Journal of Official Statistics, Vol. 22, No. 2, 2006, pp. 329–349

Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys

Sunghee Lee[1]

Propensity score adjustment (PSA) has been suggested as an approach to adjustment for volunteer panel web survey data. PSA attempts to decrease, if not remove, the biases arising from noncoverage, nonprobability sampling, and nonresponse in volunteer panel web surveys. Although PSA is an appealing method, its application in web survey practice is not well documented, and its effectiveness is not well understood. This study attempts to provide an overview of the PSA application by demystifying its performance for web surveys. Findings are three-fold: (a) PSA decreases bias but increases variance, (b) it is critical to include covariates that are highly related to the study outcomes, and (c) the role of nondemographic variables does not seem critical to improving PSA.

Key words: Web survey; propensity score adjustment.

1. Introduction

The field of survey methodology is experiencing a challenging expansion – web surveys. Conceptual ideal survey methods may function properly when the data collection channels reflect what exists in society. This is because survey methods inevitably manifest the society, as Dillman (2002, p. 6) indicated: "our survey methods are more a dependent variable of society than an independent variable." The development of the web has changed the structure of our daily communication channels. Exchanging electronic mails (e-mails) and sending instant messages over the web are now regarded as ordinary activities in most developed countries. Taking advantage of e-mail and web-based technology when developing surveys brings unique challenges to the field of survey methodology.

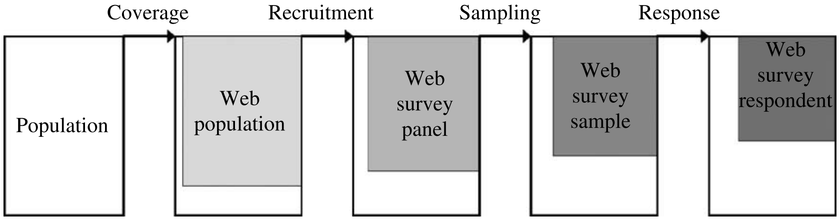
Echoing the recent changes in the communication structure and the dependence of survey modes on the existing communication structure, web surveys have been adopted rapidly as a survey medium (Taylor and Terhanian 2003). The flexibility of the web as a survey medium has resulted in various types of web surveys (see Schonlau et al. 2002; Manfreda 2001; Couper 2000 for a review). It should be noted that the scope of web surveys discussed in this study is restricted to volunteer panel web surveys, the most

[1] UCLA Center for Health Policy Research, 110960 Wilshire Boulevard, Suite 1550, Los Angeles, CA 90024, U.S.A. Email: slee9@ucla.edu

Acknowledgments: The author wishes to acknowledge the generous help from Dr. Matthias Schonlau. The author also thanks Drs. Richard Valliant, J. Michael Brick, Michael P. Couper, Robert Mislevy, and Trivellore Raghunathan for their supervision and Kristen Olson, Ting Yan, Ivan Carrillo, and Emilia Peytcheva for their helpful comments.

widely used web survey method. Some private survey firms, such as Harris Interactive (HI), exclusively use volunteer panels for their survey operations.

Figure 1 depicts the overall selection process for volunteer panel web surveys. In these surveys, not everyone in the general population is covered; only those who have web access are eligible to join the panel. The panel recruitment is done via some type of advertisement, such as banner ads, pop-up ads, or e-mails. Because participation is completely voluntary, this particular survey type is classified as a "volunteer panel" method. The self-selected participants provide basic information about themselves, but neither this information nor the identification of the participants is verified. Once the panel frame is built, survey researchers draw samples from the frame of panel members whose background characteristics match those of the target populations, and actual surveys are conducted among these people. Only a subset of the selected sample completes the survey tasks, and the typical response rates reported for these web surveys are around 20 to 25 percent (Terhanian 2000).



Source: Lee 2004

Fig. 1. Volunteer panel web survey protocol

2. Impediments in Volunteer Panel Web Surveys and Their Remedy

However, the remarkable popularity of web surveys is not indicative of their scientific quality in relation to total survey error (Groves 1989), as the selection mechanisms from one stage to another, shown in Figure 1, are unknown and uncontrolled. The greatest threat to volunteer panel web surveys is uncertain and incomplete coverage of the general population. Consequently, it is impossible to construct scientific sampling frames unless the population of interest is the volunteer panel. Drawing samples with a known probability becomes impractical. The fact that only a selected proportion of the general population has web access makes its representativeness of the general population questionable. Moreover, the people whose e-mail addresses are used for web surveys and/or the final set of respondents who comply with the web survey request may not represent the target population. Therefore, estimates from this web survey targeting the general population may suffer from a combination of noncoverage, nonprobability sampling, and nonresponse. One simple term summarizing these problems is selection bias, as the selection mechanism is not guaranteed to work in a randomized fashion.

Based on this argument, researchers at HI suggest applying a widely practiced technique for selection bias, propensity score adjustment (PSA) (e.g., Taylor 2000; Terhanian and Bremer 2000), as a way to produce post-survey adjustment weights that ideally surmount selection bias in web surveys. Traditional post-survey adjustments, such as post-stratification, have been shown to be limited in correcting for biases in web surveys (Lee 2003; Vehovar and Manfreda 1999). Therefore, an alternative technique, such as propensity score adjustment (PSA), is needed to facilitate the improved use of web survey data.

A few studies have examined the application of PSA for volunteer panel web surveys (e.g., Schonlau et al. 2004; Danielsson 2002; Varedian and Forsman 2002; Taylor et al. 2001; Taylor 2000; Terhanian et al. 2000), but more in-depth evaluation is needed. First, the resemblance between web surveys and the situations from which PSA originated needs to be scrutinized. Second, the mathematics behind the PSA for web survey data needs to be clearly presented. Third, adjusted web estimates in those studies have often been compared to estimates from other surveys, typically telephone surveys conducted parallel to the web surveys. Because both estimates are subject to all survey errors in Groves (1989), the source of any observed differences cannot be determined. The errors in web survey estimates may be assessed more effectively by using a population with known characteristics. Fourth, existing studies have focused only on bias properties of the estimates. The other component of survey errors, variance, has not been examined, although PSA is likely to increase variability. In general, weights add an extra component to the variability of the estimates and, thus, decrease the precision. Therefore, it is important to examine both aspects of errors in evaluating the performance of PSA. The last issue is that some of the existing studies favored web surveys by comparing the web polling estimates and the election outcomes. Drawing conclusions about the general quality of web surveys from these studies may be flawed, for example if web survey respondents are more likely to vote than telephone survey respondents. If true, this fact alone may make web surveys favorable for predicting election results because respondents' likelihood of voting may determine the election outcomes, not because web surveys are a more valid survey method.

This study attempts to investigate the viability of adopting PSA to overcome the limitations of past research. To bridge the gap between the original setting for PSA and the current web survey practice, Section 3 contains an overview of PSA. Section 4 contains a detailed description of the process of applying PSA to web survey data. An example of applying PSA is described in Section 5, followed by the illustration and discussion of the results in Section 6. The article concludes with procedural remarks about PSA.

3. Propensity Score Adjustment for Observational Studies

Propensity score adjustment was introduced as a post-hoc approach to alleviate the confounding effects of the selection mechanism in observational studies by achieving a balance of covariates between comparison groups (see Rosenbaum and Rubin 1983, 1984, and D'Agostino 1998 for a review). Group comparisons are a popular method of presenting scientific research outcomes. For example, a newspaper article may claim that a survey has found that those who consume a glass of wine daily have a lower risk of heart attack than those who do not. The result seems reasonable prima facie. A closer examination may reveal that this claim relies on a critical assumption – the level of wine consumed is the sole factor influencing the differential risk of heart attack between the two groups. This assumption may not be valid because wine drinkers and nondrinkers may differ on other characteristics that are also risk factors, such as age, gender, race, education, or health status. This kind of a direct comparison can be justified only when the researchers assign study subjects randomly to the comparison groups and calculate the treatment effects from those groups. The comparison groups may differ not only by wine consumption but also by other characteristics.

The desired effect of wine consumption, τ, which is the theoretical difference in heart attack risk between the two populations of the wine consumer group, τ_1, and the nonconsumer group, τ_0, may not be readily derived from their survey estimates, r_1 and r_0. This is because the expected group difference in the observational study is biased unless the group assignment (wine consumption), the study variable (heart attack risk) and the auxiliary variables (e.g., age, gender, race, education, and health status) are independent. Specifically, $E(r_1) - E(r_0) = \tau_1 - \tau_0 + (\bar{u}_1 - \bar{u}_0)$, where $\bar{u}_1$ and $\bar{u}_0$ are some function of the auxiliary variables, $\mathbf{x}$, for respective groups. The survey estimates of the heart attack risk differences may contain an artifact arising from the unbalanced covariates.

Randomization of group assignment, although desirable, is impractical, unethical, or impossible at times. For example, randomization of the level of wine consumption may be possible in a lab experiment, but the generalizability of its findings may be limited. Such an experiment might be unethical to construct in view of the fact that the treatment factor may directly affect health. Observational studies become the only approach in this case, because it is not possible to force one group of people to drink a glass of wine every day and the others not. The control is out of the researcher's hands, and those nonrandomized conditions may confound the study outcome as described above. Thus the researcher is restricted to what is available.

PSA can be applied to reduce selection bias. The following summary of the general approach of PSA is based on Rosenbaum and Rubin (1983). A propensity score is defined as the conditional probability of receiving treatment given the vector of the individual's covariates and is calculated for each individual. It is often estimated in a logistic regression model as

$$\ln\left[\frac{e(\mathbf{x})}{1 - e(\mathbf{x})}\right] = \alpha + \boldsymbol{\beta}^T f(\mathbf{x}) \qquad (1)$$

where $e(\mathbf{x}) = \Pr(z = 1|\mathbf{x})$ is the propensity score of receiving the treatment ($z = 1$) given a set of covariates, $\mathbf{x}$, and $f(\mathbf{x})$ is some function of the covariates. For a given propensity score, covariates and study results become independent of the assigned treatment, i.e., $\mathbf{x} \perp z|e(\mathbf{x})$ and $(r_1, r_0) \perp z|e(\mathbf{x})$ (see Theorems 1 and 3 in Rosenbaum and Rubin 1983). It is assumed that every unit in the population has a nonzero propensity score, $0 < e(\mathbf{x}) < 1$. These two conditions comprise the "strong ignorability" in Rosenbaum and Rubin (1983) that PSA assumes.

In practice, $e(\mathbf{x})$ is unknown, because we have only observed covariates, $\mathbf{x}_{obs}$, not the full range of covariates, $\mathbf{x}' \equiv ((\mathbf{x}_{obs})', (\mathbf{x}_{unobs})')$, including both observed and unobserved covariates. Instead, $\hat{e}(\mathbf{x}_{obs})$ is used, estimated by applying expression (1) to observed data. As long as $\mathbf{x}_{obs}$ represents $\mathbf{x}$, meaning that $\hat{e}$ contains all potential confounders, the adjustment based on the propensity score leads to unbiased estimates of the treatment effect, such that

$$\begin{aligned} &E_{e(\mathbf{x})}[E\{r_1|e(\mathbf{x}), z=1\} - E\{r_0|e(\mathbf{x}), z=0\}] = E_{e(\mathbf{x})}[E\{r_1|e(\mathbf{x})\} - E\{r_0|e(\mathbf{x})\}] \\ &\quad = E(r_1 - r_0) = \tau_1 - \tau_0 = \tau \end{aligned} \tag{2}$$

A large literature applying PSA can be found in biostatistics for comparisons of nonrandomized treatment and control groups (e.g., Lieberman et al. 1996; Frigoletto et al. 1995; Stone et al. 1995; Lavori 1992; Cook and Goldman 1989; Lavori and Keller 1988). These studies proved the usefulness of PSA in balancing covariates. However, the degree of bias reduction in the treatment effects brought by PSA was only speculated, as the true treatment effects were unobtainable.

Other methods for reducing selection bias in observational studies can be found in econometrics, such as Heckman's parametric selection bias model (Heckman 1979) and the instrumental variable approach (Angrist, Imbens, and Rubin 1996). However, these econometric selection methods require less realistic distributional assumptions and are very sensitive to model specification details (Crown 2001; Obenchain and Melfi 1997). These limitations lower the applicability of econometric selection methods.

4. Propensity Score Adjustment for Volunteer Panel Web Surveys

The idea of adopting PSA for web survey data was first introduced by HI (e.g., Taylor 2000; Terhanian and Bremer 2000), regarding uncertainties about selection bias arising in their survey protocol. PSA for volunteer panel web surveys starts with the assumption that there are reference survey data (Terhanian and Bremer 2000). The reference survey is conducted parallel to the web survey in terms of target population and time. It is supposed to have more desirable properties, such as the power of probability sampling through a traditional survey mode such as in-person or telephone interviews, and higher response rates. The reference survey serves as a benchmark for the web survey. The benchmarking is carried out via PSA by balancing the covariate distributions of the web sample to match those of the reference sample.

Suppose that there are two samples: (a) a volunteer panel web survey sample (s^W) with n^W units each with a base weight of d_j^W, where $j = 1, \ldots, n^W$, and (b) a reference survey sample (s^R) with n^R units each with a base weight of d_k^R, where $k = 1, \ldots, n^R$. Note that d_j^W values may not be inverses of selection probabilities because probability sampling is not used. First, the two samples are combined into one, $s = (s^W \cup s^R)$ with $n = n^W + n^R$ units. We calculate propensity scores from s. The propensity score of the ith unit is the likelihood of the unit participating in the volunteer panel web survey ($g = 1$) rather than the reference survey ($g = 0$), where $i = 1, \ldots, n$, given auxiliary variables. Therefore, g in PSA applied to web survey adjustment may be labeled as sample origin instead of treatment assignment. Propensity scores are defined as $e(\mathbf{x}_i) = \Pr(i \in s^W|\mathbf{x}_i, i = 1, \ldots, n)$ and estimated in a logistic regression as in Equation (1) using covariates observed commonly in the web and the reference survey, $\mathbf{x}_{obs}$. Critical assumptions in doing this are (a) that given a set of covariate values, a person must have some nonzero probability of being in the web survey and (b) that probability must be estimable from the combined sample, s.

Based on the predicted propensity score, $\hat{e}(\mathbf{x}_{obs})$, the distribution of the web sample units is rearranged so that s^W resembles s^R. Mechanically, this is first done by sorting s by $\hat{e}(\mathbf{x}_{obs})$ and partitioning s into C subclasses, where each subclass has about the same number of units. Alternatively, one might use only s^R in this subclassification. However, the aim of this study is to evaluate the current practice, which uses s. Based on Cochran (1968), the conventional choice is to use five subclasses based on quintile points. Ideally, all units in a given subclass will have about the same propensity score or, at least, the range of scores in each class is fairly narrow. This is so that Equation (2) will apply approximately. In the cth subclass, denoted as s_c, there are $n_c = n_c^W + n_c^R$ units, where n_c^W is the number of units from s^W, and n_c^R from s^R. The total number of units s remains the same because

$$\sum_{c=1}^{C}\left(n_c^W + n_c^R\right) = \sum_{c=1}^{C} n_c = n$$

Second, we compute the following adjustment factor:

$$f_c = \frac{\sum_{k\in(s_c^R)} d_k^R \Big/ \sum_{k\in(s^R)} d_k^R}{\sum_{j\in(s_c^W)} d_j^W \Big/ \sum_{j\in(s^W)} d_j^W} \tag{3}$$

where s_c^R and s_c^W are the sets of units in the reference sample and web sample of the cth subclass. If the base weights in Equation (3) are the inverses of selection probabilities, it can be expanded to:

$$f_c = \frac{\sum_{k\in(s_c^R)} d_k^R \Big/ \sum_{k\in(s^R)} d_k^R}{\sum_{j\in(s_c^W)} d_j^W \Big/ \sum_{j\in(s^W)} d_j^W} \equiv \frac{\hat{N}_c^R/\hat{N}^R}{\hat{N}_c^W/\hat{N}^W}$$

The adjusted weight for unit j in class c of the web sample becomes

$$d_j^{W.PSA} = f_c d_j^W = \frac{\hat{N}_c^R \Big/ \hat{N}^R}{\hat{N}_c^W \Big/ \hat{N}^W} d_j^W \tag{4}$$

When the base weights are equal for all units or are not available, one may use an alternative adjustment factor as follows:

$$f_c = \frac{n_c^R \Big/ n^R}{n_c^W \Big/ n^W} \tag{5}$$

The adjustment using Equation (5) does not allow population totals to be estimated because the weights are not appropriately scaled, unless the population sizes for both the reference survey and the web survey are known.

The weights using Equation (4) may make the distribution of the web survey sample equal to that of the reference survey sample in terms of propensity scores. For example, using the adjusted weights, the estimated number of units in class c from the web sample is

$$\begin{aligned}\hat{N}_c^{W.PSA} &= \sum_{j \in (s_c^W)} d_j^{W.PSA} \\ &= \hat{N}^W \frac{\hat{N}_c^R}{\hat{N}^R}\end{aligned}$$

In words, the estimated number of units from the web survey, $\hat{N}^W$, is distributed among the classes according to the distribution from the reference survey, $\hat{N}_c^R / \hat{N}^R$. The estimator for the mean of a study variable, y, from s^W becomes

$$\hat{\bar{y}}^{W.PSA} = \frac{\sum_c \sum_{j \in (s_c^W)} d_j^{W.PSA} y_j}{\sum_c \sum_{j \in (s_c^W)} d_j^{W.PSA}} \tag{6}$$

Note that the reference sample is not used for estimating $\hat{\bar{y}}^{W.PSA}$. It is needed only in the adjustment process and, thus, is required to have only the covariate data, not necessarily the variables of interest. The same reference sample can be used for the adjustment of more than one web survey as long as its target population coincides with that of the web samples and the temporal circumstances are equivalent. The size of the reference sample is often smaller than that of the web sample. If the reference sample needed to be larger than the web sample, and a reference sample had to be conducted for every web survey, the cost-effectiveness of web surveys would be lost.

PSA for selection bias is not generally necessary in survey data analysis because most scientific surveys draw randomized samples. In theory, survey estimates are expected to be design unbiased or consistent with the distribution of characteristics within a population. On the other hand, PSA is not novel to survey statistics, especially to post-survey adjustment. PSA has been used to derive adjustment weights for reducing biases in survey estimates arising from coverage problems (Duncan and Stasny 2001), late response (Czajka et al. 1992), and nonresponse (Smith et al. 2000; Vartivarian and Little 2003).

The set of covariates typically includes similar kinds of demographic variables to those used in traditional post-stratification adjustment. HI includes both demographic and nondemographic variables in the propensity models (Terhanian et al. 2000; Taylor et al. 2001). The importance of covariates in PSA should be understood in relation to the substantive study variable, y, and the sample origin variable, g (Rosenbaum and Rubin 1984). Rubin and Thomas (1996) suggest including all covariates, even if some are not statistically significant in predicting the outcome variables, unless they are unrelated to the treatment outcomes or inappropriate for the model. However, in practice, covariates are usually selected on the basis of some statistical procedures, such as stepwise selection (e.g., Rosenbaum and Rubin 1983, 1984). The importance of including nondemographic variables in PSA for web surveys is unclear due to two facts: (a) inclusion of more variables automatically increases the predictive power of the model and

(b) nondemographic (e.g., attitudinal) covariates can often be explained by demographic variables to certain degree.

In her simulation study, Drake (1993) showed that the impact of misspecifying propensity score models, such as mistakenly adding a quadratic term or dropping a covariate, is not very serious. In fact, the misspecification of the propensity score model led to only a small bias compared to the misspecification of the response model that was used to simulate the response distribution.

5. Methods

5.1. Pseudo-population Data Preparation

The aim of this study was to assess the effectiveness of PSA for the reduction of selection biases in volunteer panel web surveys. This was facilitated in multiple realizations of samples, which made simulation using pseudo-populations a logical approach.

The creation of the full pseudo-population was carried out in connection with the 2002 General Social Survey (GSS). The GSS is an ongoing biennial survey conducted by the National Opinion Research Center with core funding from the National Science Foundation. The survey measures contemporary American society by targeting noninstitutionalized adults 18 years of age and older. A representative national sample was drawn using multistage area probability sampling. One special feature of the 2002 GSS is that it collected information about whether people use the web, including e-mail. The reported response rate for the 2002 GSS is 70%, which resulted in a data set (U) containing 2,746 cases with complete information on four stratifying variables – age, gender, education, and race – and the web usage variable as in Table 1.

We can examine the age, gender, education, and race distributions of the 2002 GSS full sample and web users, and the HI web survey respondents (see Table 1 Parts A, B, and C). There is a noteworthy gap not only between the GSS sample and the two web samples but, surprisingly, also between the two web samples. The GSS full sample includes fewer young people and fewer with higher education than the two web samples. The most remarkable disparity between the HI sample and the two GSS samples is in the educational attainment level. Although less than half of the GSS full sample and GSS web users have some college or higher education, this group provides 90 percent of the HI respondent data. Also, the HI sample includes more minorities, especially educated minorities, than the GSS samples. If a sample distributed like the HI respondents is to provide unbiased estimates for the general population or even the population with web access, some major weighting adjustment will be required.

PSA is feasible when all cases in the merged data have information on the covariates included in the propensity score models. Otherwise, propensity scores for the units where some of the covariates are missing cannot be computed, which hinders the adjustment. To compensate for this situation, missing values on the 14 covariates included in the propensity score models in Table 2 were imputed within the cell defined in Table 1 using the hot-deck method. Because a larger population will facilitate testing of methods by simulation, the full pseudo-population (P^F) was created by bootstrapping U with simple random sampling with replacement for a size of 20,000. Because the GSS was conducted

Table 1. Distribution of Age, Gender, Education, and Race of GSS Full Sample, GSS Web User, and Harris Interactive Survey Respondents

		High school or less		Some college or above		Total by age %	Sum %
		White %	Nonwhite %	White %	Nonwhite %		
A. GSS Full sample ($n = 2,746$)[a]							
≤ 40 yrs	Female	9.76	6.61	5.51	1.79		
	Male	9.65	4.18	4.41	1.37	43.28	
41 yrs +	Female	16.75	4.75	8.39	1.44		
	Male	13.14	3.15	7.75	1.37	56.74	
Total by education and race		49.30	18.69	26.06	5.97		
Total by education			67.99		32.03		100.00
Sum						100.00	
B. GSS web users ($n = 1,692$)[b]							
≤ 40 yrs	Female	11.68	6.08	7.97	2.62		
	Male	10.52	3.22	6.69	2.01	50.79	
41 yrs +	Female	11.50	2.31	11.01	1.64		
	Male	9.49	1.46	10.16	1.64	49.21	
Total by education and race		43.19	13.07	35.83	7.91		
Total by education			56.26		43.74		100.00
Sum						100.00	
C. Harris Interactive Respondents ($n = 8,195$)							
≤ 40 yrs	Female	2.03	1.64	13.28	13.37		
	Male	0.85	0.61	7.58	9.09	48.45	
41 yrs +	Female	2.45	0.48	15.58	4.58		
	Male	1.70	0.24	20.82	5.71	51.56	
Total by education and race		7.03	2.97	57.26	32.75		
Total by education			10.00		90.01		100.00
Sum						100.00	

[a] This sample size reflects the exclusion of 19 cases where some of the four covariates are missing.

[b] This is the subset of web users from the original 2002 GSS sample.

in the face-to-face mode, it becomes possible to draw face-to-face reference samples with known probabilities from P^F.

As discussed earlier, U contains information about e-mail and web usage. Based on this information, people who are classified as web users in P^F were retained for the pseudo-web population (P^W) for a size of 12,306. The proportion of the web users in P^F is the same as that in U at 61%. This pseudo-web population allows us to draw different types of web samples, especially ones resembling HI web survey respondents, where web usage is a prerequisite for the panel members in those surveys.

Table 2. P-values of the Auxiliary Variables in Logit Models Predicting y_{blks} (Warm Feelings toward Blacks) and y_{vote} (Voting Participation in 2000 Presidential Election)[a]

Covariate	Description	Type	*p*-value	
			y_{blks}	y_{vote}
Demographic				
Age	Age in years	Continuous	$<.0001$	$<.0001$
Educ	Education in years	Continuous	$<.0001$	$<.0001$
Newsize	Size of the residential area	Continuous	.2006	.1804
Hhldsize	Household size	Continuous	.8318	.3496
Income	Family income	Continuous	.4548	.0002
Race	Race	4 categories	$<.0001$	.0002
Gender	Gender	2 categories	$<.0001$	.1568
Married	Marital status	2 categories	.0616	.0280
Region	Region of the residential area	4 categories	.0391	.2017
Nondemographic				
Class	Self-rated social class	Continuous	.1435	$<.0001$
Work	Employment status	2 categories	.6502	.1680
Party	Political party affiliation	3 categories	.2174	$<.0001$
Religion	Having a religion	2 categories	.1197	.8480
Ethnofit	Opinion toward ethnic minorities	Continuous	$<.0001$	–

[a] These analyses were done using the original GSS sample ($n = 2,746$).

5.2. Sampling and Adjustment in Simulation

Using the two pseudo-populations, a reference sample and two types of web samples were drawn in each simulation. The reference survey sample (s^R) was drawn from P^F by simple random sampling for the size of $n^R = 200$. Because the 2002 GSS was conducted in the face-to-face mode, these reference samples will serve as face-to-face reference samples with known probabilities of selection. Note that this study also tested another set of simulations with a larger reference sample size ($n^R = 400$) holding all other conditions equal. This produced identical results.

Two types of web samples were drawn from P^W by Poisson sampling with selection probabilities equal to cell proportions in Table 1, Part B and Table 1, Part C. For example, for the first web sample, white females with a high school education or less and who were 40 years old or less were selected with a probability of 0.1168. Thus, the two samples were allocated according to the covariate distributions from Table 1, Part B and Table 1, Part C, where each cell serves as a stratum. The first web sample, $s^{W.ST}$, was assumed to resemble the pseudo-web population (Table 1, Part B). The second web sample, $s^{W.HI}$, mimicking the respondents in a HI volunteer panel web survey was drawn on the basis of subclass proportions from a real HI web survey dataset in Table 1, Part C (obtained via a personal communication with Matthias Schonlau, see Schonlau et al. 2004). Both web samples were drawn for the desired size of $n^{W.ST} = n^{W.HI} = 800$ (actual web sample sizes varied around 800, as Poisson sampling was used). This procedure of selecting the three samples (s^R, $s^{W.ST}$ and $s^{W.HI}$) was repeated 2,000 times.

This study examined two variables: (a) y_{blks}, the proportion of people indicating warm feelings toward blacks, and (b) y_{vote}, the proportion of people who voted in the 2000

Presidential Election. The estimates of y_{blks} and y_{vote} from the simulated web samples are corrected by applying PSA in Equation (6).

There are 14 covariates used for adjusting y_{blks} and 13 for y_{vote}, where nine of each set of all covariates are demographic and the remainder are nondemographic characteristics. The demographic/nondemographic nature of a given covariate is tentatively determined on the basis of whether the variable is typically used in survey post-stratification or not. As shown in Table 2, the significance of these covariates in predicting y_{blks} and y_{vote} differs greatly. Some of the variables are continuous, whereas others are categorical with different numbers of categories.

On the basis of the characteristics of the covariates (demographic or nondemographic) and a cut-point of $p = .05$ (significant or nonsignificant), different propensity score models were developed focusing on the relationship between the substantive study variables and the covariates. The first model, which served as the base model, *D1*, included all demographic variables as main effects in a logit model such that

$$\begin{aligned} DI : \ln\left(\frac{\Pr(g=1)}{1-\Pr(g=1)}\right) \\ &= \alpha + \beta_1 age + \beta_2 educ + \beta_3 newsize + \beta_4 hhldsize + \beta_5 income \\ &\quad + \beta_6 race + \beta_7 gender + \beta_8 married + \beta_9 region \end{aligned}$$

where g is the sample origin. The subsequent models used logit models with covariates' main effects only, as shown in Table 3. This allowed us to detect the importance of including significant and/or nondemographic covariates in the propensity score model. The respective effectiveness of different models is compared in the following section.

5.3. Assessment of Propensity Score Adjustment Performance

The performance of PSA is evaluated with respect to the following four criteria: bias, reduction of bias, standard error, and increase of standard error. It is important to understand the trade-off between bias reduction and variance increase because the variability introduced by the weights may increase the variance associated with the estimates, whereas applying adjustment in the estimation may reduce biases in the estimates.

5.3.1. Bias and Percent Bias Reduction

The "bias" measure of the web survey estimates compared to the reference survey estimates takes the following form:

$$\begin{aligned} bias(\bar{y}^W) &= \left[\sum_{m=1}^{M} y_m^W - \sum_{m=1}^{M} y_m^R\right] \Big/ M \\ &= \bar{y}^W - \bar{y}^R \end{aligned}$$

where y_m^R and y_m^W are the reference and web estimates from the mth simulation.

Table 3. *Composition of Propensity Score Models*[a]

Covariate	Propensity score models					
Demographic (*D*)	*D1*		*D2*		*D3*	
	All covariates		Significant covariates		Nonsignificant covariates	
	y_{blks}	y_{vote}	y_{blks}	y_{vote}	y_{blks}	y_{vote}
Age	√	√	√	√		
Educ	√	√	√	√		
Newsize	√	√			√	√
Hhldsize	√	√			√	√
Income	√	√		√	√	
Race	√	√	√	√		
Gender	√	√	√			√
Married	√	√		√	√	
Region	√	√	√			√
Nondemographic (*N*)	*N1*		*N2*		*N3*	
	All covariates		Significant covariates		Nonsignificant covariates	
	y_{blks}	y_{vote}	y_{blks}	y_{vote}	y_{blks}	y_{vote}
Class	√	√		√	√	
Work	√	√			√	√
Party	√	√		√	√	
Religion	√	√			√	√
Ethnofit	√	–	√	–	√	–
Demographics and nondemographics (*A*)	*A1*		*A2*		*A3*	
	All covariates		Significant covariates		Nonsignificant covariates	
	y_{blks}	y_{vote}	y_{blks}	y_{vote}	y_{blks}	y_{vote}
	D1 + *N1*	*D1* + *N1*	*D2* + *N2*	*D2* + *N2*	*D3* + *N3*	*D3* + *N3*

[a] Included covariates are indicated by check marks.

Note: Propensity Model 4, not shown in the table, is the combination of *D1* and *N2*.

Following Expression (2.6) in Rubin (1973), the percent reduction in bias (*p.bias*) is calculated as

$$p.bias(\bar{y}^{W.PSA}) = \left[\frac{|bias(\bar{y}^{W.U})| - |bias(\bar{y}^{W.PSA})|}{|bias(\bar{y}^{W.U})|}\right] \times 100 \tag{7}$$

where $\bar{y}^{W.PSA}$ and $\bar{y}^{W.U}$ are the simulation means of the propensity score adjusted (*PSA* is substituted by model names hereafter) and the unadjusted web estimates, respectively. We expected the unadjusted estimates to have larger bias than the adjusted ones. The larger the *p.bias*, the more effective the PSA was in reducing the bias. A negative *p.bias* indicates that the adjustment increased the bias of the estimates, meaning the adjusted estimates are of lower quality in an absolute sense than the unadjusted estimates.

5.3.2. Standard Error and Percent Standard Error Increase

The variability in estimates is calculated by the standard error (*se*) of the simulation mean as

$$se(\bar{y}^W) \approx \sqrt{\sum_{m=1}^{M} \left(y_m^W - \bar{y}^W\right)^2 \Big/ M}$$

This statistic allows us to examine the magnitude of added variability on the estimates due to the PSA. In order to measure for increased variability resulting from the adjustment, the percent increase in standard error (*p.se*) is computed as follows:

$$p.se(\bar{y}^{W.PSA}) = \left[\frac{|se(\bar{y}^{W.PSA})| - |se(\bar{y}^{W.U})|}{|se(\bar{y}^{W.U})|}\right] \times 100$$

6. Results

The simulation was conducted over 2,000 times following the procedures described in Section 5. Table 4 shows respective unadjusted means of y_{blks} and y_{vote} from three samples – s^R, $s^{W.ST}$, and $s^{W.HI}$ – over all simulations, calculated as

$$\bar{y} = \sum_{m=1}^{M} y_m \Big/ M$$

where y_m is an estimate from the *m*th simulation and $m = 1, \ldots, M$.

Web estimates deviated from benchmark reference sample estimates, implying that people in the web samples were more likely to express warm feelings toward blacks and were more likely to have participated in the election than people in the reference sample. This result seems plausible when one considers the cell proportions in Table 1 used to create $s^{W.ST}$ and $s^{W.HI}$, because these are more likely to have higher proportions of minorities and people with higher education than s^R. Estimates from $s^{W.HI}$ and $s^{W.ST}$ were more biased than s^R estimates. For example, the bias detected in the estimate for y_{vote} from $s^{W.HI}$ is 16.7 percentage points.

6.1. Performance of Propensity Score Adjustment

The bias reduction was carried out with PSA as described in Section 4. First, the base model (*D1*) was applied, and adjustment weights were computed and incorporated in the

Table 4. Simulation Means of Estimates by Different Samples Before Adjustment

	s^R	$s^{W.ST}$	$s^{W.HI}$
y_{blks}: Proportion of warm feelings toward blacks (M = 2,000)	0.612	0.636	0.675
y_{vote}: Proportion of voters in 2000 election ($M = 1,971$)[a]	0.650	0.715	0.817

[a] In simulations for y_{vote}, 29 simulations were not completed due to zero cases in subclasses in s_c^R defined by propensity scores, causing an inability to derive weights using Equation (4).

estimation. Table 5 compares unadjusted and *D1* adjusted estimates to reference sample estimates. Adjusted web survey estimates appear to be closer to the reference sample values than the adjusted estimates. For example, the *D1* propensity score adjusted mean ($y.D1$) for y_{blks} was 0.623 based on $s^{W.ST}$ samples, which is closer to the reference sample mean ($y.R$: 0.615) than the unadjusted mean ($y.U$: 0.636). By incorporating adjustment weights, the web estimates are closer to the reference sample values than the unadjusted estimates.

Table 5 presents simulation estimates of y_{blks} and y_{vote} and their evaluation statistics when no adjustment and adjustment by *D1* model were applied for both $s^{W.ST}$ and $s^{W.HI}$ (see Appendix for the same information for adjustment using all propensity models in Table 3). When *D1* adjustment was applied, biases and deviations in web estimates from the reference sample estimates decreased dramatically. The greatest advantage using PSA occurred in the samples mimicking Harris Interactive respondents – the larger bias reduction occurred for $s^{W.HI}$ compared to $s^{W.ST}$ for both study variables. This echoes the statement in Cochran et al. (1954, p. 246) that "adjustment will only be seriously helpful when the sampling procedure is not random . . ." Nonetheless, the adjusted estimates have larger standard errors, showing that the reduction in bias came at the cost of increased variability. Although percentages may seem to suggest that variance increases surpassed bias reduction, comparison on the basis of absolute values reveals the benefit of the adjustment.

6.2. *Effect of Covariates in Propensity Score Models*

The role of covariates was examined exclusively using $s^{W.HI}$. First, the significance of the covariates was examined by forming different models. As shown in Table 3, three models are related only to demographic variables: *D1*, the base propensity model,

Table 5. Reference Sample, Unadjusted, and Propensity Score Adjusted Web Sample Estimates for y_{blks} and y_{vote}

	$s^{W.ST}$					$s^{W.HI}$				
	estimate	*bias*	*p.bias*	*se*	*p.se*	*estimate*	*bias*	*p.bias*	*se*	*p.se*
y_{blks}										
($M = 2{,}000$)										
y.R	0.612			0.034		0.612			0.034	
y.U	0.636	0.024		0.016		0.675	0.064		0.016	
y.D1	0.623	0.012	52.4%	0.022	38.1%	0.638	0.026	58.6%	0.032	100.0%
y_{vote}										
($M = 1{,}971$)										
y.R	0.650			0.034		0.650			0.034	
y.U	0.715	0.065		0.015		0.817	0.167		0.013	
y.D1	0.709	0.059	9.7%	0.022	46.7%	0.724	0.074	55.7%	0.031	138.5%

Note: *y.R*: Reference sample estimate.
y.U: Unadjusted web sample estimate.
y.D1: Web sample estimate after PSA using Model *D1*.

contains all demographic covariates, *D2* contains only significant demographic covariates, and *D3* contains only nonsignificant demographic covariates.

The unadjusted ($y.U$) and the adjusted web estimates using *D1*, *D2*, and *D3* ($y.D1$, $y.D2$, and $y.D3$) are plotted against the reference sample estimate ($y.R$) for y_{blks} in Figure 2 and for y_{vote} in Figure 3, for all simulated samples. A diagonal reference line is drawn in each panel. If the propensity score adjusted web sample estimates were always equal to the reference sample estimates, then all points would fall on this reference line. Clusters of dots approaching the reference line indicate that the disparity of web estimates is diminished. Widely dispersed clusters are evidence of increased variability.

Figures 2 and 3 convey the same messages. Among the three adjustments, *D1* and *D2* outperform *D3*. When the propensity score model was composed of only highly predictive covariates (*D2*), the level of adjustment was comparable to that of the base model that includes all variables (*D1*). The PSA based on weakly predictive covariates (*D3*) did not improve the point estimates to any degree. The figures also illustrate the increased variability of estimates when using PSA weights. Once the weights are incorporated, the scatter plots in the two center panels show higher variability. In particular, estimates from the better performing models show widely scattered distributions. In the case of the propensity model *D3* for y_{blks}, the adjustment increased variability *without* decreasing the deviation to any degree, which ultimately decreased the quality of estimates in an absolute sense.

Next, we examined the effect of including nondemographic (or attitudinal) variables in the propensity score model by comparing four different models: all demographic covariates (*D1*), all nondemographic covariates (*N1*), all covariates ($A1 = D1 + N1$), and all demographic and important nondemographic covariates (*4*). The distributions of the adjusted estimates using these models are displayed in Figure 4 along with those of the reference sample estimates ($y.R$) and the unadjusted estimates ($y.U$).

For both study variables, the reference sample estimates ($y.R$) were more widely distributed than the unadjusted web sample estimates ($y.U$). This is not surprising because the web samples were four times larger than the reference samples. However, the distributions of $y.U$ did not contain $y.R$ simulation means. For y_{vote}, the distributions of $y.U$ and $y.R$ were almost nonoverlapping. Among the four adjustment models, those including demographic variables (*D1*, *A1,* and *4*) produced less biased web estimates than one with only nondemographic variables (*N1*). The marginal effect of adding

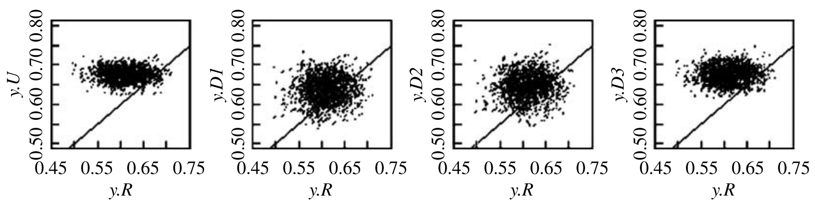


Fig. 2. Relationship between the distributions of the different web sample estimates and the reference sample estimates for y_{blks}

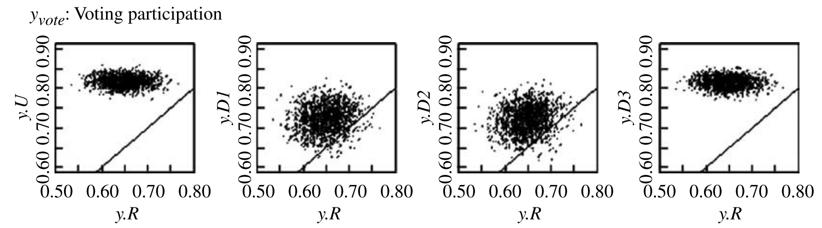


Fig. 3. Relationship between the distributions of the different web sample estimates and the reference sample estimates for y_{vote}

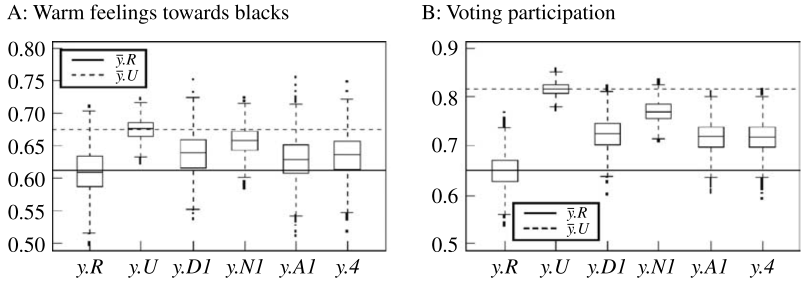


Fig. 4. Distributions of the web estimates by different propensity score adjustments

nondemographic variables can be detected by comparing the box plots for *A1* and *D1* (Figure 4). Figure 4 shows that this effect is minimal because the performance of *A1* and *D1* are comparable. Although the distributions of the adjusted estimates differ noticeably, none of the methods successfully removed the deviation. Addition of only significant nondemographic covariates (*4*) on *D1* was comparable to adding all nondemographic covariates (*A1*).

7. Discussion

This study illustrated the exclusive application of propensity score adjustment for volunteer panel web surveys. The adjustment decreased but did not eliminate the biases in the web sample estimates compared to benchmark sample estimates. Also, this bias reduction came at the cost of increased variance. The variance of the web sample estimates increased when adjustment weights were applied, and especially when the adjustments were more effective in reducing biases. The increase in variance was found primarily in propensity models containing demographic covariates, which indicates the significance of these covariates in predicting the propensity score. It is notable that the variability of effective model estimates can be as large as that of the reference sample estimates, meaning that the precision obtained from the larger sample size in web surveys is lost.

The relationship between the covariates and the study variables was found to be important in forming propensity models, because the propensity models with weakly predictive covariates did not decrease the bias but did increase the variance. It seems to be a reasonable practice to include all available covariates from a given dataset, as Rubin and Thomas (1996) suggest. However, the assertion that including nondemographic variables in the propensity models is useful was not verified – compared to demographic variables, the value of including nondemographic variables was limited in this study. This may be due to the nature of the two study variables, warm feelings toward blacks and voting behavior, which are highly correlated to demographic variables such as race and education.

Although the aim of the study, providing an overview of PSA for volunteer panel web surveys, was met, this article has not addressed certain areas. First, the study compared web survey estimates only to reference sample estimates, which were also subject to sampling and nonresponse error. A logical approach to this issue seems to be to combine the PSA weights with additional weights that project the adjusted web samples to the general population. For example, general regression estimation or some other type of calibration can be used to generate additional weights. The combination of the two weights may reduce selection bias in web surveys to a greater degree. Second, this study demonstrated the main effects of covariates in propensity models. One of the advantages of using PSA weighting over traditional weighting methods is the flexibility in model formation. Propensity model refinement (e.g., including higher order interactions among the covariates or using more covariates) may provide a clearer insight into variable selection. Third, the significance of the covariates in this study was examined only in relation to the substantive study variables, y, not to the sample origin variable, g. Covariate and model selection may be modified by incorporating both y and g, allowing an extensive examination of the role of covariates. Fourth, the value of nondemographic covariates was not confirmed in this study because the web samples were drawn on the basis of the distribution of demographic variables, and these variables were also included in the adjustment. One may consider another way of drawing web samples or conducting a series of web surveys on substantive variables whose true values are either known or obtainable. Fifth, the subclassification based on the propensity scores for voting behavior was not completed in 29 out of 2,000 simulations due to subclasses having zero cases in the reference sample. Suppose that the reference sample data were originally collected for the general population and that only a subset (e.g., veterans of the military) were to be used as the reference population for a web survey targeting a subgroup of the general population. In this situation, reference samples may contain only a small number of cases. One may consider either dividing the merged sample into a small number of subclasses (with adequate numbers in each subclass) or conducting a larger reference survey so that the reference samples for any likely web survey target populations have sufficient observations for subclassification. Finally, the study examined empirical variance in simulation. There is no clear approach for deriving a variance estimator that accounts for the complexity of multiple weights in PSA. Although not discussed in the current study, this is crucial in increasing the adaptability of PSA. These limitations remain to be explored in future research.

Appendix. Reference Sample and Unadjusted and Propensity Score Adjusted Web Sample Estimates for y_{blks} and y_{vote}

.

	$s^{W.ST}$					$s^{W.HI}$				
	estimate	***bias***	***p.bias***	***se***	***p.se***	***estimate***	***bias***	***p.bias***	***se***	***p.se***
y_{blks}										
y.R	0.612			0.034		0.612			0.034	
y.U	0.636	0.024		0.016		0.675	0.064		0.016	
y.D1	0.623	0.012	52.4%	0.022	37.5%	0.638	0.026	58.6%	0.032	100.0%
y.D2	0.622	0.010	57.1%	0.021	31.3%	0.645	0.034	47.0%	0.031	93.8%
y.D3	0.637	0.025	− 4.7%	0.018	12.5%	0.675	0.063	0.4%	0.021	31.3%
y.N1	0.620	0.008	65.7%	0.020	25.0%	0.657	0.046	28.3%	0.022	37.5%
y.N2	0.622	0.010	58.6%	0.018	12.5%	0.658	0.046	27.3%	0.017	6.3%
y.N3	0.632	0.020	17.5%	0.018	12.5%	0.672	0.061	4.8%	0.021	31.3%
y.A1	0.616	0.004	82.0%	0.023	43.8%	0.629	0.017	72.6%	0.032	100.0%
y.A2	0.617	0.005	79.4%	0.022	37.5%	0.642	0.030	52.2%	0.032	100.0%
y.A3	0.636	0.024	1.7%	0.019	18.8%	0.669	0.057	10.0%	0.021	31.3%
y.4	0.619	0.007	71.3%	0.023	43.8%	0.635	0.023	63.9%	0.032	100.0%
	$s^{W.ST}$					$s^{W.HI}$				
	estimate	***bias***	***p.bias***	***se***	***p.se***	***estimate***	***bias***	***p.bias***	***se***	***p.se***
y_{vote}										
y.R	0.650			0.034		0.650			0.034	
y.U	0.715	0.065		0.015		0.817	0.167		0.013	
y.D1	0.709	0.059	9.7%	0.022	46.7%	0.724	0.074	55.7%	0.031	138.5%
y.D2	0.711	0.062	5.4%	0.021	40.0%	0.721	0.072	57.2%	0.032	146.2%
y.D3	0.720	0.070	−7.1%	0.016	6.7%	0.814	0.164	1.7%	0.014	7.7%
y.N1	0.695	0.045	30.5%	0.019	26.7%	0.771	0.121	27.5%	0.020	53.8%
y.N2	0.694	0.044	32.0%	0.019	26.7%	0.764	0.115	31.4%	0.019	46.2%
y.N3	0.719	0.069	−5.6%	0.016	6.7%	0.821	0.172	−2.6%	0.013	0.0%
y.A1	0.702	0.052	19.9%	0.024	60.0%	0.718	0.069	58.9%	0.032	146.2%
y.A2	0.706	0.057	13.5%	0.023	53.3%	0.716	0.066	60.4%	0.032	146.2%
y.A3	0.724	0.074	−13.4%	0.017	13.3%	0.818	0.169	−0.7%	0.014	7.7%
y.4	0.703	0.053	18.8%	0.024	60.0%	0.718	0.068	59.2%	0.032	146.2%

8. References

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association, 91, 444–472.

Cochran, W.G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. Biometrics, 24, 295–313.

Cochran, W.G., Mosteller, F., and Tukey, J.W. (1954). Statistical Problems of the Kinsey Report (on Sexual Behavior in the Human Male). Washington D.C.: American Statistical Association.

Cook, E.F. and Goldman, L. (1989). Performance of Tests of Significance Based on Stratification by a Multivariate Confounder Score or by a Propensity Score. Journal of Clinical Epidemiology, 42, 317–324.

Couper, M.P. (2000). Web Surveys: A Review of Issues and Approaches. Public Opinion Quarterly, 64, 464–494.

Crown, W.H. (2001). Antidepressant Selection and Economic Outcome: A Review of Methods and Studies from Clinical Practice. The British Journal of Psychiatry, 179, s18–s22.

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992). Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics. Journal of Business and Economic Statistics, 10, 117–132.

D'Agostino, R.B., Jr. (1998). Propensity Score Methods for Bias Reduction for the Comparison of a Treatment to a Non-randomized Control Group. Statistics in Medicine, 17, 2265–2281.

Danielsson, S. (2002). The Propensity Score and Estimation in Nonrandom Surveys – An Overview. Available at http://www.statistics.su.se/modernsurveys/publ/11.pdf

Dillman, D.A. (2002). Navigating the Rapids of Change: Some Observations on Survey Methodology in the Early 21st Century. Draft of Presidential Address to American Association for Public Opinion Research Annual Meeting. Available at http://survey.sesrc.wsu.edu/dillman/papers.htm

Drake, C. (1993). Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. Biometrics, 49, 1231–1236.

Duncan, K.B. and Stasny, E.A. (2001). Using Propensity Scores to Control Coverage Bias in Telephone Surveys. Survey Methodology, 27, 121–130.

Frigoletto, F.D., Lieberman, E., Lang, J.M., Cohen, A.P., Barss, V., Ringer, S.A., and Datta, S. (1995). A Clinical Trial of Active Management of Labor. New England Journal of Medicine, 333, 745–750.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley and Sons.

Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47, 153–162.

Lavori, P.W. (1992). Clinical Trials in Psychiatry: Should Protocol Deviation Censor Patient Data? Neuropsychopharmacology, 6, 39–48.

Lavori, P.W. and Keller, M.N. (1988). Improving the Aggregate Performance of Psychiatric Disgnostic Methods When Not All Subjects Receive the Standard Test. Statistics in Medicine, 7, 723–737.

Lee, S. (2006). An Evaluation of Nonresponse and Coverage Errors in a Web Panel Survey. Social Science Computer Review, 24 (forthcoming).

Lee, S. (2004). Statistical Estimation Methods in Volunteer Panel Web Surveys. Unpublished Doctoral Dissertation. University of Maryland, Joint Program in Survey Methodology.

Lieberman, E., Lang, J.M., Cohen, A.P., D'Agostino, Jr. R, Datta, S., and Frigoletto, Jr. F.D. (1996). Association of Epidural Analgesia with Caesareans in Nulliparous Women. Obstetrics and Gynecology, 88, 993–1000.

Manfreda, K.L. (2001). Web Survey Errors. Unpublished Doctoral Dissertation. University of Ljubljana (Slovenia), Faculty of Social Science.

Obenchain, R.L. and Melfi, C.A. (1997). Propensity Score and Heckman Adjustments for Treatment Selection Bias in Database Studies. Proceedings of the American Statistical Association, Biopharmaceutical Section, 297–306.

Rosenbaum, P.R. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41–55.

Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association, 79, 516–524.

Rubin, D.B. (1973). Matching to Remove Bias in Observational Studies. Biometrics, 29, 581–592.

Rubin, D.B. and Thomas, N. (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice. Biometrics, 52, 254–268.

Schonlau, M., Fricker, R.D., Jr., and Elliott, M.N. (2002). Conducting Research Surveys via E-mail and the Web. Santa Monica, CA: RAND.

Schonlau, M., Zapert, K., Simon L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R., and Berry, S. (2004). A Comparison Between a Propensity Weighted Web Survey and an Identical RDD Survey. Social Science Computer Review, 22, 128–138.

Smith, P.J., Rao, J.N.K., Battaglia, M.P., Daniels, D., and Ezzati-Rice, T. (2000). Compensating for Nonresponse Bias in the National Immunization Survey Using Response Propensities. Proceedings of the American Statistical Association, Section on Survey Research Methods, 641–646.

Stone, R.A., Oborsky, S., Singer, D.E., Kapoor, W.N., and Fine, M.J. (1995). Propensity Score Adjustment for Pretreatment Differences between Hospitalized and Ambulatory Patients with Community-Acquired Pneumonia. Medical Care, 33, AS56–AS66.

Taylor, H. (2000). Does Internet Research Work? Comparing Online Survey Result with Telephone Survey. International Journal of Market Research, 42, 58–63.

Taylor, H. and Terhanian, G. (2003). The Evaluation of Online Research and Surveys Over the Last Two Years. Unpublished Manuscript.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J.W., and Terhanian, G. (2001). The Record of Internet-Based Opinion Polls in Predicting the Results of 72 Races in the November 2000 US Elections. International Journal of Market Research, 43, 127–135.

Terhanian, G. (2000). How to Produce Credible, Trustworthy Information through Internet-Based Survey Research. Paper presented at the annual meeting of the American Association for the Public Opinion Research, Portland, OR.

Terhanian, G. and Bremer, J. (2000). Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research. Research paper: Harris Interactive.

Terhanian, G., Bremer, J., Smith, R., and Thomas, R. (2000). Correcting Data from Online Survey for the Effects of Nonrandom Selection and Nonrandom Assignment. Research paper: Harris Interactive.

Varedian, M. and Forsman, G. (2002). Comparing Propensity Score Weighting with Other Weighting Methods: A Case Study on Web Data. Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg Beach, FL.

Vartivarian, S. and Little, R. (2003). On the Formation of Weighting Adjustment Cells for Unit Nonresponse. University of Michigan Department of Biostatistics Working Paper Series.

Vehovar, V. and Manfreda, K.L. (1999). Web Surveys: Can the Weighting Solve the Problem? Proceedings of the American Statistical Association, Section on Survey Research Methods, 962–967.

Received March 2004
Revised August 2005