Survey Research Methods (2007)
Vol.1 , No.3 , pp. 155-163
ISSN 1864-3361
http://www.surveymethods.org

Are "Webographic" or attitudinal questions useful for adjusting estimates from Web surveys using propensity scoring?

Matthias Schonlau
RAND Corporation

Arthur van Soest
RAND and Tilburg University

Arie Kapteyn
RAND Corporation

Inference from Web surveys may be affected by non-random selection of Web survey participants. One approach to reduce selection bias is to use propensity scores and a parallel phone survey. This approach uses demographic and additional so-called Webographic or lifestyle variables to balance observed differences between Web survey respondents and phone survey respondents. Here we investigate some of the Webographic questions used by Harris Interactive, a commercial company specializing in Web surveys. Our Webographic questions include choice of activities such as reading, sports and traveling and perceptions about what would constitute a violation of privacy. We use data from an existing probability sample of respondents of age at least 40 who are interviewed over the phone, and a corresponding sample of respondents interviewed over the Web. We find that Webographic questions differentiate between on and offline populations differently than demographic questions. In general, propensity score adjustment of variables in the Web survey works quite well for a number of variables of interest (including home ownership and labor force participation). For two outcomes, (having emotional problems and often experiencing pain) the process of adjusting for demographic variables leads to the discovery of an instance of Simpson's paradox, implying a differential mode effect or differential selection. We interpret this mainly as the result of a mode effect, where sensitive questions are less likely to be affected by social desirability over the Internet than over the phone.
Keywords: propensity scoring, Web survey, selection bias, Webographic variables, lifestyle variables, Simpson's paradox

Introduction

Demographic questions are asked to understand whether important segments of the population respond differently. They frequently do. To the extent that some segments of the population (e.g. racial/ethnic groups) are underrepresented in a survey, nonresponse or post stratification weights are introduced to correct for the imbalance. The weighted sample is then representative of the population with respect to those demographic variables.

Selection bias arises both from non-coverage and non-response bias. Non-coverage bias in a Web survey arises because part of the population of interest has no access to the Web. Since fewer people have no access to a phone, we expect Web survey respondents to be less representative of a general population than phone survey respondents. This particularly applies if the population of interest is an older part of the total population, such as the 40+. In cases where Web surveys use convenience samples additional bias may be expected due to self-selection. Unit non-response is common to all socio-economic surveys, including Web surveys as well as phone surveys.

Because Web survey respondents may be less representative, it is reasonable to look for additional imbalances that are not already captured by demographic variables. Harris Interactive, a commercial company specializing in Web surveys, has experimented with scores of additional so-called Webographic questions. Other researchers call them also "lifestyle" (Varedian and Forsman 2003) or "attitudinal" (Lee 2004, 2006) questions. These questions supposedly capture the differences between the online and the off-line populations and allow adjusting for the non-coverage selectivity in Web surveys. Harris Interactive does not publish the questions it uses. However, Webographic questions can easily be detected in Harris Interactive surveys: they are questions that are unrelated to the survey topic, usually near the end of a survey. Examples of these questions, which are also used in the analysis in this paper, are given in Appendix.

It is not clear to what extent phone survey respondents and Web survey respondents really differ with respect to Webographic questions or whether differences disappear after adjusting for differences in the distribution of demographic variables. It is also not clear how well the adjustment using webographic questions works for various outcomes. These are the issues we address in this paper. In section 2 we provide a brief review of the empirical literature on uses of propensity

Contact information: Matthias Schonlau, matt@rand.org

scoring for survey reweighting. Section 3 contains details on the weighting method for propensity scoring that we use and a description of the data. We use data from an existing probability sample of respondents of age at least 40 who are interviewed over the phone, and a corresponding sample of respondents interviewed over the Web. Sections 4 and 5 contain the results and a discussion.

Background

Propensity scoring was originally developed to draw inferences in observational studies (Rosenbaum and Rubin 1983). To adjust for nonresponse bias one needs to model the probability of response. To make the sample of respondents representative of the complete population of respondents and non-respondents, each respondent is assigned a weight equal to the inverse of the estimated probability of response $1/\hat{p}$. This scheme gives greater weight to respondents that are more similar to non-respondents (Little and Rubin 1987). Existing applications include the following. Smith et al. (2000) adjust for bias due to non-response by constructing strata from propensity scores. They estimate the percent of children vaccinated with four critical vaccines. Using information about responding and non-responding providers they construct 5 propensity strata and adjust the weights within each cell. Duncan and Stasny (2001) adjust for bias due to non-coverage in a telephone survey by arguing that households without telephone service are similar to transient households – households with interrupted telephone service. They compute the propensity of a household being transient and then assign transient households higher weights based on propensity scores. Garren and Chang (2002) also adjust for coverage bias using a related method.

DeVries et al. (2005) use propensity scoring to assess the mode effect between a phone and a mail survey for the CAHPS hospital survey. They use variables from administrative data as covariates for the logistic regression that models the probability of being in a particular mode.

Several authors (Isaksson and Forsman 2003, Lee 2006, Schonlau et al. 2004a, 2004b, Varedian and Forsman 2003) use propensity scoring to reduce the bias in estimates based on convenience samples. Many Web surveys form convenience samples. To adjust the estimates in this way requires a reference survey; often a Random Digit Dialing (RDD) phone survey is used for this purpose. Propensity scoring is applied, with propensity weights estimated on the basis of the combined sample of the Web survey and the reference survey. Questions that are asked in both the phone and the Web survey are used as covariates in a logistic regression. These questions may include Webographic or lifestyle questions that capture differences between the online and offline population.

Two important questions arise: (1) Does the propensity score adjustment reduce or eliminate the selection bias in estimates of population statistics based upon a Web survey? And (2) what questions should be asked in both the Web survey and the reference survey? The literature seems to suggest that the adjustment almost always reduces the bias but does not necessarily eliminate it completely (e.g., Lee 2006). Schonlau et al. (2004b) compare estimates from an RDD phone survey with propensity-adjusted estimates from a Web survey conducted by Harris Interactive. They find that the phone survey estimates were not significantly different from the adjusted Web survey estimates for only 8 out of 37 questions investigated.

Schonlau et al. (2004a) analyze data from the 2002 Health and Retirement Study (HRS), a CAPI/CATI survey representative for the US 50+ population and their spouses. A subset of HRS respondents with Internet access subsequently also responded to a Web survey. Using demographic variables and some other common variables, Schonlau et al. (2004a) are able to adjust for selectivity in many but not in all cases. In particular, the percentage of households owning stock, predicted based on the Web survey, remained significantly different from the estimate based on the 2002 HRS.

Varedian and Forsman (2003) experiment with propensity score weighting in the context of a marketing survey about the use of hygiene products and attitudes toward local banks. They find that none of the weighting schemes had pronounced effects on any of their estimates.

Isaksson and Forsman (2003) study political polls for the 2002 election in Sweden. They find that propensity adjustment based on a set of lifestyle questions reduces the absolute differences with the actual election results more than the usual post stratification by sex, age, class, and dwelling (big-city vs. not). Moreover, adding the latter four variables to the propensity adjustment did not appreciably improve the bias reduction compared to using the lifestyle questions only.

The second question is what questions should be used for the propensity adjustment. All researchers adjust for differences in the distributions of some demographic variables. Schonlau et al. (2004a) find that at a minimum a set of demographic variables are needed to adjust for selection bias and also find self assessed health status useful. As mentioned, in addition to age, gender and region (in Sweden), Varedian and Forsman (2003) emphasize the importance of including lifestyle questions that are meant to capture a respondent's "modernity". Lee (2006) uses the "lifestyle" (non-demographic) variables self-rated social class, employment status, political party affiliation, having a religion and opinion toward ethnic minorities as variables for propensity scoring. Lee finds, however, that this particular set of non-demographic variables makes little difference. She points out that most of her non-demographic variables are not significantly related to the two outcomes, "warm feeling towards blacks" and whether one voted in the 2000 election.

While Harris Interactive does publish successful applications (Taylor et al. 2001), the company does not publish its research on the use of this method or which Webographic questions are valuable. Presumably, however, Webographic questions that continue to be used are useful. The secrecy leads to oblique references in the literature. For example, Danielsson (2004) reports "In some preliminary papers from Harris Interactive (not to be quoted) the effects of using propensity scores in Web-surveys are reported and the results are surprisingly good (but the method used is not clearly described)." The general propensity scoring method Harris

Interactive used, without details on the specific Webographic questions, was described in Schonlau et al. (2004b).

Method

We consider the combined sample of Web survey respondents and phone survey respondents. The propensity score is the conditional probability that the i^{th} respondent is a Web survey respondent in the combined sample of phone and Web respondents:

$$p_i = P(Z_i = 1|X_i)$$

where Z_i is an indicator variable for participation in the Web survey. The covariates X_i contain information that is collected in both the phone survey and the Web survey. In this paper we explore the value of answers to demographic and Webographic questions as covariates.

The propensity score is a balancing score, that is, respondents in both surveys with the same propensity score have the same distribution of X (Rosenbaum and Rubin 1983). This is a theoretical property. In practice, $p(X)$ is estimated, usually with a logit model or another parametric model involving functional form assumptions. One can test for lack of balance in X, for example, by testing whether the means of covariates after the propensity score adjustment differ in the two surveys. If they do, this points at inappropriate adjustment, e.g., because the functional form of $p(X)$ is misspecified.

For the inference to work for a variable of interest Y we must assume strong ignorability. Let Y_1 be the response to a particular question Y of an individual to a Web survey and Y_0 the response of the same individual to the same question in the phone survey. The two answers may differ because of a mode effect. Note that each respondent only responds to one survey and therefore only one of the two outcomes is observed. Membership of the Web survey, Z, is defined to be strongly ignorable with respect to Y if, conditional on X, (Y_0, Y_1) and Z are independent. Strong ignorability implies that there are no unobserved questions/variables that explain selection into the Web sample that are also related to the question of interest Y. After propensity adjusting for selectivity we can test whether the adjusted sample means of Y_1 in the Web survey and Y_0 in the phone survey differ. This is a joint test of no mode effects $(Y_1 = Y_0)$ and strong ignorability.

In this paper we focus on this test and on the extent to which adjustment using propensity scores based upon different sets of covariates X (including and not including Webographic questions) brings the web survey answers in line with the phone survey answers. We do not attempt to choose the weights so as to make the sample representative for a population of interest. This would also require, for example, adjusting for unit non-response, which we do not aim for (and which would be difficult due to lack of information on non-respondents). If we find that ignorability is rejected, it might also be due to non-ignorability of unit response, but we find this less plausible than other explanations, i.e., non-ignorable coverage (Web access) or mode effects $(Y_0 \neq Y_1)$.

We conducted a phone survey (Spring 2004) and a Web survey (Fall 2003) containing the same questions; 516 respondents completed the phone survey and 1128 the Web survey. Respondents of the Internet survey are participants of the RAND American Life Panel (ALP). Both samples have been drawn as part of a grant provided by the National Institute on Aging in the United States.

The respondents in the ALP are recruited from among individuals age 40 and older who are respondents to the Monthly Survey (MS) of the University of Michigan's Survey Research Center (SRC). The MS is the leading consumer sentiment survey that incorporates the long-standing Survey of Consumer Attitudes (SCA) and produces, among others, the widely used Index of Consumer Expectations. Each month, the MS interviews approximately 500 households, of which 300 households are a random-digit-dial (RDD) sample and 200 are re-interviewed from the RDD sample surveyed six months previously. At the end of the re-interview respondents who meet the age criterion are asked to join the phone sample or the ALP.[1]

As is clear from this description, building up the two samples took some time. Initially respondents for the phone sample were recruited in a way that made them representative of the population over 40. That is, a substantial part of the respondents in the phone sample did have Internet access. It also implied that many respondents without Internet access were not used in the phone sample, because to keep the phone sample representative and to recruit enough respondents for the Internet sample the number of respondents with Internet access became the bottleneck. It turned out that this led to a slower than expected build-up of the samples. To speed up the process of building the sample it was then decided to allocate all willing respondents with Internet access to the Internet sample and to fill the telephone sample primarily with willing respondents without Internet access. As a result of this, the difference between the phone sample and the Internet sample is bigger than it would have been, if the phone sample were fully representative of the population over 40. This increases the challenge of reweighting such that the Internet sample can reproduce variable distributions observed in the phone sample.

The sample design therefore implies the following (ignoring unit non-response, as explained above). The population consists of two parts, those with and those without Web access. The Web sample is a simple random sample from respondents with Web access. The phone sample is a stratified sample from the subpopulations with and without Web access (with unobserved inclusion probabilities that are probably different). The indicator variable Z indicates whether a respondent is in the Web sample. Because of the design, the null hypothesis that Z is ignorable for a variable of interest Y is the same as the null hypothesis that Web access is ignorable for Y. However, because the phone sample consists of respondents both with and without internet access under the alternative

[1] The description is correct for the period over which the data used in this paper were collected. Since the Fall of 2006, respondents over age 18 are eligible to join the ALP; moreover respondents without Internet access are given the opportunity to join the ALP, in which case they are provided with Internet access by means of a Web TV (also called Internet player).

hypothesis any differences in adjusted sample means will be reduced compared to the case where the phone sample would be a simple random sample from the non Web access subpopulation (but the difference will not be reduced to zero).

It is worth noting furthermore that by concentrating on older ages, the challenge for Web surveys to produce population representative measures is increased greatly. Internet access falls sharply with age, so Web surveys cover a relatively smaller part of the population at older ages (Cheeseman Day et al. 2005).

Our logistic regression model to compute the propensity score includes the following demographic variables: gender, log10 income, (log10 income) squared, age/10, education (less than high school, high school but less than college, college or more), primary language is English, born in the US, and self assessed health status (excellent and very good [combined], good, fair, poor), coded as indicator variables. Self assessed health status was included because we previously found it to be an important predictor (Schonlau et al. 2004a). Even though health status is not a demographic variable, we refer to the set of variables as demographic variables to avoid overly cumbersome language. Dummy variables for whether the primary language is English and whether the respondent was born in the US are included as well. Race and ethnicity are not available.

Webographic variables fall into one of four categories which we label attitudinal variables, factual variables, privacy variables and variables related to knowing gay people. The questions themselves are shown in Appendix. These questions continue to be used by Harris Interactive.

Propensity scores can be used in different ways. We use the inverse propensity scores as weights (Rosenbaum 1987). One further subtle correction is needed to the weights, since in our case propensity scores refer to a population of Web and phone survey respondents combined. We want to compare weighted Web sample based estimates with (unweighted) phone sample estimates, and do not aim at making the estimates representative of the complete population.[2] Therefore we multiply the weights with the probability of being in the phone survey $(1 - p_i)$. Specifically, for the Web survey estimates we use the weights $w_i = (1 - p_i)/p_i$ if respondent i is a Web survey respondent. This argument follows Hirano and Imbens (2001, Section 2.5) and Hirano et al. (2003, Section 4.3).

We test for evidence of lack of balance by testing for differences in means. Because significance is affected by the use of unequal weights we also investigate reduction in effect sizes (Cohen 1988). The effect size for variable k is defined as $\mu_k^{web} - \mu_k^{phone} / stddev_k$ where the standard deviation is computed from the pooled data and where μ_k^{web} and μ_k^{phone} are the sample means in the two samples. Comparing effect sizes before and after adjustment indicates whether standardized differences in means are reduced.

We investigate adjustments for several outcomes Y: home ownership, whether or not the respondent is working, has a stressful job, a doctor has ever told the respondent he/she has an emotional or psychological problem, whether he or she often experiences pain, as well as the respondent's average daily number of servings of fruit and vegetables, the number of hours a week with moderate activity, and the number of days on which the respondent watched news in the last 30 days.

For each outcome we compute phone survey estimates and unadjusted and adjusted Web survey estimates with varying sets of adjusting variables X. Adjustments are based on demographic variables only, Webographic variables only, or on both sets of variables. We also explore how leaving some Webographic variables out of X affects the estimates. Specifically, different groups of variables (factual, attitudinal, perception of violation of privacy, knowing anyone who is gay) are left out alternatively to study how the adjusted estimates change.

Results

Table 1 explores the balance of the covariates used in the propensity scores by presenting means in the phone survey, in the Web survey, as well as means of the Web survey estimates adjusted for demographics only and adjusted for both demographic and Webographic variables. Table 1 also gives effect sizes before the adjustment and after the adjustment for demographic and Webographic variables.

Web survey respondents are more likely to take chances and to say that they feel alone. They travel more often, read a book more often and participate in sport more often. They perceive violations of privacy more often, in particular for airport search and credit card storage. They are more likely to know someone who is gay.

If the adjustment works, that is, if balance is achieved, then the adjusted estimates should not differ significantly from the phone survey estimates. Estimates corresponding to differences significant at $\alpha = 0.05$ are denoted by an asterisk. The phone survey estimates differ significantly from many of the unadjusted Web survey estimates. After adjusting for demographic variables the demographic estimates are no longer significantly different. However, several imbalances among the Webographic variables remain. No significant imbalances remain after adjusting for both sets of variables. Because the use of the propensity weights inflates standard errors this may be artificial. Hence we also want to see a reduced effect size. The average effect size across all variables is reduced from 0.20 to 0.03.

Table 2 provides several estimates of the prevalence or means of a number of variables of interest: CATI estimates, unadjusted Web estimates, adjusted Web estimates using only the demographic variables, and Web estimates adjusted using both the demographic and the Webographic variables. Web based estimates that differ significantly ($\alpha = 0.05$) from the phone survey estimates are denoted by an asterisk. About half

[2] Testing whether the Web survey adjusted estimates using weights $1/p_i$ would balance the combined sample of unweighted Web sample and phone sample would be equivalent. Because the combined sample does not represent any population of interest, the interpretation of the adjusted estimates using weights $1/p_i$ would be less clear, however.

Table 1: Balance of Phone and Web survey estimates before and after adjusting for demographic and combined demographic and Webographic questions.

		Phone	Web			Effect sizes	
			Unadjusted	Adjusted (Demographic)	Adjusted (Demographic and Webographic)	Phone vs Unadjusted Web	Phone vs Adjusted (Demographic and Webographic)
Age		64.6	54.0*	60.8	60.9	−0.13	−0.04
log hh income		4.3	4.8*	4.2	4.3	0.49	−0.01
log hh income squared		19.1	23.2*	18.6	19.0	0.62	−0.01
		(%)	(%)	(%)	(%)		
Male		39.1	46.9*	39.7	42.6	0.11	0.05
Language not English	yes/no	6.2	3.0*	4.4	3.8	−0.11	−0.08
Born in the US	yes/no	93.6	91.4	94.5	94.7	−0.06	0.03
Education	some college	61.1	21.3*	61.8	60.8	−0.63	0.00
	≥ 4 yr college	5.4	28.0*	4.8	5.1	0.45	−0.01
Self assessed health	good	29.6	27.0	29.5	31.3	−0.04	0.03
	fair	23.2	10.0*	26.0	28.2	−0.25	0.10
	poor	11.5	1.9*	10.6	8.2	−0.28	−0.09
Attitudinal Questions	eager to learn	57.1	60.4	59.6	56.9	0.05	0.00
	takes chances	53.9	60.9*	62.1*	52.4	0.10	−0.02
	often feel alone	54.1	60.6*	58.4	53.5	0.09	−0.01
In the last month have you...	traveled?	29.8	53.5*	44.2*	32.9	0.35	0.05
	participated in sport?	11.7	25.5*	17.0	11.7	0.26	0.00
	read a book?	63.2	77.5*	76.7*	71.0	0.22	0.12
Violation of Privacy	airport search	28.9	18.4*	14.5*	23.6	−0.17	−0.09
	cookies	72.1	71.9	65.4	68.4	0.00	−0.06
	phone calls	70.5	72.6	77.0	74.7	0.03	0.06
	aids screening	38.8	31.6*	25.5*	40.1	−0.11	0.02
	credit card storage	84.3	65.3*	71.9*	87.0	−0.32	0.04
Know anyone who is gay	no	40.4	19.7*	28.2*	39.4	−0.33	−0.02
	family	19.5	22.2	19.5	16.7	0.05	−0.05
	closefriend	21.6	18.5	13.9*	21.4	−0.05	0.00
	acquaintance	40.2	44.2	35.1	38.8	0.06	−0.02
	other	19.3	25.3*	23.8	19.3	0.10	0.00

Estimates that are significantly different ($\alpha = 5\%$) from the phone survey estimates are denoted by an asterisk.
Demographic and Webographic questions are separated by a horizontal line.

Table 2: Phone survey estimates and unadjusted and adjusted estimates Web survey estimates.

	own house (%)	working (%)	job stressful (%)	emotional/ psych problem (%)	often experiences pain (%)	servings of fruit/vegetable per day	hours a week moderate activity	watch news program last 30 days
Phone	72.1	33.9	63.3	11.5	37.4	3.5	3.5	23.1
Unadjusted Web	88.5*	65.8*	72.1*	14.4	41.1	3.8	3.8	20.0*
Adjusted using Demographics only	82.3*	34.6	69.1	17.6*	54.7*	3.9	3.4	21.8
Adjusted only Webographics	87.5*	63.0*	68.9	12.8	39.6	3.7	3.6	20.8
Adjusted using Demographics and Webographics	78.4	32.6	67.8	20.3*	57.5*	3.8	3.4	22.5

Estimates that are significantly different ($\alpha = 5\%$) from the phone survey estimates are denoted by an asterisk.

Table 3: Percentage of respondents reporting a doctor has ever told the respondent he/she has an emotional or psychological problem

self assessed health status	phone (%)	Web (%)	difference (%)
excellent/very good	3.3	10.1	6.8
good	10.5	15.6	5.1
fair	16.8	30.4	13.6
poor	28.8	52.4	23.6
overall (weighted)	11.5	14.4	2.9

Estimates are shown overall and by each self assessed health status category by survey mode.

of the unadjusted Web estimates do not differ significantly from the phone estimates. Estimates that differ imply that Web survey respondents are more likely to own a house, to be working, to consider their job stressful, and to watch slightly fewer news programs.

Adjustments tend to reduce discrepancies between Web and phone, but not always. For all but two variables ("emotional/psychological problem" and "often experiences pain") adding the Webographic variables either reduces discrepancies - rendering the Web based estimate statistically insignificantly different from the phone estimate - or does not affect the significance of an already insignificantly different estimate. The adjustment for "working" works only when the demographic variables are included. Most - but not all - of that adjustment is due to the respondent's age. The estimate of homeownership requires both demographic and Webographic variables to achieve insignificance. None of the adjustments substantially change the estimates for the number of fruit servings and hours a week of moderate activity.

Adjusting on the basis of demographic variables for the two remaining outcomes, "emotional/ psychological problem" and "often experiences pain" (either with or without Webographic variables) increases the discrepancy. We explain this in more detail by concentrating on just one of the variables, self assessed health status. Table 3 gives the percentage of respondents who reported emotional or psychological problems by self assessed health status for both phone and Web survey respondents.

In each category the percentage of Web survey respondents reporting emotional/psychological problems is higher, by 5.1%-points to 23.6%-points. However, the combined mean difference is only 2.9%. This phenomenon is known as Simpson's paradox (Simpson 1951) and is due to the fact that the distribution of self assessed health status is very different for Web and phone respondents. Specifically, relatively fewer Web respondents report to be in fair or poor health.

The combined mean difference is the difference between the unadjusted Web survey and the phone survey. The propensity adjustment assigns Web respondents with poor or fair health a greater weight because they are under represented, thereby exacerbating rather than reducing differences.

We see at least two possible explanations for these differences and both are mode effects. The first one would be due to differential recency effects over the phone and over the Internet. Particularly among elderly respondents, recency effects (a tendency to recall the last item mentioned in a list) may be relatively prominent over the phone: Since "fair" and "poor" are the last two items of the five point health scale, these may be more likely to be chosen in a phone interview than in a Web interview, where recency effects are less likely. This would exaggerate the health differential between Internet respondents and phone respondents. Secondly, social desirability may make a respondent less inclined to admit to emotional and psychological problems in a phone interview than in a self-administered interview over the Internet. Both effects would work in the direction of generating Simpson's paradox. To a perhaps somewhat lesser extent a similar explanation may apply to the other outcome, "often experiences pain". In addition, some response differences between the web and the phone survey might be due to different fielding periods. For example, respondents may perceive pain ("Do you often experience pain?") differently in the spring than in the fall.

Table 4: Adjusted estimates with different groups of variables are left out to study their effect.

Adjusted using Demographics and Webographics *except*	own house (%)	working (%)
Factual	78.0	30.8
Attitudinal	78.7	33.3
Know gay	79.2	36.0
Factual + attitudinal	79.0	31.6
Privacy	80.9*	33.1
Self-Assessed Health Status	81.4*	35.9
Know gay + Privacy	82.4*	36.8
Benchmark:		
Phone Survey estimates	72.1*	33.9

Estimates that are significantly different (α = 5%) from the phone survey estimates in Table 2 are denoted by an asterisk.

For home ownership the adjustment works only when Webographic variables are included. Table 4 explores how estimates are affected by removing subsets of the Webographic variables (and health status). For example, to explore the effect of the privacy questions, we remove the privacy questions from the full set of variables to see how the estimates change. For homeownership removing either the privacy questions or health status would result in an estimate significantly different from the phone survey estimate. We also explored the variable working; none of the subsets of variables affects the estimate much.

Discussion

Demographic variables differentiate between on and offline populations differently than Webographic variables, because adjusting for demographic variables does not balance Webographic variables. The most imbalanced Webographic variables after adjusting for demographic variables are: (not)

knowing anyone who is gay, perceived privacy violation by airport searches, storage of credit card information and aids screening, having read a book in the last month, having traveled in the last month and taking chances. If one had to choose a subset of Webographic questions because of cost or other constraints (e.g. phone interview time) the variables that are most imbalanced after the adjustment for demographic variables are primary candidates. Among the Webographic questions in our study privacy appears to affect the estimates more than other Webographic questions.

The question about knowing gay persons is somewhat unusual. For respondents who wonder why the question is being asked Harris Interactive now has an html link to the following text:

> *"Collecting data from all respondents on this question is important so that we can better and more reliably report differences and similarities between people of different sexual orientations. We understand that you might be concerned about sharing this information. Please be assured that the responses you provide are kept completely confidential. Any identifying information will be separated from your answers. Results are reported using the average, or pooled answers to the questions, instead of the responses of any one individual."*

Harris interactive uses a similar text link for some other questions including income.

Because Webographic variables appear to be useful in some cases there seems to be no downside to adding the variables when they are available. This finding is also supported in the propensity scoring literature. Rubin and Thomas (1996) advocate including all available covariates into the propensity model because any omitted variable may lead to bias. In a simulation study Drake (1993) shows that the propensity method is relatively robust against misspecifying the propensity score models (e.g. adding a quadratic term).

The discovery of instances of Simpson's paradox for two of the outcome variables is a reminder that the adjustment can only work when there is no mode effect and strong ignorability holds for propensity scoring. If our suspicion that social desirability plays a role in explaining the differences is correct the implication is not necessarily that the Web leads to unrepresentative results, but rather that for certain topics the Web may be better suited than phone interviews. Regardless of whether our conjecture that Simpson' paradox here is due to social desirability biases is correct, the results confirms that when no adjustment was made, 4 out of 8 variables showed significantly different estimates; and when adjustments were made, this number decreased.

It is unlikely that one set of Webographic variables can remove bias for all outcomes. Future work may need to concentrate on establishing variables that work for common outcomes in different substantive areas.

Acknowledgements

Support for this research comes from grant R01AG20717 from the National Institute of Aging of the U.S. National Institutes to RAND (Arie Kapteyn, P.I.) and from the University of Michigan's Survey Research Center (Robert J Willis, P.I.).

References

Cheeseman Day, J., Janus, A., & Davis, J. (2005). *Computer and internet use in the united states: 2003*. US Census Bureau.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Laurence Erlbaum & Associates.

Danielsson, S. (2004). *The propensity score and estimation in nonrandom surveys: an overview.* Department of Statistics, University of Linköping; Report no. 18 from the project "Modern statistical survey methods." Retrieved April, 2006. Available from `http://www.statistics.su.se/modernsurveys/publ/11.pdf`

De Vries, H., Elliot, M. N., Hepner, K. A., Keller, S. D., & Hays, R. D. (2005). Equivalence of mail and telephone responses to the cahps hospital survey. *Health Services Research*, *40*(6), 2120-2139.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*(4), 1231-1236.

Duncan, K. B., & Stasny, E. A. (2001). Using propensity scores to control coverage bias in telephone surveys. *Survey Methodology*, *27*(2), 121-130.

Garren, S. T., & Chang, T. C. (2002). Improved ratio estimation in telephone surveys. adjusting for noncoverage. *Survey Methodology*, *28*(1), 63-76.

Hirano, K., & Imbens, G. (2001). Health services and outcomes. *Research Methodology*, *2*, 259-278.

Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161-1189.

Isaksson, A., & Forsman, G. (2003). *A comparison between using the web and using the telephone to survey political opinions.* In Proceedings of the Section on Survey Research Methods. (American Statistical Association)

Lee, S. (2004). *Statistical estimation methods in volunteer panel web surveys.* (Unpublished Doctoral Dissertation. University of Maryland, Joint Program in Survey Methodology)

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, *22*(2), 329-349.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. New York: John Wiley and Sons.

Rosenbaum, P. R. (1987). Model based direct adjustment. *Journal of the American Statistical Association*, *82*, 387-394.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*, 254-268.

Schonlau, M., Soest, A. V., Kapteyn, A., Couper, M., & Winter, J. (2004a). *Adjusting for selection bias in web surveys using propensity scores: the case of the health and retirement study (hrs).* In Proceedings of the Section on Survey Research Methods. CD-ROM. (American Statistical Association)

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K., Marcus, S., Adams, J., et al. (2004b). A comparison between a propensity

weighted web survey and an identical rdd survey. *Social Science Computer Review*, *22*(1), 128-138.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B*, *13*(2), 238-241.

Smith, P. J., Rao, J. N. K., Battaglia, M. P., Daniels, D., & Ezzati-Rice, T. (2000). *Compensating for nonresponse bias in the national immunization survey using response propensities.* In Proceedings of the Section on Survey Research Methods, pp. 641-646. Available from `http://www.cdc.gov/nis/pdfs/estimation_weighting/smith2000.pdf` (American Statistical Association)

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W., & 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*(2), 127-135.

Varedian, M., & Forsman, G. (2003). *Comparing propensity score weighting with other weighting methods: A case study on web data.* (Presented at the 2003 AAPOR meetings)

Appendix: Webographic questions

The questions were first used by Harris Interactive with the following differences: Harris Interactive asks the attitudinal questions on a Likert scale. The last response option in the question about violation of privacy (Electronic storage of credit card numbers) was added. For the same questions a response option "none of these" was removed.

Attitudinal Questions

Do you often feel alone? (yes/no)
Are you eager to learn new things? (yes/no)
Do you take chances? (yes/no)

Factual Questions

In the last month have you traveled? (yes/no)
In the last month have you participated in a team or individual sport? (yes/no)
In the last month have you read a book? (yes/no)

Privacy

Which of these practices, if any, do you consider to be a serious violation of privacy?
Please check all that apply.

1. Thorough searches at airport checkpoints, based on visual profiles
2. The use of programs such as 'cookies' to track what an individual does on the Internet
3. Unsolicited phone calls for the purpose of selling products or services
4. Screening of employees for AIDS
5. Electronic storage of credit card numbers by Internet stores

Know anyone who is gay

Do you know anyone who is gay, lesbian, bisexual, or transgender?
Please check all that apply.

1. Yes, a family member
2. Yes, a close personal friend
3. Yes, a co-worker
4. Yes, a friend or acquaintance (not a co-worker)
5. Yes, another person not mentioned
6. No