WORKING
PAPER

Selection Bias in Web Surveys and the Use of Propensity Scores

MATTHIAS SCHONLAU
ARTHUR VAN SOEST
ARIE KAPTEYN
MICK P. COUPER

WR-279

April 2006

LABOR AND POPULATION

Selection bias in Web surveys and the use of propensity scores

Matthias Schonlau[1], Arthur van Soest[2], Arie Kapteyn[1], Mick Couper[3]

[1]RAND

[2]RAND and Tilburg University

[3]University of Michigan

Corresponding author: Matthias Schonlau, RAND, 201 N Craig St, Suite 202, Pittsburgh, PA, 15213

matt@rand.org

Abstract

Web surveys have several advantages compared to more traditional surveys with in-person interviews, telephone interviews, or mail surveys. Their most obvious potential drawback is that they may not be representative of the population of interest because the sub-population with access to Internet is quite specific. This paper investigates propensity scores as a method for dealing with selection bias in web surveys. Our main example has an unusually rich sampling design, where the Internet sample is drawn from an existing much larger probability sample that is representative of the US 50+ population and their spouses (the Health and Retirement Study).

We use this to estimate propensity scores and to construct weights based on the propensity scores to correct for selectivity. We investigate whether propensity weights constructed on the basis of a relatively small set of variables are sufficient to correct the distribution of other variables so that these distributions become representative of the population. If this is the case, information about these other variables could be collected over the Internet only. Using a backward stepwise regression we find that at a minimum all demographic variables are needed to construct the weights. The propensity adjustment works well for many but not all variables investigated. For example, we find that correcting on the basis of socio-economic status by using education level and personal income is not enough to get a representative estimate of stock ownership. This casts some doubt on the common procedure to use a few basic variables to blindly correct for selectivity in convenience samples drawn over the Internet. Alternatives include providing non-Internet users with access to the Web or conducting web surveys in the context of mixed mode surveys.

1. Introduction

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

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

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

A distinction needs to be made between coverage error, non-response error, and the usual random sampling error. Coverage error and non-response errors may lead to biased estimates, whereas sampling error is due to random variation. Even a perfect random sample will lead to sampling error because only a subset of the population is sampled. As a consequence, the population characteristic of interest and corresponding sample statistic used to estimate it differ. For example, the population mean can be estimated using the sample mean. The deviation in case of a perfectly random sample is the sampling error. Non-response error arises when a selective group does not answer a given question of interest (item non-response) or does not participate in the survey at all (unit non-response). In this case the sample mean ignoring the nonrespondents typically may lead to a biased estimator for the mean of the population (the population includes the non-respondents), to the extent that respondents differ from non-respondents on the variable of interest (e.g., Groves and Couper, 1988). Coverage error arises when the survey is designed such that a specific part of the target population is not included on the frame. For example, if respondents are selected by randomly dialing telephone numbers (RDD, random digit dialing), households with no phone will never be selected. The target population, however, usually also includes the households without phone. If, for example, the purpose is measuring average household income and the households without a phone have lower average income than others, ignoring the non-phone group will lead to an upward bias of average household income in the population. Of course the size of the bias would probably be limited in this example, because the group of households without a phone connection is small.

For Internet surveys, however, problems with a good coverage of the target population play a much larger role. If the sampling design implies that only those with an Internet connection are selected, the coverage error can be substantial. Because Internet use is not yet equally spread among all socio-economic groups the coverage problem is likely to lead to biased estimates on variables related to SES. This may be a particular problem if the target population is the elderly population, where Internet access

is less wide spread than in the population as a whole. One way to address this problem is to provide households without Internet access with the tools to get access, relieving the coverage error. This is, for example, done by Knowledge Networks in the US and CentERdata in the Netherlands. While this avoids the coverage error it is still subject to nonresponse error. For Knowledge Networks multiple levels of nonresponse errors lead to overall response rates of substantially less than 30%.

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

Today there are many Internet-based samples used to conduct surveys of various kinds. Typically, no attempt is made to make these samples cover more than the population of active Internet users. For example, prospective respondents may be recruited by email or by placing banners on frequently visited Web sites. There are obvious problems with such samples (cf. Couper, 2000) which are often ignored (Schonlau et al., 2002). Not only are the respondents a selective sample of the population at large, they are the most savvy computer users and therefore may be expected to be much quicker at understanding and answering Internet interviews than others. Because they may respond differently, one needs to find a way to validly generalize from such a sample to a broader group of respondents.

An important tool to correct for the selection effect in operational studies is the use of propensity scores (Rosenbaum and Rubin, 1983, 1984; Little and Rubin, 2002). Harris Interactive (Taylor, 2000) uses the propensity scoring methodology to reweight a convenience Web sample based on a monthly random phone sample using various sets of about five "webographic" variables. "Webographic" questions are questions that are thought to best capture the differences between the general population and people able and willing to answer Web surveys. The use of propensity scores for surveys requires two samples: a random sample for calibration and a second sample that is calibrated.

This paper investigates the usefulness and validity of propensity scores to correct for the selective nature of an Internet sample drawn from the Health and Retirement Study (HRS). The unique feature of the Internet sample is the sampling design, which is particularly appropriate for studying this issue. The HRS is a representative survey of elderly cohorts in the US (with sampling weights to correct for unit non-response and age-stratified sampling). The Internet sample is randomly drawn from a subsample of the HRS respondents who reported to have Internet connection. The HRS is unique in that a vast amount of information is already available on all respondents, irrespective of whether they were included in the Internet sample or not, which greatly enhances the scope for reweighting. This helps enormously to analyze selection into the Internet survey and to study how well propensity scores constructed on the basis of a small set of HRS variables perform in correcting for selectivity in other variables.

The remainder of this paper is organized as follows. Section 2 provides background information on the core HRS and the HRS Internet sample. In section 3, we discuss propensity scoring and our methods to investigate whether propensity scoring is a useful way to correct for selection effects. Section 4 presents empirical results. Section 5 concludes.

2. Background

The University of Michigan Health and Retirement Study (HRS) surveys more than 22,000 Americans over the age of 50 every two years. The study paints a portrait of an aging America's

physical and mental health, health insurance coverage, use of health care, socio-economic status, income, wealth and portfolio choice, labor market position, job characteristics, family networks, and family transfers. It started in 1992 with the 1931-1941 birth cohort (see Juster and Suzman, 1995, for more information on the first wave). In 1993, the first wave of AHEAD took place, with respondents of age 70 and older and their spouses. As of 1998, AHEAD was merged with HRS. Other cohorts were added also, so that the 1998 sample covered the US population aged 51 and older and their spouses. In this project we will use the HRS wave of 2002, covering the 55+ population and their spouses. The first wave of HRS was conducted using computer assisted personal interviews (CAPI). Follow-up surveys are mainly done by telephone, but respondents over 80 years old and households who have no phone are interviewed in person.

In order to use a sample to draw inference on the population of interest, the sample design needs to have certain characteristics. The ideal textbook case is a random sample in which each member of the population is drawn with the same probability, independently of other members of the population. This design is rare in the practice of social science surveys, due to, for example, regional stratification and unit non-response. It then helps to have information on the stratified design and to have at least some information on the unit non-respondents, such as their age and gender, and preferably some indicator of their socio-economic status (possibly derived from the neighborhood where they live), or to have an external source which can be used to determine the size of population segments characterized by age, ethnicity, gender, and perhaps other characteristics. Such information can be used to construct sampling weights for all observations in the sample. Under the assumption that unit non-response (and stratification) is not related to the variables of interest conditional on the information incorporated in the sampling weights, the sampling weights can be used to correct for unit non-response (and stratified sampling) in statistical inference. For example, the population mean of a variable of interest can be consistently estimated by its weighted sample mean.

The HRS uses sampling weights based upon an external source, the March samples of the Current Population Survey (CPS). Weights are constructed first at a household level, using initial sampling probabilities and the birth years and race/ethnicity of the male and female household members, and then at the respondent level (see http://hrsonline.isr.umich.edu/meta/tracker/desc/wghtdoc.pdf for details). Thus the only information used in the weights is birth year, gender, race, ethnicity, and marital status. The analysis in the current paper is at the respondent level and only uses the respondent level sampling weights. It is a maintained assumption that these weights appropriately correct for the non-random nature of the core HRS for all our variables of interest. What we focus on is the other stage: the Internet sub-sample of the core HRS.

Because of the cost effectiveness and other advantages of Internet interviewing, the University of Michigan and RAND set up a pilot project with the overall goal to explore the feasibility of using the Internet to supplement interviewer-administered data collection in the HRS and to explore a variety of methodological issues related to Web-based measurement. With this in mind, the following question about willingness to participate in a future Web survey was added in 2002:

> *"We may want to try out a procedure for asking questions of some of the participants of this study, using the Internet. Would you be willing to consider answering questions on the Internet, if it took about 15 minutes of your time?"*

This question was only asked to respondents who in an earlier question had indicated that they had access to the Internet. The 2002 wave of the HRS contained 16,698 respondents (excluding respondents with zero respondent level weight). Of these, 29.7% reported to have Internet access. Of the group with

Internet access, 73.3% indicated willingness to participate in a Web survey. A random sample of those who indicated willingness was sent a mailed invitation to participate in a Web survey. 78.4% of them completed the Web survey. Overall, this leads to 1,893 respondents who completed the Web survey. Thus several sequential selection processes play a role: Internet access, willingness to participate given access, random selection into the group that gets the letter inviting them to participate, and non-response given willingness to participate. Except for the random selection step, all these steps induce potential selection effects. Couper et al. (2004) look at the several stages in detail. We aggregate the various steps and focus on those of the 16,698 who eventually participated in the Web Survey. We call respondents who volunteered to participate and completed the Web survey "Web responders". Because the survey was designed as a survey of households we analyze data using household level weights. There are 11316 households. In households with more than one respondent we choose the financial respondent; in case there was more than one we choose a financial respondent at random. Throughout we compare estimates based on Web responders only (the small slice in Figure 1), adjusted estimates based on Web responders only (the small slice projected out to the full pie) and adjusted estimates based on non-Web responders (the large slice in Figure 1 projected out to the full pie). Because the two sub-samples need to be disjoint to compare the estimates we do not use estimates corresponding to the full sample. However, the difference between the estimates based on the full sample and the adjusted estimates based on non-web responders is negligible.

<Go to Figure 1 >

Nowhere in this paper do we use the data of the actual Web survey; we only use the information of who responded to the survey. This makes it possible to study selection issues without having to account for potential mode effects – the possibility that answers to the same question may differ depending on whether the question is asked by phone or over the Internet (cf., e.g., Schwarz and Sudman, 1992). Mode effects are certainly another major issue in selecting the mode of interviewing, but are irrelevant for the research questions on correcting for selection addressed in the current paper.

Most Web surveys do not have an underlying sampling frame like the core HRS. Usually, a convenience sample, rather than a random sample or a probability sample is selected. Drawing inferences from convenience samples, including estimates of population frequencies and percentages, is a hard problem, which is often neglected (also see Schonlau et al., 2002). Biostatisticians have long been accustomed to drawing inferences from observational studies because the randomization required for experiments can be unethical when dealing with human subjects or difficult to achieve in practice. Propensity scoring (Rosenbaum and Rubin, 1983, Rosenbaum, 2002) is commonly used to draw inferences from observational data.

Harris Interactive, a commercial Web survey company, has adopted this approach for the use of Web surveys. The Harris approach involves partitioning the propensity score into a categorical variable. This and other variables are then used for post-stratification. The Harris Interactive approach is described in more detail in Schonlau et al. (2004). Application of propensity scores in the context of Web surveys is also described by Danielsson (2004). The central issue is whether and under what circumstances propensity adjusted estimates are comparable to those based on random samples. An integral component of the issue is what questions should be asked to capture the difference between the Web responders and the population of interest. As mentioned, Harris Interactive calls these elusive

questions "webographic" questions, comprising both demographic and lifestyle questions. Other researchers call them "lifestyle" or "attitudinal" questions.

Forsman and Varedian (2004) investigate the efficacy of propensity score weighting in the context of a marketing survey about the use of hygiene products and attitudes toward local banks. A phone survey (N=347) and a Web survey (N=4724) were conducted in a northern European country. Their survey included lifestyle questions that were trying to capture a respondent's "modernity". They use logistic regression on lifestyle questions and demographic questions to capture the selection effect. They conclude that the estimates obtained from Web and RDD phone surveys are different. Further, various different weighting schemes did not change the results very much.

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

For the 2002 wave of the HRS, which did not have the specific webographic questions asked by Harris Interactive, we use demographic questions, health related questions, and others that were available in the 2002 wave of the HRS. We investigate which variables need at a minimum be included in constructing the sampling weights so that the sampling weights can correct for selection bias in all variables of the complete set. The idea behind this is that the variables not included in this minimal set only have to be asked in the Web survey. We will explain this in detail in terms of propensity scores in the next section.

3. Propensity scoring

We first illustrate how propensity scoring works with a simple example. Suppose we are interested in estimating the prevalence of diabetes and would only know diabetes prevalence for the respondents in the Web survey, not for the others in the HRS. The prevalence of diabetes in the sample of Web responders is 11.3%. In the Internet sample, 27.8% of African American respondents have diabetes compared to only 10.7% of others. The proportion of African Americans in the Internet sample is 3.8%, compared to 14.7% in the HRS. Under the assumptions that 1) the HRS is representative for the population and 2) participation in the Web survey may depend on race but, conditional on race, participation in the Web survey is not related to whether or not the respondent has diabetes, these percentages can be used to compute the following adjusted estimate of prevalence of diabetes in the population:

$$0.147 * 0.278 + (1-0.147) * 0.166 = 0.132 \qquad (1.1)$$

This is a weighted average of the diabetes prevalence dummy over the Internet sample, where the weights are 147/38 for African Americans and (1000-147)/(1000-38) for others (before normalizing them so that they add up to one). Thus the weights correct for under-sampling of African Americans in the Web survey. The weights can also be written as the reciprocals of the "propensity scores," the probabilities of Web survey participation for African Americans and others.

The adjusted estimate, 13.2% is larger than the unadjusted estimate, 11.3%. In this example, we actually know prevalence of diabetes for the complete HRS sample, and we can compare the adjusted

Web survey estimate with the direct estimate based upon HRS. The latter is 18.2%. Although the adjustment (from 11.3% to 13.2%) goes in the right direction, it is far from complete. The remaining discrepancy suggests that the assumption that conditioning on race only is enough to get independence of Web survey participation and prevalence of diabetes is not justified. We will need more than just one single conditioning variable to construct the weights. This is typical for most situations – we will usually need much more than a single conditioning variable on for constructing appropriate weights. The correction procedure in terms of propensity scores remains the same – use weights constructed as reciprocals of the propensity scores. The propensity scores can be estimated with parametric or nonparametric methods.

For the general case, let Y be a (vector of) variable(s) of interest (e.g., a set of health conditions), and let X be a set of conditioning variables (e.g., race, gender, age). Let I denote the dummy variable indicating whether the respondent is a Web responder $(I=1)$ or not $(I=0)$. The propensity score is defined as $P(I|X)$. To use the propensity scores in constructing sampling weights, we need the assumption of conditional independence (CI):

Y and I are conditionally independent given X (CI-a)

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

$Y|X,I=1$ has the same distribution as $Y|X,I=0$ for almost all X (CI-b)

Or as

$P(I=1|X,Y) = P(I=1|X)$ for almost all X and Y (CI-c)

Under CI, we can use weights to correct for the fact that $P(I|X)$ depends on X as in the example above. Formally we have, if, for example, we want to estimate the population mean $E(Y)$ of Y:

$$E(Y) = E[E(Y|X)] = E[E(I/P(I=1|X)|X).E(Y|X)] = E[E(I/P(I=1|X).Y|X)] =$$

$$= E[P(I=1)/P(I=1|X).E(Y|X,I=1)] = E[w(X).E(Y|X,I=1)] = E[w(X)Y|I=1]$$

This is a weighted mean over the subpopulation participating in the Web survey, where the weight $w(X)=P(I=1)/P(I=1|X)$ is proportional to the reciprocal of the propensity score (see, for example, Little and Rubin, 2002).

To estimate $E(Y)$, we also need to account for the sampling weights in the HRS. For observation i, this is denoted by w_i^{HRS}. Under CI and the maintained assumption that the HRS sampling weights are appropriate to make HRS representative of the population of interest (cf. Section 2), a consistent estimator for $E(Y)$ will be given by:

$$\frac{1}{N_I}\sum_{i=1}^{N_I} w_i^{HRS}\hat{w}(X_i)Y_i$$

where the summation is over the N_I observations in the Web survey, and where $\hat{w}(X_i)$ is a consistent estimator of the weight, based upon a (non-parametric or parametric) estimator of the propensity score.

The crucial assumption is the assumption of conditional independence. Note that if this holds for a given set of conditioning variables *X*, it will also hold for any larger set. This leads to the idea of selecting a minimal set *X* such that CI holds for a large enough set of *Y* variables of interest. Once such a set *X* is found, it is sufficient to have Web survey observations on *Y* and propensity scores based upon *X*. This is where the potential efficiency gain of Internet surveying is situated. We know that Web survey participation is selective, but if we can find a relatively small set of conditioning variables *X*, we can still use a Web survey to draw population inference on *Y*. All we need is a representative survey measuring *X*, to construct the propensity scores.[1]

The HRS has too many variables to consider all *Y* variables that could potentially be of interest and all *X* variables that could be conditioned upon. Instead, we have pre-selected a number of potential variables with information in various domains (health, economic status, family composition). Variables of potential interest may also be included in the minimal set of *X*, so there is no reason to distinguish a priori between *X* and *Y* variables. In order to find a minimal set *X*, we use (CI-c) and perform a backward stepwise logistic regression procedure, excluding variables that are not significant in explaining Web survey participation (the significance level for removing a variable is 0.05). The variables that remain statistically significant given the other covariates constitute the minimal set. Once the minimal set *X* is selected, propensity scores $\hat{p}_i = \hat{P}(I = 1 \mid X_i)$ follow directly as the logit predictions of Web survey participation, and the weights in the two sub-samples can be constructed as $w_i = w_i^{HRS} / \hat{p}_i$ for Web survey participants and $w_i = w_i^{HRS} / (1 - \hat{p}_i)$ for non Web survey participants.

The stepwise logistic regression has the usual drawbacks of stepwise procedures and results may be sensitive to, for example, the significance level that is used. Moreover, the logistic specification may not be valid. On the other hand, due to the curse of dimensionality, a fully non-parametric model is infeasible. Instead of considering alternative specifications or formally testing whether the logit specification is sufficient for modeling Web survey participation, we directly test whether the propensity scores achieve what they are designed for: constructing weights to correct for selective participation in the Web survey. While such a test is not possible with the usual convenience samples, it is here, since both *Y* and *X* are not only available for the Web survey, but for the complete HRS.

Under the null hypothesis (CI), consistent estimates of features of the population distribution of *Y* based upon the HRS sub-samples of Web responders and non-Web responders should not be significantly different. These consistent estimates can be obtained using the weights described above for the Web responders, and in a similar way for the non-Web responders. The commonly-used Kolmogorov-Smirnov statistic for equality of distributions is based on the maximum difference between two cumulative distribution functions (CDF). To our knowledge this test has not yet been developed for weighted data and can thus not be applied here. The Cramer-von Mises test is based on the integrated squared difference between two CDF's. It could be adapted for use with weights; however, the implementation would require numerical integration. Instead, we test the hypothesis that the two distributions are equal as follows: For the non-Web responders, we divide the empirical distribution of the variable of interest *Y* into 10 deciles (accounting for the weights). For the Web responders, we use the same cutoff values as for the non-Web responders and divide the distribution of the variable of

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

interest into the same 10 categories. Under the null hypothesis, the categories for the Web responders also should each contain approximately 10% of the observations (again, accounting for the weights). We use Pearson's chi-squared test adjusted for use with weights (Rao and Scott, 1981 and 1984) to test independence in the two-way table.

The same test can be applied when the variable of interest is categorical (such as a dummy variable). Instead of using 10 categories we then simply use the number of categories of the categorical variable. For example, for a dummy variable the test is based on a 2 by 2 table (with weights). Throughout we use Stata for the analysis. Pearson's chi squared test for weighted data is implemented in Stata's procedure for survey data "svytab".

4. Results

We ran a backward stepwise logistic regression of the indicator variable *I* whether the respondent participated in the Web survey on a number of variables: demographics (race/ethnicity, gender, education, age), personal income, an indicator of home ownership, self-assessed health, and various measures of difficulties with activities of daily living (ADL). The variables retained in the regression are shown in Table 1.[2] All variables considered except income are dummy variables, making it easy to present the results in terms of comparable odds ratios. Age was transformed into a small number of categorical dummies, allowing for non-linear and non-monotonic effects.

All demographic variables retained in the backward stepwise regression except gender were highly significant. In spite of its low significance level, we retained a gender dummy among the conditioning variables, since this is almost always one of the variables used to construct sampling weights. Furthermore, we decided to include or exclude variables based upon the same underlying survey question as a group. For example, self-assessed health is measured on a five-point scale (from poor to excellent), leading to four dummy variables. The dummies were jointly significant and we therefore included all of them, even though not all of them were individually significant. Remarkably, among the activities of daily living only "difficulty with grocery shopping" was significant.

The odds ratios according to the final logistic regression results are presented in Table 1. The t-values and corresponding p-values refer to the test that the corresponding regression coefficient is zero, i.e., that the odds ratio is equal to 1. As expected, participation in the Web survey (which requires Internet access) falls with age. Non-Hispanic Whites participate at higher rates than Hispanics, African Americans, and other races. Web survey participation rises substantially with education level: someone with less than high school has a probability to participate in the Web survey which is only about one fourth of the probability of someone with high school and the same other characteristics, and about one tenth of the probability of someone with at least a college degree. Large and significant effects are also found for marital status dummies, with the largest Web survey participation probability for married people. The probability to participate in the Web survey also rises significantly with income. The odds ratio on log income[3] implies that multiplying income by 10 increases the odds of participation by 1.66.

---

[2] The variables that were dropped all pertain to activities of daily living: (difficulty with) dressing; walking across room; bathing/showering; eating; getting in/out of bed; using the toilet; preparing hot meals; using the phone; taking medications.

[3] All log transformations in this paper are computed as $\log_{10}(y+\sqrt{1+y^2})$ where y is the variable to be trasnformed. This transformation ensures that the argument of the log is never negative.

Almost 12% in the HRS report having no personal income; these are mainly spouses of main earners or former main earners. The estimates indicate that the probability for these people is similar to that of people with an income of about 2650 US dollars, below the 20th percentile of the income distribution. Finally, a strong negative effect of health problems is found. People who report that they are in fair or poor health have about half the probability to participate in the Web survey of otherwise similar people in excellent health.

<Go to Table 1>

Table 2 gives the estimates of the population means of a number of health related variables. The first column is based on the non-Web responders reweighted to make this sample representative of the population. (We are presenting adjusted results of non-Web responders rather than results of the entire population because the statistical tests require that the sample of Web responders must not overlap with the second sample. Most respondents are non-Web responders. The adjustment is minor.) The second column is based upon Web responders only, not correcting for selectivity of Web survey participation. The fifth column has the propensity-score adjusted estimates for Web responders. The variables are grouped into comorbidities, mental health variables, and activities of daily living (ADL). All of these variables are binary indicator variables. This makes it easy to present results in a single table. The bar chart in Figure 2 displays the same estimates graphically for the comorbidities and the mental health variables.

Table 2 also presents the differences between the estimates based on non-Web responders and Web responders, not adjusting (third column) and adjusting (sixth column) with propensity scores, and the p-values of the corresponding Pearson's chi-squared test discussed in the previous section, separately for each of the health variables considered.

<Go to Table 2>

<Go to Figure 2>

The unadjusted estimates based upon Web responders in the second column are almost always significantly different from those of the adjusted non-web responders in the first column. The exceptions are the onset of cancer or psychological problems and reporting to be happy. This shows that some correction is necessary to adjust for selectivity of Internet access. In all cases where the difference is significant, the difference in prevalence rates suggests that Web responders suffer less from the health problem considered than the rest of the 50+ population. Web responders have lower prevalence of chronic diseases, fewer symptoms of mental health problems, and fewer limitations in their activities of daily living. This is in line with the results in Table 1, where we saw that poor or fair self-assessed health (as well as difficulties with grocery shopping) greatly reduced the Web survey participation probability.

The propensity score adjusted estimates of the means in column 5 are typically substantially higher and much closer to those of the non-Web responders. All differences except one are 4%-points or less. The only exception is the variable indicating psychological or emotional problems, where we find that correcting for selective Internet access leads to an overestimate of the prevalence rate. In general, however, we can conclude that the propensity score based corrections work for a broad range of health related variables. This implies that, once selectivity through the set of variables in Table 1 (including only a few health variables) is controlled for, other health variables can be studies using the Internet sample only. In this case the propensity scores do the job for which they are constructed.

Note that not only the size of the difference changes due to the use of the propensity weights, but also the level of significance if the difference is kept constant. A difference of two percentage points with the unadjusted estimates may be significant whereas that same difference may not be significant when using weights. This is because the unequal selection probabilities increase the probability design effect (Kish, 1965). The probability design effect, DE_p, can be computed for the Web survey responders as

$$DE_p = \frac{N_I \sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{N_I} w_i \right)^2}$$

The design effect for web responders is much larger if propensity score adjustments are used than if only the HRS sampling weights are used (6.37 versus 1.41),[4] suggesting that selective participation in the Web survey leads to a substantial reduction in effective sample size for the purpose of making statistical inference (a loss of precision in estimation and a reduction of power for testing).

Table 3 does the same thing as Table 2 for a different set of variables, related to health behavior and the amount and composition of household wealth. The conclusions are different from those in Table 2. The only variable in Table 3 for which the propensity score adjustment really "works" is the dummy for whether the respondent performs vigorous activity or not. The unadjusted Web-based estimate of the mean of this variable is much too high, and adjusting with the propensity score reduces it substantially, making it close to insignificantly different from the sample mean of the non-Web responders. The dummy "smoke now" also gives insignificant differences after adjustment, but here the difference was already insignificant before propensity scores were used. For the dummy "smoke ever," the adjustment goes in the wrong direction, increasing the difference between the estimates of the population mean based on Web-responders based and non-Web responders. In all other cases, including all four asset variables considered, the propensity score adjustment reduces the gap but is insufficient to make the difference insignificant. For example, ownership of stocks is 35% in the non-Web sample, compared to 52% before correction and 42% after correction in the Web sample. Thus the propensity weights correction helps but not enough.

To gain some further insight into what causes the differences in assets across RDD and Web responders, Figure 3 illustrates the differences on between the distributions of assets across Web and non-Web responders. The distribution of log amounts held in checking accounts in the non Web responders has a large dispersion with, in particular, many small values. The unadjusted estimated based upon the Web responders only has a much smaller dispersion and a higher mean (the right hand figure). The adjusted estimate in the middle panel is in between the two in both respects. Again, this suggests that the adjustment helps (i.e., goes in the right direction) but is incomplete. The propensity scores do a good job in improving the estimate of the upper tail of the distribution, but not the lower tail. The distribution of the stocks includes a large spike at zero (households without stock) and is therefore different from the distribution of the checking accounts. Nonetheless, a similar conclusion is obtained for stocks. Web responders hold larger amounts on average, and, in particular, less often hold small amounts than would be predicted on the basis of income, home ownership, self-assessed health and the other variables in the propensity scores.

[4] The design effect for the complete HRS sample (using the HRS sampling weights) is 1.42.

All this leads to the conclusion that, even controlling for socio-economic status through income and education variables, households with Internet access more often hold stocks than households without Internet access. It implies that collecting data on asset ownership for the Internet sample and adjusting the estimates using propensity scores (based upon the small set considered here and not incorporating asset ownership information) is not sufficient to analyze asset ownership in the population of interest. A similar conclusion applies to health related behavior: the limited set of conditioning variables used to build the propensity scores is not enough to control for differences in all health related behavior between Web users and non users. One possible explanation of why the adjustment did not work for some variables is that variables for which the adjustment did not work all related to events in the past ("ever had psychological problems", "ever smoke", "ever drink alcohol"). . Other variables such as "ever had a stroke" or "ever had a heart attack" also occurred in the past but arguably greatly affect the present.

< Go to Figure 2 >

5. Discussion

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

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

Starting with a limited subset of all the variables, we performed a stepwise logistic regression explaining Web survey participation, to determine which variables maybe left out when constructing the weights. We are able to drop a number, but not very many variables. We also found that the a priori selected set of variables was not sufficient to correct for selectivity in all variables of interest. For example, we corrected for selection on the basis of socio-economic status by using education level as well as personal income to construct the weights. In spite of this correction, we still found that ownership of shares of stock or stock mutual fund is substantially overestimated when using Web responders only. The implication is that Web survey information on ownership of stocks is not enough to estimate the ownership rate in the population of interest, even in the presence of a representative survey of other socio-economic variables. This conclusion differs from that of Berrens et al. (2003) who find that the correction using propensity scores based upon "webographic" questions works well for analyzing political variables. We find that the corrections generally work well for health variables, but not for past health behavior (smoking and drinking) or, particularly, financial assets.

Thus our results cast some doubt in general on the ability to correct for selectivity on the basis of a small set of basic variables. Unfortunately, the HRS does not contain these so-called webographic

variables used to construct the weights in several recent Internet convenience samples, but it would be interesting to check their performance in the same way.

If propensity scores cannot be used to construct for selectivity in the distribution of the variables of interest, this underlines the necessity of getting broader coverage of Internet surveys or the continued search for suitable webographic variables. Perhaps broader coverage will happen automatically over the next ten years, given the speed with which Internet access has spread in recent years. Particularly for elderly cohorts, however, alternatives may still be necessary. One obvious solution is to provide non-Internet users access to the Internet by giving them the necessary equipment. A prominent example is the CentERpanel collected by CentERdata in the Netherlands (http://www.uvt.nl/centerdata/en/). Another example is Knowledge Networks (http://www.knowledgenetworks.com/). Both companies provide a so-called set-top box (or Web-TV) to households without Internet access that can be used to connect to the Internet, using a TV-set as a monitor. (A TV set is provided as well if necessary). Although this does not alleviate other common problems like unit non-response and panel attrition, the approach provides much broader coverage (at potentially much higher cost) and much better chances of appropriately correcting using propensity weights based upon a few basic variables.

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

Acknowledgement

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

Berrens, R.P., A.K. Bohara, H. Jenkins-Smith, C. Silva, and D.L. Weimer (2003), The advent of Internet surveys for political research: A comparison of telephone and Internet samples, *Political Analysis*, 11(1), 1-22.

Birnbaum, M.H. (2004), Human research and data collection via the Internet, *Annual Review of Psychology*, 55, 803-832.

Blumberg, S.J., J.V. Luke, and M.L. Cynamon (2004), Has cord-cutting cut into random-digit-dialed health surveys? The prevalence and impact of wireless substitution. In S.B. Cohen and J.M. Lepkowski (eds.), *Eighth Conference on Health Survey Research Methods*. Hyattsville, MD: National Center for Health Statistics, pp. 137-142.

Cochran W.G. (1977), *Sampling Techniques* (3rd ed), New York: John Wiley & Sons.

Couper, M.P. (2000), Web surveys. A review of issues and approaches, *Public Opinion Quarterly*, 64, 464-494.

Couper M.P., Kapteyn A, Schonlau M, Winter J. (2004), Noncoverage and nonresponse in an Internet survey. In *Proceedings of the International Conference on Social Science Methodology* (RC33), Amsterdam, August 2004.

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

De Heer, W. (1999), International response trends: results of an international survey, *Journal of Official Statistics,* 15(2), 129-142.

Groves, R.M. and M.P. Couper (1998), *Nonresponse in Household Interview Surveys,* New York: John Wiley & Sons.

Juster FT, Suzman R. (1995), An overview of the Health and Retirement Study, *The Journal of Human Resources*, 30, Special Issue on the Health and Retirement Study: Data Quality and Early Results, S7-S56.

Kish L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

Little, R.A. and D.B. Rubin (2002), *Statistical analysis with missing data,* New York: John Wiley & Sons.

Link, M.W., and R.W. Oldendick (1999), Call Screening: Is it really a problem for survey research?, *Public Opinion Quarterly,* 63, 577-589.

Oldendick, R.W., and M.W. Link (1994), The answering machine generation: who are they and what problem do they pose for survey research?, *Public Opinion Quarterly,* 58, 264-273.

Piekarski, L., G. Kaplan, J. Prestegaard (1999), Telephony and telephone sampling, paper presented at the Annual Conference of the American Association for Public Opinion Research, St. Petersburg, FL.

Rao J.N.K. and A.J. Scott (1981), The analysis of Categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76(374), 221-230.

Rao J.N.K. and A.J. Scott (1984), On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data, *Annals of Statistics,* 12, 46-60.

Rosenbaum P.R. (2002), *Observational Studies,* 2nd ed. New York: Springer-Verlag.

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

Rubin D.B. and N. Thomas N. (1996), Matching using estimated propensity scores: relating theory to practice, *Biometrics*, 52, 249-64.

Schonlau, M., R. Fricker and M. Elliott (2002). *Conducting Research Surveys via Email and the Web*, Santa Monica, CA: RAND.

Schonlau M, Zapert K, Payne Simon L, Sanstad K, Marcus S, Adams J, Spranca M, Kan H-J, Turner R, Berry S. (2004) A comparison between a propensity weighted web survey and an identical RDD survey, *Social Science Computer Review,* 22(1):128-138.

Schonlau M, B.J. Asch BJ, and C. Du (2003), Web surveys as part of a mixed mode strategy for populations that cannot be contacted by e-mail. *Social Science Computer Review,* 21(2): 218-222.

Schwarz, N., and S. Sudman (Eds.) (1992), *Context Effects in Social and Psychological Research*. New York, NY: Springer Verlag.

Taylor, H. (2000), Does Internet research "work?": comparing on-line survey results with telephone surveys, *Journal of Marketing Research,* 42(1), 51-64.

Varedian M and G. Forsman (2003), Comparing propensity score weighting with other weighting methods: A case study on Web data" In *Proceedings of the Section on Survey Statistics*, American Statistical Association; 2003, CD-ROM.

Czaja, S.J., and J. Shari (1998), Age differences in attitudes toward computers, *Journal of Gerontology* 53B, 329-340.

Morrell, R.W., C.B. Mayhorn, and J. Bennett (2000), A survey of World Wide Web use in middle-aged and older Adults, *Human Factors,* 42, 175-182.

Table 1

		Odds Ratio	t	p
Race /Ethnicity	White	1.00	NA	NA
	African American	0.30	-8.38	0.00
	Other Race	0.42	-3.06	0.00
	Hispanic	0.29	-5.53	0.00
Gender	Female	1.00		
	Male	0.99	-0.09	0.93
Education	<high school	0.28	-7.46	0.00
	high school	1.00		
	some college	1.71	6.44	0.00
	>= college	2.63	11.93	0.00
Age	<55	1.18	1.59	0.11
	55-65	1.00		
	65-75	0.69	-4.85	0.00
	>75	0.25	-11.60	0.00
Marital Status	Married	1.00		
	separated, divorced, widowed	0.64	-5.60	0.00
	never married	0.60	-2.75	0.01
Income	log 10 income	1.66	5.59	0.00
	Indicator (Income==0)	5.93	4.31	0.00
Self Assessed	Excellent	1.15	1.59	0.11
Health	Very Good	1.00		
	Good	0.81	-2.66	0.01
	Fair	0.59	-4.57	0.00
	Poor	0.66	-2.06	0.04
Difficulty with …	Grocery Shopping	0.50	-2.97	0.00
House Owner	Home Owner	1.30	2.43	0.02

Table 1: Logistic regression of participation in the Web survey on various covariates.

Table 2

		Adjusted	Unadjusted			Adjusted		
		Non Web Responders	Web Responders	p	diff	Web Responders	p	diff
Comorbidites	High Blood pressure	52%	44%	0.00	-0.03	49%	0.27	-0.03
	Lung disease	10%	7%	0.01	-0.03	10%	0.90	0.00
	Heart disease	24%	16%	0.00	-0.08	20%	0.21	-0.03
	Stroke	7%	4%	0.00	-0.03	6%	0.38	-0.02
	Cancer	13%	12%	0.40	-0.01	14%	0.65	0.01
	Diabetes	16%	12%	0.00	-0.04	18%	0.38	0.02
	Arthritis	59%	47%	0.00	-0.11	63%	0.22	0.04
	Ever had psych problems	16%	16%	0.97	0.00	22%	0.04	0.07
Mental Health	Depressed	18%	9%	0.00	-0.08	15%	0.60	-0.02
	Lonely	20%	11%	0.00	-0.09	20%	0.87	0.01
	Happy	87%	89%	0.09	0.02	87%	0.64	0.01
	Sad	22%	15%	0.00	-0.07	24%	0.66	0.02
	Effort	25%	12%	0.00	-0.13	25%	0.96	0.00
	Sleep was restless	28%	24%	0.00	-0.04	30%	0.66	0.02
	Could not get going	23%	15%	0.00	-0.08	26%	0.43	0.03
Activities of	.. Dressing	8%	3%	0.00	-0.05	11%	0.37	0.04
Daily living	..Walking across Room	6%	2%	0.00	-0.04	5%	0.64	-0.01
(Difficulty with ..)	.. Bathing/Showering	5%	1%	0.00	-0.04	3%	0.09	-0.02
	.. Eating	2%	0%	0.00	-0.02	1%	0.06	-0.01
	.. Getting in/out Bed	5%	2%	0.00	-0.03	3%	0.10	-0.02
	.. Using the Toilet	5%	2%	0.00	-0.03	3%	0.37	-0.01
	.. Preparing hot meals	5%	1%	0.00	-0.04	3%	0.50	-0.01
	.. Grocery shopping	8%	2%	0.00	-0.06	7%	0.44	-0.02
	.. Using the phone	2%	1%	0.00	-0.02	1%	0.04	-0.01
	.. Taking medications	2%	1%	0.01	-0.01	1%	0.17	-0.01
	Managing money	4%	2%	0.00	-0.03	2%	0.01	-0.02
	Walking several blocks	29%	13%	0.00	-0.15	28%	0.83	-0.01
	Walking one block	13%	4%	0.00	-0.09	15%	0.75	0.01
	Sitting for 2 hours	20%	13%	0.00	-0.07	23%	0.45	0.03
	Getting up from chair	39%	28%	0.00	-0.11	43%	0.39	0.04
	Climbing sev. Flt stairs	44%	30%	0.00	-0.15	44%	0.89	0.00
	Climbing one flt stairs	16%	7%	0.00	-0.09	16%	0.98	0.00

Table 2: Differences in prevalence of comorbidities, symptoms of mental health problems, and limitations in activities of daily living. Unadjusted estimates refer to those that only use sampling weights. Adjusted estimates refer to those that in addition use propensity weights. Grey shaded areas indicate significance at 5%. The ten activities of daily living above the double line were included in the backward stepwise regression, those below were not. The only activity of daily living retained in the model is in italics.

Table 3

		adjusted	unadjusted			adjusted		
		Non Web responders	Web responders	p	diff	Web Responders	p	diff
Health behavior	R smoke ever	57%	61%	0.01	0.04	65%	0.02	0.08
	R smoke now	15%	13%	0.15	-0.02	12%	0.08	-0.03
	R ever drink alcohol	47%	67%	0.00	0.19	61%	0.00	0.13
	Vig phys activity 3+ wk	42%	51%	0.00	0.10	43%	0.79	0.01
Assets	Has checking account	88%	97%	0.00	0.09	96%	0.00	0.08
	Owns stock	35%	52%	0.00	0.17	42%	0.04	0.08
	Assets Stock (log)	1.60	2.58	0.00	0.99	2.07	0.01	0.47
	Assets Checking (log)	3.56	4.15	0.00	0.59	3.79	0.02	0.23

Table 3: Differences in health related behavior and ownership and amounts held of financial assets. Unadjusted estimates refer to those that only use sampling weights. Percentages refer to the sample means of dummy variables. Adjusted estimates refer to those that in addition use propensity weights. Grey shaded areas indicate significance at 5%.

Figure 1

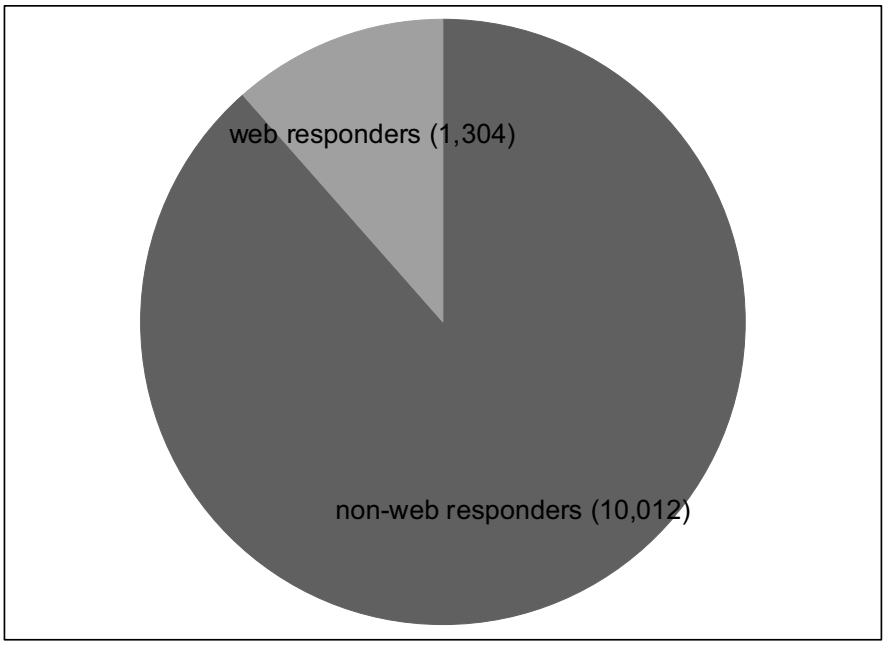

Figure 1: Web responding households and non-Web responding households together form a random sample. The adjusted estimates based on non-Web responders adjust for the missing slice in the pie; the adjustment is minor. The adjusted estimates based on Web responders use the slice to project to the entire pie; clearly a much harder task.

Figure 2

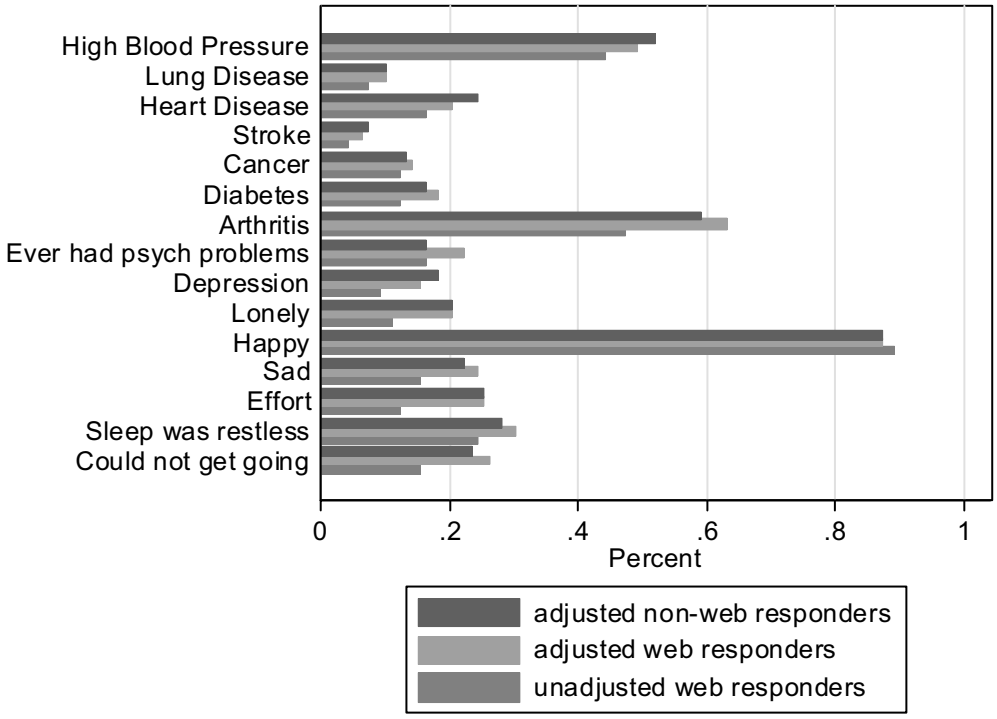

Figure 2: Prevalence of physical health conditions and mental health indicators. The estimate for non-Web responders and the unadjusted estimate for Web responders use only the sampling weight. The estimates of adjusted Web responders uses both sampling and propensity weights.

Figure 3

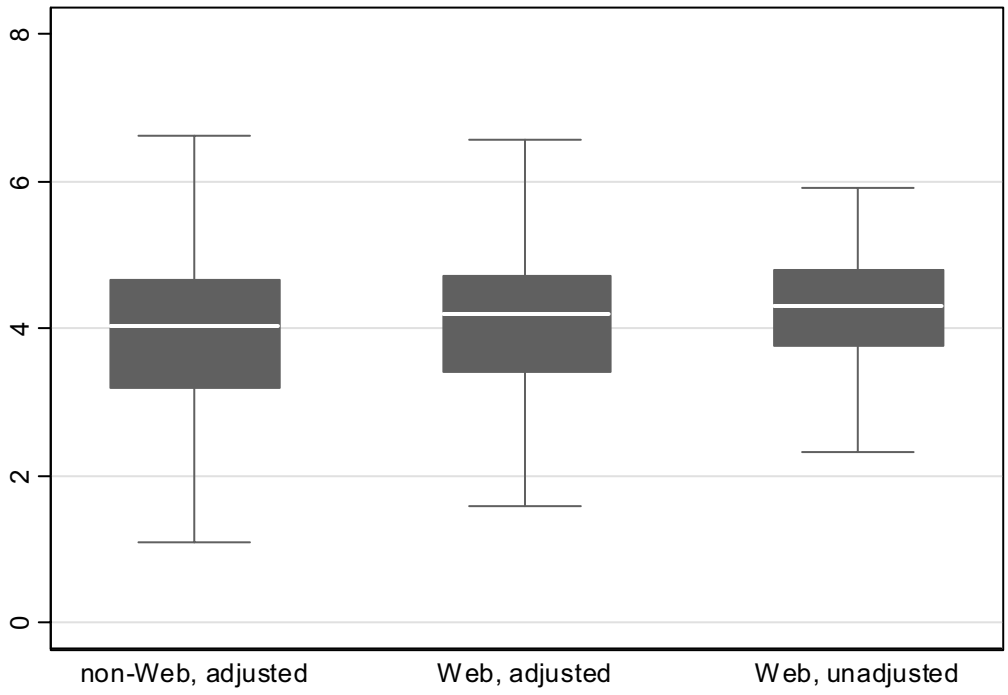

Figure 3: Parallel box plots for log assets (checking account). The adjustment works in the right direction and also expands the range of distribution.