking at all possible combinations, we find that the null hypothesis is rejected in 53 percent of all cases. This confirms that the adjustment helps but does not completely resolve all selection problems.

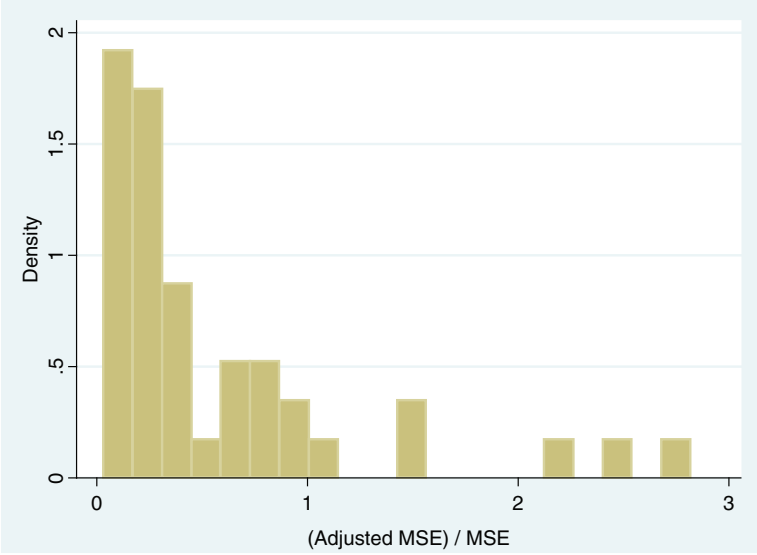
There are two instances for which the adjusted difference is reduced but remains larger than 6 percentage points: "R [Respondent] ever drinks alcohol" (15 percentage points/9 percentage points difference) and "Owns stock" (11 percentage points/7 percentage points difference). For the two continuous variables in our investigation, stocks and checking, the differences after the adjustment are greatly reduced but remain substantial and significant. In general, however, we can conclude that the corrections work for a broad range of health-related variables. This implies that once selectivity through the set of variables in Table 1 (including only self-assessed health dummies and no other health variables) is controlled for, other health variables can be studied using the Internet sample only. The adjustment does not work well for assets.

All this leads to the conclusion that even controlling for SES through income and education variables, households with Internet access more often hold stocks than households without Internet access. It implies that collecting data on asset ownership for the Internet sample and adjusting the estimates using propensity scores (based on the small set considered here and not incorporating asset ownership information) is not sufficient to analyze asset ownership in the population of interest. A similar conclusion applies to health-related behavior: The limited set of conditioning variables used to build the propensity scores is not enough to control for differences in all health-related behavior between respondents with Internet access and those without.

Discussion

While a reduction in bias is desirable, it does not necessarily imply a reduction in mean square error (MSE). The adjustment weights induce probability design effects that inflate the variances. In our case, the adjustment raises the probability design effect (Kish 1965) from 1.4 (HRS weights only) to 6.7 (combining HRS weights and inverse propensity scores for Internet access). Still, because of the large sample size, the estimated variances of the means are small and the bias dominates the MSE both

Figure 1
Histogram of (Adjusted MSE)/MSE Using Inverse Propensity Scoring as Weights



Note: MSE = mean square error.

before and after the adjustment. Figure 1 shows a histogram of the ratio (adjusted MSE)/MSE for each estimate in Tables 3 and 4. Like the bias, the MSE is generally reduced (the ratios are smaller than 1 for most estimates). Lee (2004, 2006) also finds reduced bias at the expense of increased variance. Similar to what we find, she also finds that increases in variance sometimes lead to an increase of the MSE (Lee 2004).

Web surveys have several advantages compared to more traditional surveys with in-person interviews, telephone interviews, or mail-outs. Their most obvious potential drawback is that they may not be representative of the population of interest because the subpopulation with access to the Internet may be quite specific. In this article, we investigated selectivity and how to deal with it using an unusually rich sampling design, where the Internet sample is drawn from an existing much larger probability sample that is representative of the U.S. population aged 50 and older and their spouses.

We used this to estimate propensity scores to correct for the selection effect. We investigated whether a relatively small set of variables is sufficient to correct the distribution of other variables. The idea is that if the small set is sufficient, then for new surveys, we need only a representative sample with information on the small set of variables. The other questions can be asked exclusively over the Internet. This would be very useful because of the higher cost per time unit of phone or personal interviews, because many types of questions are easier to ask over the Internet, because of the ability to exploit graphical possibilities, and because of other advantages of Internet interviewing such as shorter turnaround time and so forth.

We find that the estimated bias is almost always reduced, but significant differences in many cases remain in this large sample. For example, we still find that ownership of shares of stock or mutual funds is substantially overestimated when using respondents with Internet access only. The implication is that Web survey information on ownership of stocks is not enough to estimate the ownership rate in the population of interest, even in the presence of a representative survey of other socioeconomic variables. This conclusion differs from that of Berrens et al. (2003), who find that the correction using propensity scores based on webographic questions works well for analyzing political variables. We find that the corrections generally work well for health variables, but not for past health behavior (smoking and drinking) or, particularly, financial assets. An obvious difference between the samples looked at by Berrens et al. and our samples is that we are looking at the population aged 50 and older, among which Internet access is much less prevalent than among the population at large. One may suspect, therefore, that if Internet access among older age groups increases, propensity score reweighting or matching will become more effective.

If there is some unobserved characteristic that drives both selection (e.g., through Internet access) in a way unrelated to the propensity variables and an outcome variable of interest, then no weighting scheme will fix the problem. The use of webographic variables can be seen as an attempt to capture some of these otherwise unobserved characteristics. Unfortunately, the HRS does not contain the so-called webographic variables used to construct the weights in several recent Internet convenience samples, but it would be interesting to check their performance in the same way. Part of the challenge will be to identify which outcome variables can be adjusted for with a given set of propensity variables.

If propensity scores cannot be used to correct for selectivity in the distribution of the variables of interest, this underlines the necessity of getting broader coverage of Internet surveys or the continued search for suitable

webographic variables. Perhaps broader coverage, including older age groups, will happen automatically over the next 10 years, given the speed with which Internet access has spread in recent years. Particularly for elderly cohorts, however, alternatives may still be necessary. One obvious solution is to provide non-Internet users with access to the Internet by giving them the necessary equipment. A prominent example is the CentERpanel, collected by CentERdata in the Netherlands (http://www.uvt.nl/centerdata/en). Other examples are Knowledge Networks (http://www.knowledgenetworks.com) and the RAND American Life Panel (http://www.rand.org/labor/roybalfd/american_life.html). All three organizations provide a so-called set-top box (or Web TV) to households without Internet access that can be used to connect to the Internet, using a TV set as a monitor. (A TV set is provided as well if necessary.) Although this does not alleviate other common problems such as unit nonresponse and panel attrition, the approach provides much broader coverage and much better chances of appropriately correcting with propensity weights based on a few basic variables.

Another approach is to conduct Web surveys as part of a mixed-mode strategy with the intention to capture the part of the sample that is unable or unwilling to respond on the Web through another mode. While the administrative overhead increases, a mixed-mode strategy can be less expensive than, say, a mail-only survey (Schonlau, Asch, and Du 2003).

Notes

1. The Health and Retirement Study (HRS) documentation refers to these as sampling weights. We will refer to them as HRS weights.

2. There could be an efficiency gain in selecting a minimum set X and constructing propensity scores and weights for each separate (set of) variable(s) Y of interest (Rubin and Thomas 1996). We do not pursue this here.

3. Valliant (2004) showed that standard error may be underestimated due to the presence of multiple weighting steps.

References

Battaglia, Michael P., Donald J. Malec, Bruce D. Spencer, David C. Hoaglin, and Joseph Sedransk. 1995. "Adjusting for Noncoverage of Nontelephone Households in the National Immunization Survey." Pp. 678-83 in *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Bellemare, Charles and Sabine Kroeger. 2007. "On Representative Social Capital." *European Economic Review* 51:183-202.

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.

Birnbaum, Michael H. 2004. "Human Research and Data Collection via the Internet." *Annual Review of Psychology* 55:803-32.

Blumberg, Stephen J., Julian V. Luke, and Marcie L. Cynamon. 2004. "Has Cord-Cutting Cut Into Random-Digit-Dialed Health Surveys? The Prevalence and Impact of Wireless Substitution." Pp. 137-42 in *Eighth Conference on Health Survey Research Methods*, edited by S. B. Cohen and J. M. Lepkowski. Hyattsville, MD: National Center for Health Statistics.

Butz, William P. and Barbara B. Torrey. 2006. "Some Frontiers in Social Science." *Science* 312:1898-900.

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

Couper, Mick P. 2000. "Web Surveys. A Review of Issues and Approaches." *Public Opinion Quarterly* 64:464-94.

———, Arie Kapteyn, Matthias Schonlau, and J. Winter. 2007. "Noncoverage and Nonresponse in an Internet Survey." *Social Science Research* 36:131-48.

Czajka, John L., Sharon M. Hirabayash, Roderick J. A. Little, and Donald B. Rubin. 1992. "Projecting From Advanced Data Using Propensity Modeling: An Application to Income and Tax Statistics." *Journal of Business and Economic Statistics* 10:117-32.

Danielsson, Stig. 2004. "The Propensity Score and Estimation in Nonrandom Surveys: An Overview." Modern Statistical Survey Methods Project Report No. 18, Department of Statistics, University of Linkoeping. Retrieved August 2004 from http://www.statistics.su.se/modernsurveys/publ/11.pdf

De Heer, Wim. 1999. "International Response Trends: Results of an International Survey." *Journal of Official Statistics* 15:129-42.

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, Berkeley University. Available from 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.

Duncan, Kristin B. and Elizabeth A. Stasny. 2001. "Using Propensity Scores to Control Coverage Bias in Telephone Surveys." *Survey Methodology* 27:121-30.

Garren, Steven T. and Ted C. Chang. 2002. "Improved Ratio Estimation in Telephone Surveys Adjusting for Noncoverage." *Survey Methodology* 28:63-76.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley.

——— and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley.

Health and Retirement Study. 2002. *Sampling Weights Revised for Tracker 2.0 and Beyond*. Ann Arbor, MI: University of Michigan, Survey Research Center. Available from http://hrsonline.isr.umich.edu/meta/tracker/desc/wghtdoc.pdf

Huggins, Vicki J. and Karol Krotki. 2001. "Implementation of Nationally Representative Web-Based Surveys." In *Proceedings of the Annual Meeting of the American Statistical Association*. Available from http://www.amstat.org/Sections/Srms/Proceedings/y2001/Proceed/00299.pdf

Iannacchione, Vincent G. 2003. "Sequential Weight Adjustments for Location and Cooperation Propensity for the 1995 National Survey of Family Growth." *Journal of Official Statistics* 19:31-43.

Isaksson, Annika and Gosta Forsman. 2003. "A Comparison Between Using the Web and Using the Telephone to Survey Political Opinions." Pp. 100-106 in *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association. Available from http://www.amstat.org/sections/SRMS/Proceedings

Juster, F. Thomas and Richard Suzman. 1995. "An Overview of the Health and Retirement Study." *Journal of Human Resources* [Special Issue on the Health and Retirement Study: Data Quality and Early Results] 30:S7-56.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley.

Lee, Sunghee. 2004. "Statistical Estimation Methods in Volunteer Panel Web Surveys." Ph.D. dissertation, Joint Program of Survey Methodology, University of Maryland.

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

Link, Michael W. and Robert W. Oldendick. 1999. "Call Screening: Is It Really a Problem for Survey Research?" *Public Opinion Quarterly* 63:577-89.

Little, Roderick A. and Donald B. Rubin. 2002. *Statistical Analysis With Missing Data*. New York: John Wiley.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9:403-25.

Oldendick, Robert W. and Michael W. Link. 1994. "The Answering Machine Generation: Who Are They and What Problem Do They Pose for Survey Research?" *Public Opinion Quarterly* 58:264-73.

Rosenbaum, Paul R. 1987. "Model-Based Direct Adjustment." *Journal of the American Statistical Association* 82:387-94.

———. 2002. *Observational Studies*. 2d ed. New York: Springer-Verlag.

——— and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.

Rubin, Donald B. and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics* 52:249-64.

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.

———, Ronald Fricker, and Marc Elliott. 2002. *Conducting Research Surveys via Email and the Web*. Santa Monica, CA: RAND.

———, Kinga Zapert, Lisa Payne Simon, Katherine Sanstad, Sue Marcus, John Adams, Mark Spranca, Hongjun Kan, Rachel Turner, and Sandra Berry. 2004. "A Comparison Between a Propensity Weighted Web Survey and an Identical RDD Survey." *Social Science Computer Review* 22:128-38.

Schwarz, Norbert and Seymour Sudman, eds. 1992. *Context Effects in Social and Psychological Research*. New York: Springer-Verlag.

Smith, Philip J., J. N. K. Rao, Michael P. Battaglia, Dani Daniels, and Trena Ezzati-Rice. 2001. "Compensating for Provider Nonresponse Using Response Propensities to Form Adjustment Cells: The National Immunization Survey." *Vital and Health Statistics* 2 (133): 1-17.

Sobel, Michael E. 1995. "Causal Inference in the Social and Behavioral Sciences." Pp. 1-38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by Gerhard C. Arminger, Clifford C. Clogg, and Michael E. Sobel. New York: Plenum.

Taylor, Humphrey. 2000. "Does Internet Research 'Work'? Comparing On-Line Survey Results With Telephone Surveys." *Journal of Marketing Research* 42:51-64.

Valliant, Richard. 2004. "The Effect of Multiple Weighting Steps on Variance Estimation." *Journal of Official Statistics* 20:1-18.

Varedian, Maria and Gosta Forsman. 2003. "Comparing Propensity Score Weighting With Other Weighting Methods: A Case Study on Web Data." Presented at the American Association for Public Opinion Research Conference, St. Petersburg Beach, FL.

Winship, Christopher and Stephen L. Morgan. 1999. "The Estimation of Causal Effects From Observational Data." *Annual Review of Sociology* 25:659-707.

_____ and Michael Sobel. 2006. "Causal Inference in Sociological Studies." Retrieved December 2006 from http://www.wjh.harvard.edu/~winship/cfa_papers/causalinference.pdf

Yoshimura, Osamu. 2004. Pp. 4660-65 in "Adjusting Responses in a Non-probability Web Panel Survey by the Propensity Score Weighting." *Proceedings of the Section on Survey Statistics*, *American Statistical Association*. Available from http://www.amstat.org/sections/SRMS/Proceedings

Matthias Schonlau is head of the statistical consulting service at RAND Corporation and an adjunct assistant professor in the Department of Psychiatry, Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine. His recent work has focused on selection bias and nonresponse in Web surveys, propensity scoring, graphics, and applications in public policy. Publication outlets include *Statistical Science*, *Survey Research Methods*, the *Journal of the American Medical Association*, and the *New England Journal of Medicine*. He is the lead author of *Conducting Research Surveys via E-Mail and the Web* (RAND, 2002).

Arthur van Soest is a professor of econometrics at Tilburg University and an adjunct senior economist at RAND Corporation. He is also one of the scientific directors of Netspar, the Network of Studies on Pensions, Aging and Retirement. His research focuses on microeconometrics, particularly applied to labor economics, economics of aging, savings and consumption, health, retirement and work disability, and subjective expectations. He is involved with several projects on innovative data collection, including Internet interviewing. He has recently published in the *Journal of Econometrics*, *Journal of Health Economics*, *Journal of the American Statistical Association*, and *Econometrica*.

Arie Kapteyn is a senior economist at RAND Corporation and director of the Labor and Population program. He is a fellow of the Econometric Society, past president of the European Society for Population Economics, and a corresponding member of the Netherlands Royal Academy of Arts and Sciences. His research covers microeconomics, public finance, and econometrics. Much of his recent applied work is in the field of aging, with papers on topics related to retirement, consumption and savings, pensions and Social Security, disability, and economic well-being of the elderly. He has a strong interest in new modes of data collection, including Internet interviewing.

Mick Couper is a research professor in the Survey Research Center at the University of Michigan and the Joint Program in Survey Methodology. He is a fellow of the American

Statistical Association. His current research focuses on the design and execution of Internet surveys and online health interventions. He also does research on the application of technology to the survey data collection process, mixed-mode survey design, and survey nonresponse. He is the author of *Designing Effective Web Surveys* (Cambridge University Press, 2008).