A COMPARISON BETWEEN USING THE WEB AND USING TELEPHONE TO SURVEY POLITICAL OPINIONS

Annica Isaksson, Linköping University, Sweden
Gösta Forsman, Swedish National Road Administration and Linköping University, Sweden

KEY WORDS: Web surveys, propensity score adjustment, political opinion poll.

1. Introduction

In Sweden, as in many other countries, the telephone is the predominant mode for surveying opinions among the general population. In recent years, however, the World Wide Web has become a viable alternative for the survey industry. Among the appealing features of Web surveys are low data collection costs and a large flexibility in the design of the questionnaire. Web surveys, however, also suffer from several methodological problems, including dubious representativity due to low Internet penetration in the general population, lack of sampling frames and high nonresponse rates. Data are often collected from easily recruited nonprobability samples, or convenience samples, of Web users. Inference from such samples relies heavily on appropriate weighting methods to reduce selection bias.

Weighting adjustment methods are often based on the assumption that a population can be divided into groups (or *weighting classes*) within which the respondents characteristics are "representative" for the entire group. If the assumption holds, the selection bias may be completely eliminated. This is, however, an optimistic – or even naive – assumption, when inference is made from nonprobability samples. A reasonable motivation for using weighting class adjustment for nonprobability samples is that the selection bias may be somewhat reduced rather than completely eliminated.

An important key to a successful weighting adjustment is the identification of appropriate weighting classes. We compare two procedures that rely on different sources of information for weighting class construction: *poststratification* and *propensity score adjustment*. In poststratification, the weighting classes are defined and limited by available external data, e.g., the distribution of the population over one or more variables, such as population distribution by age, sex and race available from standard population estimates. In propensity score adjustment, on the other hand, the weighting classes are defined by a number of covariates, observed on both the nonprobability (Web) sample and a parallel probability (telephone) sample.

In this paper, we present results from a study conducted by the Swedish commercial survey institute TEMO in a political opinion poll before the Swedish September 2002 election.

2. Propensity Score Adjustment

Propensity score adjustment was originally developed to reduce selection bias in observational studies (quasi-random experiments), see, e.g., Rosenbaum and Rubin (1983, 1984), D'Agostino and Rubin (2000). Typical for such studies is the comparison of treatment effects between two subpopulations, such as smokers and non-smokers. Propensity score weighting for survey nonresponse adjustment has been proposed by Little (1986) and Rubin and Little (1987). The application of the method on Web surveys utilizes a "control" survey based on a traditional probability sampling method (e.g., a telephone survey based on Random Digit Dialing, RDD) that run parallel to a Web survey, based on a nonprobability sample. In the original applications, participation in the telephone and Web surveys then corresponds to "treatments." The United States company Harris Interactive has promoted this method, see Terhanian et al. (2001). The use of propensity score weighting in the survey field, including nonresponse adjustment as well as nonprobability sampling, is further discussed in Danielsson (2002).

2.1 The Control Survey

Propensity score weighting relies on the existence of a control survey, i.e., a separate survey based on a representative (probability) sample and conducted successfully in a way that population parameters and distributions can be estimated unbiasedly (at least approximately). The control survey is used to adjust the Web survey data to the population level. Or, more precisely, the control survey is used to adjust the Web survey data to the control survey level, which is assumed close to the population level. Therefore, the quality of the control survey is critical for the method.

In practise, the control survey is often conducted by telephone for an RDD sample, see Terhanian et al. (2001). In the United States, RDD telephone surveys are the dominant type of probability (or approximately probability) sample surveys in market research. Any mode that permits approximately unbiased estimates of population parameters and distributions may, however, be used for the control survey.

The theory assumes that the control survey and the Web survey are conducted at the same time. For

cost reasons, however, the control survey may be conducted at certain time intervals, e.g., each month, while the Web survey is conducted daily. When the data collection is conducted, the Web sample and the control sample are merged and the propensity score is defined on the merged sample.

2.2 The Covariates

The covariates (here assumed to be categorical) are used to partition the merged sample into groups. Within these groups, the Web sample respondents and telephone (control) sample respondents are – ideally – assumed to have identical distributions of the target variable y. Thus, the choice of covariates (sometimes called "webographics") is critical for propensity score weighting. In particular, the procedure is sensitive to any kind of mode effects when the covariates are measured on the Web and the telephone samples. Typically, the "webographic" questions touch issues such as lifestyle, attitudes and self-perception.

2.3 The Propensity Score and the Forming of Weighting Classes

In each of the groups defined by the covariates, the propensity score is defined as the proportion Web sample respondents of all respondents in the group. Formally, for a given individual k, let Z_k take the value 1 if he participates in the Web survey, zero if he participates in the telephone survey. Further, let $\boldsymbol{X}$ be a vector of covariates, possibly associated with Z, available for both the control and the Web survey. The vector of covariates for individual k is $\boldsymbol{X}_k$. The main problem is to estimate the expected value π_k of Z_k: the propensity score. The prevalent strategy is to formulate a logistic regression model (see, e.g., Neter et al., 1996, eq. (14.37)) for π_k as function of $\boldsymbol{X}_k$. Under such a model, it is straightforward to estimate π_k. Finally, groups with similar estimated scores are collapsed, thus forming a few (usually five) weighting classes.

Let n_h denote the total number of responses (by telephone or Web) within weighting class h. The total number of Web responses within the class is denoted by n_{Wh}. In the estimation, within class h, the Web responses are weighted (multiplied) by n_h/n_{Wh}.

3. Poststratification

A variation of stratification, poststratification is treated in most standard textbooks on survey sampling; see, for instance, Särndal et al. (1992, ch. 7) or Cochran (1977, sec. 5A.9). As the name suggests, the idea is simply to stratify the sample *after* it has been selected, instead of before. Originally, poststratification was used simply as an alternative to stratification and for the same reasons (such as improved precision in the estimates) - not to adjust for missing data. Common reasons for not stratifying beforehand then include that practical considerations favour some simpler design, or that the stratum identity only can be established for sampled individuals.

The case when poststratification is used as a means to adjust for nonresponse bias is treated, e.g., by Särndal et al. (1992, sec. 15.6). The poststrata are usually formed to agree with response homogeneity groups. To separate the different purposes of poststratification, Kalton and Kasprzyk (1986) suggest the name "population weighting" for this application. Technically, poststratifiction, or population weighting, can also be applied to nonprobability samples to adjust for selection bias.

4. Research Questions, Data Set, and Study Design

In our study, two modes of data collection, telephone and Web, were compared, as well as the weighting procedures poststratification and propensity score adjustment. The telephone sample was the TEMO omnibus sample; an approximate probability sample of phone numbers (households) combined with a probability sample of one individual within each household. The phone number sample is a list-based sample of phone numbers according to the "Plus one" sampling technique (Lepkowski, 1988, Forsman and Danielsson, 1997). Phone number non-contacts are replaced by non-contacts from earlier waves of the omnibus survey: a technique similar to one described by Kish and Hess (1959). The selection of an individual within household is conducted according to the "last-birthday" method, which involves identifying the person in the household who had the last birthday among all eligible household members. (For details on the "last-birthday" method, and comparisons with other procedures for selection within household, see, e.g., Binson et al., 2000, Forsman,1993 or Oldendick et al., 1988).

For the Web data collection, a stratified sample of the TEMO Web panel is used. Although this panel originally was recruited from earlier waves of the TEMO omnibus, since the dropout rate is very high, we regard it as a nonprobability sample of Web users. More precisely, we treat it as a quota sample rather than a stratified probability sample with very low response rate.

The data set was collected in a political opinion poll before the Swedish September 2002 election. The telephone sample included responses from 1001 individuals; the Web sample nearly thrice as many (2921 individuals.) We consider the problem of estimating the election outcome from these data. This implies estimating party sympathies within the 'voter population': the group of Swedes who, on Election Day, are entitled to vote (that is, are at least 18 years of age and Swedish nationals), use this right, and return valid ballot-papers. More precisely, we want to estimate the proportion of this

group that will vote for each political party. Our poststrata are formed the way TEMO usually does it; that is, by sex, age class, and dwelling (big-city-dweller or not). The propensity scores, on the other hand, are estimated from a set of "webographic" questions (see appendix.) We also try using both the lifestyle questions *and* the poststratification variables when estimating the propensity scores.

The estimation task is made slightly more difficult by the fact that the exact number of voters is unknown until the election has taken place. We handle this lack of information by estimating the size of the voter population from the sample data. Our approach, which corresponds to treating a population proportion as the ratio of two unknown population totals, is well established in the statistical literature (see, e.g., Särndal et al., 1992, result 5.8.1).

Before we present our results, for reference, let us take a brief look at how the samples are composed. We restrict our attention to respondents that belong to the voter population. The proportion of male respondents is 54.0 per cent in the Web sample, 50.0 per cent in the telephone sample. The sample distributions over ages are presented in fig. 1, the sample distributions over educational levels in fig. 2. From fig. 1, the Web sample has a peak in its age distribution in the age span 50-60 years that is not present in the telephone sample. From fig. 2, the average Web respondent seems to be better educated than the average telephone respondent is.

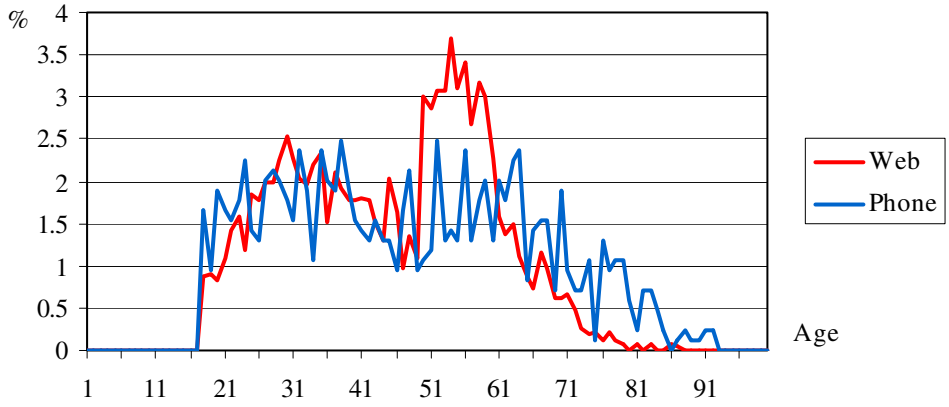

Figure 1: Sample distributions over ages (respondents belonging to the voter population only.)

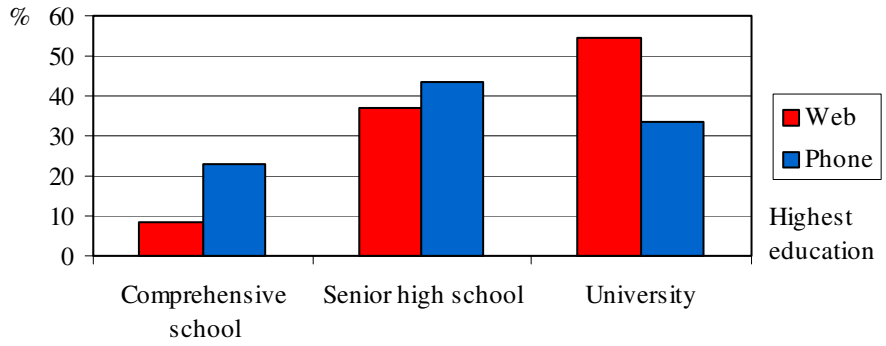

Figure 2: Sample distributions over educational levels (respondents belonging to the voter population only.)

5. Results

Our estimates based on the telephone data are presented in fig 3; our estimates based on the Web data in fig. 4. For the Web data, one possible approach is to use only the lifestyle questions to estimate the propensity scores. In fig. 4, the columns representing the resulting estimates of party sympathies are labelled "Propensity score 1." When both lifestyle questions and poststratification variables are used to estimate the propensities, we label the columns "Propensity score 2." In both fig. 3 and 4, the numbers on the horizontal axes represent the Swedish political parties (a translation of the numbers is provided in tab. 1.)

For comparison, the election outcome, as well as the unweighted sample proportions, are included in both figures. An individual's party sympathy is not necessarily stable over time. Even though the

sample data were collected close upon Election Day, the time lag is still expected to create some unavoidable discrepancies between the sample-based estimates of the election outcome, and the actual outcome.

It is interesting to compare the estimates based on the Web data not only with the election outcome, but also with the poststratified estimates based on the telephone data. To facilitate this, the poststratified telephone estimates from tab 3 reappear in fig 4.

In fig. 3, the poststratified column heights differ only marginally from the unweighted counterparts. Fig. 4 shows some promising results for the propensity score weighting technique: the propensity score estimates typically come closer to the election outcome (as well as to the poststratified telephone estimates) than does the sample proportions and the poststratified estimates. The differences between the two variations of propensity score estimates are quite small.

No.	Party name
1	Conservative
2	Liberal
3	Center
4	Christian Democrat
5	Social Democratic
6	Left Party
7	Green
8	Other

Table 1: Numbering of the Swedish political parties.

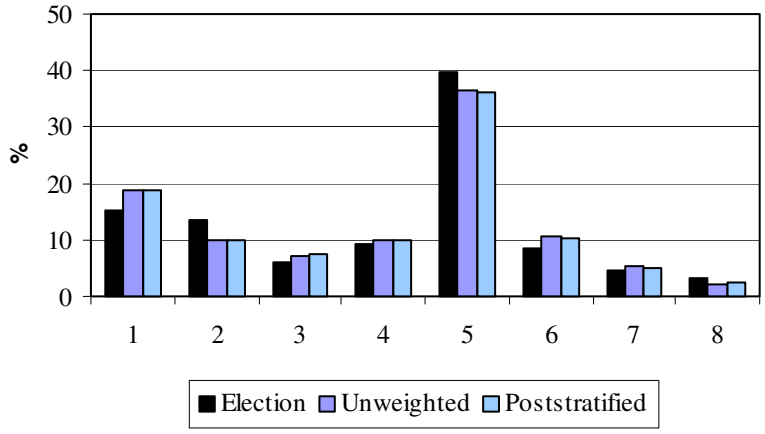

Figure 3: Poststratified estimates of election outcome, based on the telephone data (by political party.)

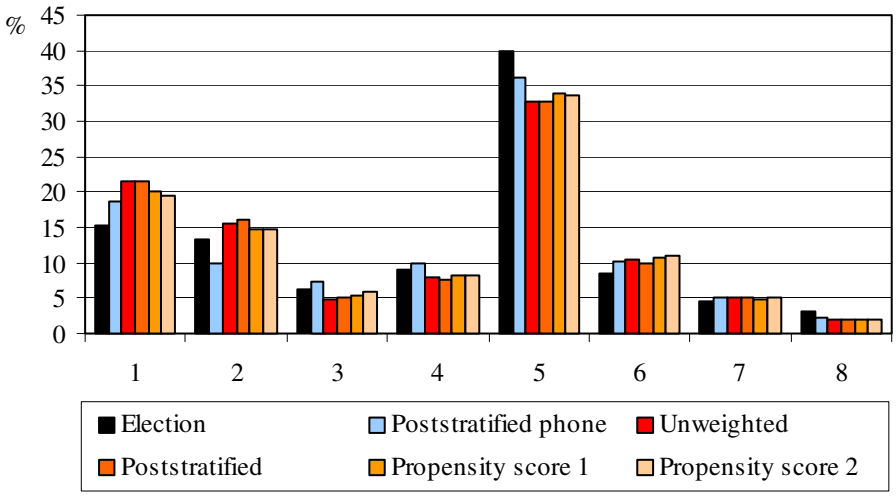

Figure 4: Poststratified, and propensity score weighted, estimates of election outcome, based on the Web data (by political party.)

As an aid in judging the accuracy of the Web estimates, we present a number of sums in tab. 2. In the table, separately for each type of estimates, we give (1) the sum of absolute differences between the Web estimates and the election outcome, and (2) the corresponding sum of absolute differences

between the Web estimates and the poststratified estimates based on the telephone data.

	Election outcome	Poststratified estimates from telephone data
Unweighted	21.51	16.95
Poststratified	21.60	17.87
Propensity score 1	17.45	13.27
Propensity score 2	17.42	12.79

Table 2: Sums of absolute differences.

6. Discussion

The findings from our study show, not surprisingly, that unweighted estimates, based on the TEMO Web panel, differ from those based on the telephone sample. The poststratification does not improve the Web estimates appreciably, whereas the propensity score adjustment at least seems to make some difference to the better. This result holds no matter if the Web estimates are compared with the poststratified estimates based on the telephone data, or with the election outcome. More research is, however, needed before we understand the usefulness of propensity score weighting. Among the research issues yet to be investigated, we identify the choice of covariates and the choice of size of the Web panel sample.

The *choice of covariates* is critical for propensity score weighting. In our study, we used covariates not particularly tested for Swedish conditions. There is a strong need for a systematic evaluation of different kinds of covariates. According to the theory of propensity score weighting, the covariates are assumed to partition the sample into groups within which Web respondents and telephone respondents have identical distributions of the target variable. We agree to the suspicion, discussed by Stenbjerre (2002), that such covariates, if they exist, may be country-specific. For example, in the Scandinavian countries, with homogeneous populations and generally high Internet penetrations, it may be harder to find discriminating covariates than in countries where a considerable part of the population has a low Internet penetration.

In our study, the *size of the Web panel sample* was 2,921. In studies reported from the United States, sample sizes are often much higher (typically well over 10,000). Since the number of cells defined by the covariates is large, and large sample sizes in the cells produce better estimates of the propensity scores, the weighting should be more efficient with larger samples. Experiments with larger samples should be possible as the TEMO Web panel increases.

Finally, there are of course other weighting adjustment methods than those used in our study. For example, the propensity score weighting procedure applied by Harris Interactive differ somewhat from the procedure used in this study. Harris Interactive uses a raking procedure (see, e.g., Oh and Scheuren, 1983) with the telephone sample distribution of the weighting classes and population distributions of background variables (such as age and gender) as marginal auxiliary information. We have not used this procedure in this study but we find it likely that the resulting estimates resembles those based on our procedure ”propensity score 2,” since the same information is used.

7. Acknowledgement

The financial support of this work by the Bank of Sweden Tercentenary Foundation (Grant no. 2000-5063) is gratefully acknowledged.

8. References

Binson, D., Canchola, J.A. and Catania, J.A. (2000). Random Selection in a National Telephone Survey: A Comparison of the Kish, Next-Birthday, and Last-Birthday Methods. *Journal of Official Statistics*, 16, 53-59.

Cochran, W.G. (1977). Sampling Techniques, 3rd edition. *New York: Wiley.*

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

D'Agostino, R.B., Jr and Rubin, D.B. (2000). Estimating and Using Propensity Scores With Partially Missing Data. *Journal of the American Statistical Association*, 95, 749-759.

Danielsson, S. (2002). The Propensity Score and Estimation in Nonrandom Surveys - an Overview. *Research Report, Department of Statistics, Linköping University.*

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

Forsman, G. (1993). Sampling Individuals Within Households in Telephone Surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Forsman, G. and Danielsson, S. (1997). Can Plus Digit Sampling Generate a Probability Sample? *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, 1, 1-16.

Kish, L. and Hess, I. (1959). A Replacement Procedure for Reducing the Bias of Nonresponse. *The American Statistician, 13, 4, 17-19.*

Lepkowski, J.M. (1988). Telephone Sampling Methods in the United States. In *Telephone Survey Methodology*, edited by R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg, 73-99. New York, Wiley.

Little, R. (1986). Survey Nonresponse Adjustments for Estimation of Means. *International Statistical Review*, 54, 139-157.

Little, R. and Rubin, D. (1987). Statistical Analysis with Missing Data. New York: John Wiley and Sons.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). Applied Linear Statistical Models. *4th ed. Chicago: Irwin.*

Oh, H.L. and Scheuren, F. (1983). Weighting Adjustment for Unit Nonresponse. In *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies* (Eds. W.G. Madow, I.Olkin and D.B. Rubin), New York: Academic Press.

Oldendick, R.W., Bishop, G.G., Sorenson, S.B. and Tuchfarber, A.J. (1988). A Comparison of the Kish and Last Birthday Methods of Respondent Selection in Telephone Surveys. *Journal of Official Statistics*, 4, 307-318.

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

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

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. *New York: Springer-Verlag.*

Stenbjerre, M. (2002). Online Live Exit Poll During Danish General Elections. *Presented at the AAPOR meetings, St Pete Beach, FL.*

Terhanian, G., Smith, R., Bremer, J., and Thomas, R.K. (2001). Exploiting Analytical Advances: Minimizing the Biases Associated with Internet-Based Surveys of Non-Random Samples. *ARF/ESOMAR: Worldwide Online Measurement, ESOMAR Publication Services*, vol 248, pp. 247-272.

Appendix: The Lifestyle Questions used for Propensity Score Weighting

Q: Does the risk that someone can abuse or spread personal information about you make you abstain from… *(Check one alternative per row)*

	Yes	No	Don't know	Don't want to answer
…buying a product or service over the phone?	□	□	□	□
…paying a restaurant bill with credit card?	□	□	□	□
…shopping on the Internet with credit card?	□	□	□	□
…using a cash machine?	□	□	□	□
…leaving information about yourself, that can be used for offering you tailored services, products or information, on Web pages?	□	□	□	□

Q: Today, many companies collect information about their customers' hobbies and lifestyle in order to tailor information, services and products accordingly. Do you consider this adaptation to the individual positive or negative, or do you not have an opinion? *(Check one)*
□Yes, individual adaptation is a positive thing
□No, individual adaptation is not a positive thing
□Don't know
□Don't want to answer

Q: Some people feel that they fail to notice things that happen around them. Do you feel this way? *(Check one)*
□Yes
□No
□Don't know
□Don't want to answer

Q: Have you, during the last month… *(Check all that apply)*
□…watched a documentary on TV?
□…gone away?
□…read a book?
□…none of the above?

Q: Some people feel that there is too much information available today through TV, radio, magazines, newspapers and the computer. Others want to have access to as much information as possible. How do you feel? *(Check one)*

□There is too much information

□I like to have access to as much information as possible

□Don't know

□Don't want to answer