Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

On Bias Adjustments for Web Surveys

Lingling Fan, Wendy Lou, and Victoria Landsman[1]

Abstract

Web surveys exclude the entire non-internet population and often have low response rates. Therefore, statistical inference based on Web survey samples will require availability of additional information about the non-covered population, careful choice of survey methods to account for potential biases, and caution with interpretation and generalization of the results to a target population. In this paper, we focus on non-coverage bias, and explore the use of weighted estimators and hot-deck imputation estimators for bias adjustment under the ideal scenario where covariate information was obtained for a simple random sample of individuals from the non-covered population. We illustrate empirically the performance of the proposed estimators under this scenario. Possible extensions of these approaches to more realistic scenarios are discussed.

Key Words: Volunteer Web surveys; Non-coverage; Weighting; Imputation.

1. Introduction

In probability-based Web surveys, each unit in the target population has a known positive probability of being sampled; thus, such surveys can be used to make valid inferences about the population. To apply them, we need to first identify the population and the sampling frame, and then generate a random sample. However, in many circumstances, it is difficult to identify units of the target population and contact a probabilistic sample from the population (Alvarez and VanBeselaere, 2005). For example, in Web surveys of all eligible voters, the sampling frame does not exist as not all potential respondents have internet access, and hence probability-based surveys cannot be used.

In non-probability-based Web surveys, drawing valid statistical inferences about the population is more difficult. Volunteer opt-in panels are commonly applied, where “open to all” recruitment is used to first create a volunteer opt-in panel and participants are then randomly selected from the panel using probability sampling. Making generalizations about the population based on volunteer survey results is potentially problematic, since the initial panel is a self-selected sample; individuals who do not have internet access or those who did not register for opt-in panels will never be sampled, causing non-coverage bias.

The purpose of this study is to explore the performance of statistical techniques for addressing non-coverage bias in volunteer Web surveys. In particular, we focus on propensity-score-based methods and imputation methods, which have been extensively used to address biased sampling in various areas of applied statistics (e.g. Valliant and Dever, 2011; Andridge and Little, 2010; Chen and Jun, 2000).

2. Volunteer Web Survey Framework

Given a finite population U of N individuals, let $\mathrm{y_k}$ denote an outcome measure and $\mathrm{X_k} = (\mathrm{X_{k1}}, \mathrm{X_{k2}}, \dots, \mathrm{X_{kp}})'$ a p-dimensional vector of explanatory variables (covariates) for unit k, $\mathrm{k} = 1, \dots, \mathrm{N}$. Two dummy variables, $\mathrm{W_k}$ and $\mathrm{V_k}$, define respectively access to the internet and willingness to volunteer for an opt-in panel: $\mathrm{W_k} = 1$ if unit k has access to the internet and $\mathrm{W_k} = 0$ otherwise; $\mathrm{V_k} = 1$ if unit k volunteers for an opt-in panel and $\mathrm{V_k} = 0$ otherwise. It is convenient to present the target population as a union of two disjoint sets: a volunteer opt-in panel, V,

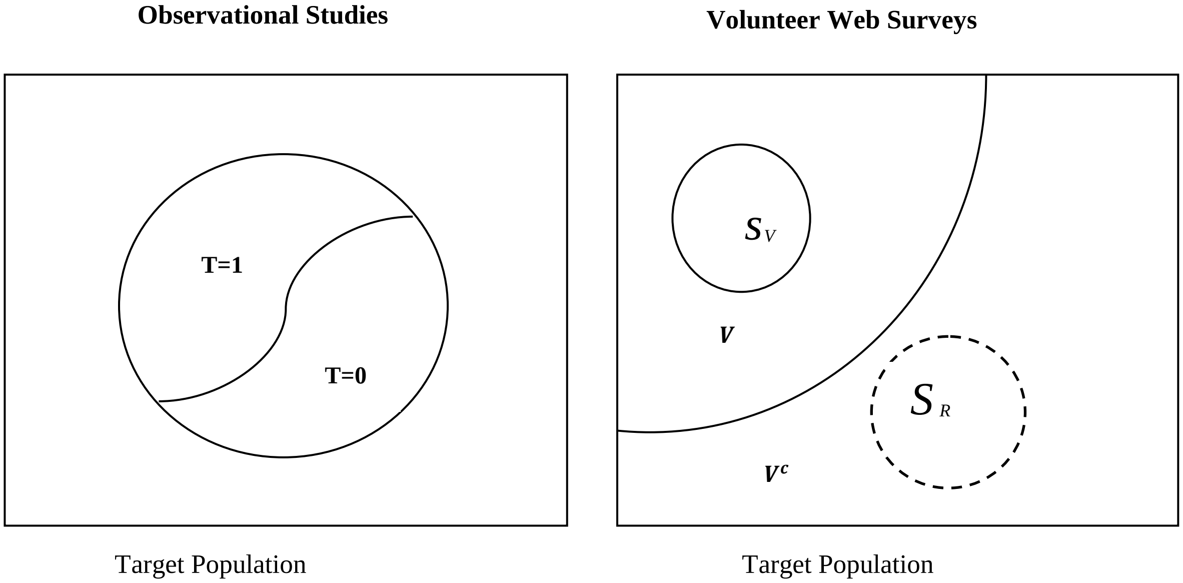
[1] Author #1 Lingling Fan, University of Toronto, 100 St. George Street, Department of Statistical Sciences, Toronto, ON, M5S 3G3; Authors #2 and #3Wendy Lou and Victoria Landsman, University of Toronto, Dalla Lana School of Public Health, Health Sciences Building, 155 College Street, Toronto, ON, M5T 3M7.

a collection of all the individuals k such that $\mathrm{V_k} = 1$ of size N_1, and a complementary set, $\mathrm{V^c}$ of size N_0. Notice that the individuals in $\mathrm{V^c}$ might or might not have internet access, whereas every individual in an opt-in panel V has internet access. Let $\mathrm{S_V}$ denote the subsample from a volunteer panel, with $\mathrm{s_k}$ being the probability that volunteer k is selected in $\mathrm{S_V}$, that is, $\mathrm{s_k = P(k \in S_V | V_k = 1,\ W_k = 1)}$; let $\mathrm{\pi_k}$ be the probability of unit k participating in the Web survey which reflects the decision of a participant k to participate in a survey. In this study, we assume that the probabilities $\mathrm{s_k}$'s are known, that there is no non-response, and that every unit with internet access registered for the volunteer panel (that is, volunteer panel and the set of units with internet access are identical). Under these three assumptions, the probabilities $\mathrm{\pi_k}$'s can be calculated by (Valliant and Dever, 2011)

$$\mathrm{\pi_k = P(V_k = 1 | X_k) s_k}$$

where the probabilities $\mathrm{P(V_k = 1|X_k)}$'s are unknown and have to be estimated.

Suppose $\mathrm{S_V}$ is a simple random sample of volunteer participants of size $\mathrm{n_1}$ from a volunteer panel V. Our goal is to estimate the population mean $\mathrm{\overline{Y}}$ for the whole population $\mathrm{U = V \cup V^c}$ from this sample. To estimate the target parameter, additional information is needed on the non-covered population $\mathrm{V^c}$. Different scenarios can be considered. In this study we assume that we select a simple random sample $\mathrm{S_R}$ of size $\mathrm{n_0}$ from $\mathrm{V^c}$ containing the covariate information only. This sample is referred to as a reference sample (Figure 1 right).


Observational Studies

Volunteer Web Surveys

Target Population

Target Population

Figure 1: Observational Studies (left) and Volunteer Web Surveys in Ideal Scenario (right)

Our simulations are motivated by Bethlehem's simulated election survey data (Bethlehem, 2010), where the target population U consists of $N = 30{,}000$ units of Age 18-80 with 19058 (63.5%) volunteers. It was found that 35.2% of units voted for the New Internet Party (NIP) among the whole population, 45.6% voted for NIP among volunteers, and young Native people were more likely to volunteer for opt-in panels and to vote for NIP.

3. Methods

3.1 Weighted Estimator

A weighted (Horvitz-Thompson-type) estimator is commonly used by survey statisticians to estimate finite population quantities from a given sample (e.g., Valliant and Dever, 2011). In the ideal scenario considered above, the weighted estimator for the population mean $\bar{Y}$ can be written using the outcome information available for the surveyed sample S_V only as

$$\hat{\bar{Y}}_{HT} = \frac{\sum_{k \in s_V} y_k \hat{w}_k}{\sum_{k \in s_V} \hat{w}_k}, \ldots\ldots (3.1)$$

where the weights are defined as $\widehat{w}_k = (\hat{\pi}_k s_k)^{-1}$.

The inverses of the s_k's are often called base weights in the survey literature (Valliant and Dever, 2011). For the scenario considered in this study, the base weights are equal to $\frac{N_1}{n_1}$ and $\frac{N_0}{n_0}$ for every individual in the surveyed and reference sample, respectively.

The probabilities π_k's reflect the self-selection mechanism of a person to participate in the opt-in panel. Since the surveyed sample S_V and the reference sample S_R are disjoint under the scenario considered, the probabilities π_k's can be viewed as the propensity scores where S_V serves as a treatment arm and S_R as a control arm (Rosenbaum and Rubin, 1983) in observational studies (see Figure 1 for comparison between observational study and volunteer Web survey). These probabilities are unknown and need to be estimated. Using the covariate information available for individuals in the reference sample, the probabilities π_k's can be estimated by fitting regression models to the joint sample $S_V \cup S_R$. Most commonly, logistic regression is the model of choice, and we choose to use it as well.

It is important to emphasize that the sampling scheme used to obtain the joint sample $S_V \cup S_R$ is informative, since the surveyed individuals and the reference sample were selected with different probabilities s_k's. In such cases, a weighted logistic regression model incorporating the inverses of s_k's (base weights) is required to obtain valid estimates for the π_k's (Valliant and Dever, 2011).

3.2 Imputed Estimator

Imputation is a commonly used method to deal with item non-response. In this method, the attempt is to create a complete data set by filling in missing values. Various imputation methods including statistical and machine learning imputation methods have been developed. Hot-deck imputation is applied widely in practice; for example, it is used very often by government statistics agencies and survey agencies. In hot-deck imputation, the missing values from a non-respondent (a recipient) are substituted by the observed values from a respondent (a donor) who has similar characteristics to the non-respondent (Andridge and Little, 2010).

Various approaches can be used to define groups of similar units, including creating imputation classes and distance metric methods. In creating imputation classes, respondents and non-respondents are classified into classes based on auxiliary variables. In the distance metric method, a distance metric is needed to measure the closeness of potential donors to recipients. The distance metric is selected based on the nature of the auxiliary variables used in imputation.

There are several hot-deck imputation methods, and we choose two of them for this study: nearest neighbor imputation (NNI) and (weighted) random hot-deck (RHD) imputation. In the RHD method, for each recipient, a donor is randomly selected within each imputation class, and the donor value is assigned to the recipient value. The only difference between RHD and weighted RHD is that base weights for units in donors set are incorporated into the imputation. In the NNI method, for each recipient j, its associated missing value y_j is imputed by a donor value, y_i, where donor i is the nearest neighbor of j measured by the X variables using the Gower distance metric (since we have both continuous and categorical imputation variables). These two commonly used hot-deck imputation methods are selected for several reasons. Firstly, both methods have several good features: the imputed values from them are actual values; the auxiliary variables involved in the matching variables are used in the imputation; they do not use an explicit model relating Y and X. Secondly, the theory of NNI estimators is well established (Chen and Jun, 2000): Chen and Jun proved that, under some regularity conditions on the distribution of the covariates and the response mechanism, the NNI sample means and any smooth function of sample means are asymptotically unbiased; NNI empirical distribution and quantile estimators are asymptotically unbiased; Chen and Jun also obtained the approximate variance of NNI estimators. Thirdly, random hot-deck imputation has some well-known properties (e.g., Rubin, 1987): for example, it can provide valid distribution and quantile estimators, but may not be as efficient as NNI imputation (Chen and Jun, 2000).

In the imputation method, we use the hot-deck imputation methods described above to impute the target variable y_j, for $j \in S_R$ in the ideal scenario (Figure 1), and then estimate the population mean using the imputed data and volunteer Web sample data as the final data:

$$\hat{\hat{\bar{Y}}}_{IMP} = \frac{N_1}{N}\bar{y}_1 + \frac{N_0}{N}\bar{y}_0, \ldots\ldots(3.2)$$

where $\bar{y}_1 = \frac{1}{n_1}\sum_{i\in S_V} y_i$, $\bar{y}_0 = \frac{1}{n_0}\sum_{j\in S_R} \hat{y}_j$, with $\hat{y}_j, j \in S_R$, being imputed values.

4. Simulation Results

To evaluate the performance of weighted and imputed estimators in estimating the population mean, relative mean bias (RMB) is used: $RMB = \frac{\bar{\theta}-\theta}{\theta} * 100\%$, $\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\theta_m$ where M is the number of simulations, θ_m is the estimated population mean calculated from the $m-th$ simulation, and θ is the true population mean. Considering the bias of the estimators depends on the sizes of the volunteer and reference samples, different combinations of volunteer and reference sample sizes were used in the simulation. The number of simulations for each combination is set to 1000 ($\mathrm{M} = 1000$), and the empirical RMB is calculated across 1000 samples. The RMB calculated based on both weighted and imputed estimators were then compared to RMB calculated based on simple random sampling (SRS) estimators from both whole population and volunteer panel. Based on the existing theoretical results, the SRS mean from the whole population and the weighted estimator (3.1) with propensity score estimated by weighted logistic regression should be approximately unbiased, while the SRS mean from volunteer panel and the weighted estimator (3.1) with propensity score estimated by unweighted logistic regression should be biased.

Table 1 and Figure 2 show the simulation results using the weighted estimator (3.1), indicating that the SRS mean from the entire population is unbiased. The SRS mean from the volunteer panel and the weighted estimator with unweighted parameter estimates are biased; the weighted estimator with weighted parameter estimates is approximately unbiased as we expected. Table 2 and Figure 3 show the simulation results using the imputed estimator (3.2), indicating that the imputed estimator with NNI, RHD, and weighted RHD (where base weights of volunteer sample are incorporated into RHD) is approximately unbiased. In addition, both estimators have very small variances compared with the SRS mean estimator from the whole population (seen in box plots in Figures 2 and 3), and by comparison, it can be seen that the imputed estimator (3.2) might have slightly better performance than the weighted estimator (3.1) in terms of bias reduction. Thus, based on the simulation studies, the bias and variance of the two estimators can be estimated empirically, as we know it is difficult to calculate the variances analytically.

Table 1: Percentage Relative Biases of Estimated Proportion of Voting for NIP in 1000 Samples Using Weighted Estimator for Different Sizes of Volunteer and Reference Samples

		SRS	Web	PROPENSITY SCORE W	PROPENSITY SCORE W-BW
n_1	n_0				
1000	1000	-0.02	29.61	-8.69	1.78
2000	1000	-0.01	29.68	4.32	1.89
3000	1000	0.00	29.72	10.69	1.80
4000	1000	0.05	29.62	14.45	1.71
2000	2000	-0.04	29.62	-8.71	1.77

SRS - SRS of size n_1 from U; Web - SRS of size n_1 from volunteer panel;
PROPENSITY SCOREW - $\hat{\bar{Y}}_{HT}$ with unweighted logistic regression; PROPENSITY SCORE W-BW -$\hat{\bar{Y}}_{HT}$ with weighted logistic regression.

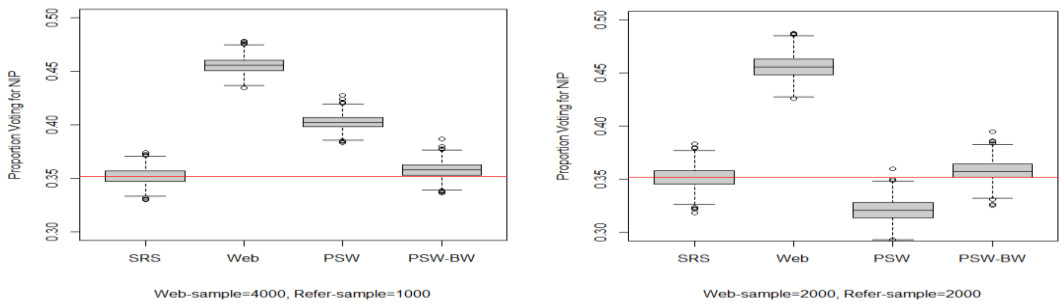


Figure 2: Simulation Results Based on Weighted Estimator

Table 2: Percentage Relative Biases of Estimated Proportion of Voting for NIP in 1000 Samples using Imputed Estimator for Different Sizes of Volunteer and Reference Samples

		SRS	Web	NNI	RHD	RHD-BW
n_1	n_0					
1000	1000	-0.35	30.40	0.30	5.04	4.67
2000	1000	-0.15	29.56	-0.30	0.79	0.77
3000	1000	-0.26	29.98	0.24	0.65	0.39
4000	1000	0.28	30.19	0.16	0.41	0.38
2000	2000	0.15	29.95	0.23	1.03	1.10

NNI - $\hat{\hat{Y}}_{IMP}$ with NNI; RHD - $\hat{\hat{Y}}_{IMP}$with RHD; RHD-BW - $\hat{\hat{Y}}_{IMP}$with weighted RHD.

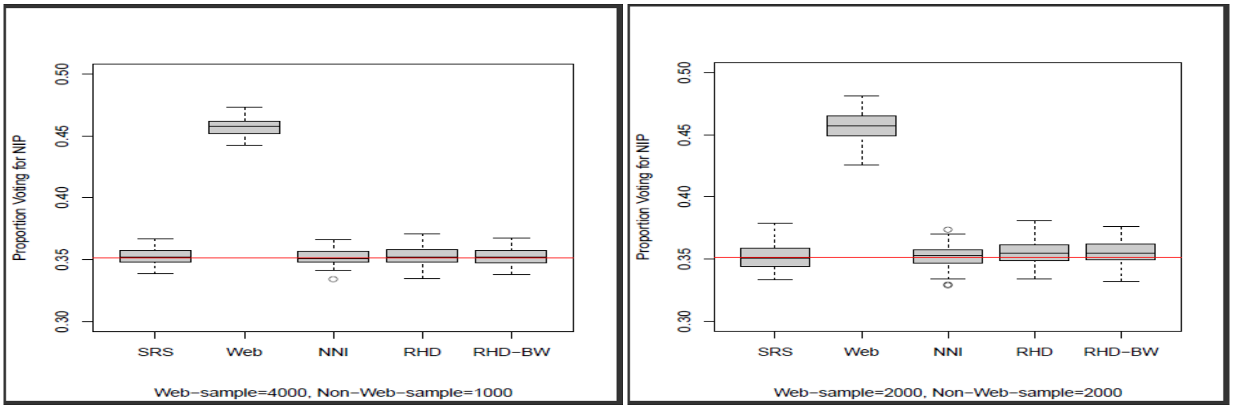
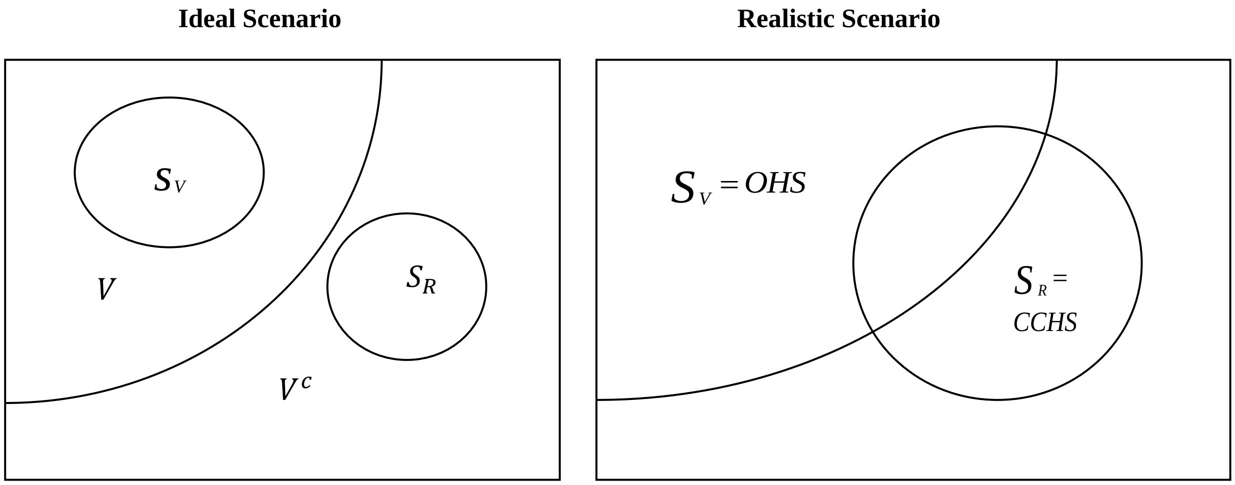


Figure 3: Simulation Results Based on Imputed Estimator

5. Discussion and Conclusion

It is well known that self-selection bias in volunteer Web surveys is more worrisome than in any other survey modes. In this study, we use propensity score weighting and imputation methods to address the non-coverage bias in volunteer Web surveys by considering the ideal scenario. Based on our simulation data set, it is found that weighted and imputed estimators perform very well in reducing non-coverage bias, and imputed estimators may perform slightly better than weighted estimators. In the weighted-estimator method, it is noted that determining the probability of unit k responding to a survey can be challenging, so we make three assumptions: (i) the probability that unit k was subsampled is known; (ii) there is no non-response; (iii) all units having internet access volunteered for the opt-in panel. Under these assumptions, we concentrate on estimating the probability of being a volunteer, which is also challenging since the base weights need to be incorporated into the regression model in order to obtain the valid estimates. However, the determination of base weights is not easy, and thus we considered a relatively simple case in this study. More complicated scenarios will be considered in the future, depending on the availability of reference samples. In the realistic scenario depicted in Figure 4, to use the propensity score method, we need to assume that the volunteer sample and reference sample are disjoint, which is a necessary but unspoken assumption in the literature on volunteer-reference surveys when propensity score model is used (Valliant and Dever, 2011). In addition, to obtain unbiased estimates, we need to determine the base weights used in the logistic regression, but their determination is difficult since reference sample consists of units from both V and V^c. In contrast, NNI and RHD imputations are easier to implement in this realistic scenario since they do not require disjoint assumption and the base weights. In addition, more approaches should be explored in the realistic scenario, such as propensity score stratification, propensity score matching, and double-robust model estimators.

The presence of missing data is especially common in volunteer Web surveys (caused by non-coverage and non-response). Numerous imputation methods have been proposed, including statistical imputation techniques (such as hot-deck imputation and regression imputation) and machine learning imputation techniques (such as CART, RF, and SVM, see Hastie, 2009), to deal with missingness. However, imputation methods—especially machine learning imputation—are not usually seen in Web surveys. Researchers often avoid using tree-based imputation methods, since they can be hard to interpret. In many cases, however, we are not interested in interpreting the trees, but only care about their ability to provide sensible imputations. Thus, the use of machine learning imputation methods, the performance of which has been proved to be excellent, in Web surveys should be explored further.


Ideal Scenario

Realistic Scenario

Figure 4: From Ideal Scenario (left) to Realistic Scenario (right)

In this study, we use hot-deck imputation to recover those missing values due to non-coverage in Web surveys. The performance of the imputed estimators based on these hot-deck imputation methods looks very promising. However, most existing imputation methods, including statistical and machine learning imputation methods, cannot incorporate survey design information such as clusters, strata, and weights into the imputation, and hence need to be modified to better incorporate survey design in the future.

References

Alvarez, R.M. and VanBeselaere, C., “Web-Based Surveys”, *Encyclopedia of Social Measurement*, 3, pp. 955-962.

Andridge, R.R. and Little, R.J.A. (2010), “A Review of Hot Deck Imputation for Survey Non-response”, *International Statistical Review*, 78(1), pp. 40-64.

Bethlehem, J. (2010), “Selection Bias in Web Surveys”, *International Statistical Review*, 78(2), pp. 161-188.

Chen, J.H. and Jun, S. (2000), “Nearest Neighbor Imputation for Survey Data”, *Journal of Official Statistics*, 16(2), pp. 113-131.

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

Hastie, T., Tibshirani, R., Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* New York, NY, Springer.

Rosenbaum, P. and Rubin, D.B. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70(1), pp. 41-55.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Valliant, R. and Dever, J. A. (2011), “Estimating Propensity Adjustments for Volunteer Web Surveys”, *Sociological Methods & Research*, 40(1), pp. 105-137.