Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Weighting for two-phase surveyed data

by Seppo Laaksonen

December, 2007

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2007
Vol. 33, No. 2, pp. 121-130
Statistics Canada, Catalogue no. 12-001-X

Weighting for two-phase surveyed data

Seppo Laaksonen [1]

Abstract

Missingness may occur in various forms. In this paper, we consider unit non-response, and hence make attempts for adjustments by appropriate weighting. Our empirical case concerns two-phase sampling so that first, a large sample survey was conducted using a fairly general questionnaire. At the end of this contact the interviewer asked whether the respondent was willing to participate in the second phase survey with a more detailed questionnaire concentrating on some themes of the first survey. This procedure leads to three missingness mechanisms. Our problem is how to weight the second survey respondents as correctly as possible so that the results from this survey are consistent with those obtained with the first phase survey. The paper first analyses missingness differences in these three steps using a human survey dataset, and then compares different weighting approaches. Our recommendation is that all available auxiliary data should have been used in the best way. This works well with a mixture of the two classic methods that first exploits response propensity weighting and then calibrates these weights to the known population distributions.

Key Words: Calibration; Internal *vs*. external auxiliary variables; Response propensity modelling method; Selective sub-sample.

1. Introduction

A standard survey is composed of one step or phase. This means that the potential survey units have first been chosen using a certain sampling design, and attempts have then been made to contact and interview these units as well as possible. However, varying amounts of non-response or other forms of missingness or data deficiencies will have occurred. Usually, addressing missingness leads to the application of post-survey adjustment methods of varying degrees of sophistication, which take advantage of available auxiliary variables. The auxiliary variables may be derived from various sources (see *e.g.*, Laaksonen 1999 or an extended version in Laaksonen 2006b, and Lundström and Särndal 2001), but for weighting purposes these are usually taken from registers, or other administrative sources or surveys. These kinds of auxiliary variables could be called *external*, if we want to distinguish them from *internal* auxiliary variables, that is, internal in the sense that the information is derived from the same survey or from its predecessor.

Internal auxiliary variables are especially used for imputations when the values for some items are missing. Such variables are also extensively used in panel surveys if a certain respondent has responded in one wave but not in another. In panel surveys, internal auxiliary information may be used both for weighting adjustments and for imputations.

This paper does not concern a standard survey as described above. It discusses two special characteristics:

(i) A survey consisting of two (or, in some sense, three) steps or phases. The first phase is like a standard survey, in which a certain number of units respond. For the second phase, we only keep in the frame the respondents who are willing to contribute to a more detailed survey. This leads first to having to distinguish such first-phase respondents who say they are willing to participate voluntarily, on one hand, and those of these respondents who actually answer the second questionnaire. This being the case, the latter subgroup will thus respond to both questionnaires.

(ii) When making attempts for post-survey adjustments, we will have the option to exploit both external and internal auxiliary variables in the second phase. The internal variables will thus be available from the first survey.

We are only considering weighting adjustments although some of our ideas could be used in imputations, too. The approach of the paper has not been much used in cross-sectional surveys although the same problem has been often met. For instance, it is typical that a face-to-face survey is conducted first and that at the end of it the interviewer will request the interviewee to respond to a self-administered additional questionnaire and, if the respondent is willing, the interviewer will hand out the questionnaire immediately for filling in, or submit it later to the volunteer. In both cases, the answers will be received by post or email. A recent example of this type is the European Social Survey (ESS) in which the supplementary questionnaire concerns especially the values of life (see www.Europeansocialsurvey.com).

1. Seppo Laaksonen, Department of Mathematics and Statistics, Box 68, FIN-00014 University of Helsinki, Finland. E-mail: Seppo.Laaksonen@ Helsinki.Fi.

Naturally, not all face-to-face respondents fill in this questionnaire.

The second-phase questions do not necessarily concern the same topic as those of the core questionnaire. Another usual strategy is to start with a broad questionnaire on a specific subject and then continue in the second phase with more detailed questions about the same subject. There can be some feedback from the first phase to the second questionnaire, and even to a sample that depends on the distribution of the key variables of the first survey (this is an example of adaptive sampling). This is often the case where there is not much experience in this type of a survey. Thus, the first survey also plays the role of a pilot survey. The so-called master samples are also close to this idea whereby the first phase survey (including a sample from administrative sources, like a micro census in some countries) may be conducted for constructing an appropriate sampling frame. In this case, the variables of the master sample are fairly limited, including usually only factual or background information.

In the case of master sampling, the objective is that the constructed sampling frame is a good representation of the target population. Hence, when going to a sample, this frame information can be well used as auxiliary data for editing, imputing and weighting of the real survey (second phase survey). Each real survey is thus a sub-sample of the master sample. We consider here a more complex case as illustrated by Table 1.

Table 1 Illustration of initial sample and three follow-up datasets

Sample with Auxiliary Variables X (*Gender*, *Age*, *Region*, *Season*) *Design* Weights for 12,554 Units without Overcoverage	First Phase Respondents with Variables Y_1 (*e.g.*, *Health*, *Outdoor*) *Basic* Weights for 10,666 Units	Volunteers for the Second Phase with Variables Y_1 Weights for 8,481 Units	Second Phase Respondents with Variables Y_2 (*e.g.*, *Skating*, *Boating*) Weights for 5,480 Units

First, there is a standard sampling procedure including some auxiliary variables X. A fairly high response rate was obtained (10,666 out of 12,554 units, about 85%) for the first survey. Some attrition occurred due to the fact that all respondents were not willing to participate in the second survey (we now have 68% of the initial sample left). Due to a rather high non-response rate in this second phase (in spite of voluntariness), our remaining sub-sample covers only 44 per cent of the initial sample. We now have the following three datasets available for the analysis:

A. First-phase respondents with survey variables Y_1
B. Second-phase respondents with survey variables Y_2
C. Both first and second-phase respondents with survey variables Y_1 and Y_2.

Most users will receive both files, A and B, and they can merge these together and obtain file C if a common identifier is available. What does a user expect having received both data files? Naturally, that the estimate for the same parameter from both files is as identical as possible, that is, the results are consistent with each other. The user obviously understands that a certain parameter estimated from the smaller file C is less accurate than that estimated from a larger file. In principle, it is possible to impute the missing values for variables Y_2, but we do not believe that it is possible to do this well, hence we approach this question by weighting. Our aim is to attempt to construct adjusted sampling weights for file B so that the analysis over variables Y_1 and Y_2 is as adequate as possible.

Several strategies can be used for this kind of weighting. Useful general aspects have been presented by, among others, Kalton and Kasprzyk (1986), Little (1986), Särndal, Swensson and Wretman (1992), Fuller, Loughin and Baker (1994), Wu and Sitter (2001), and Lundström and Särndal (2001). If we assume that the missingness only depends on the sampling design, we can construct the weights for files A and B in the respective way. For example, if stratified random sampling has been applied, the same stratification would naturally be applied to both phases. In the case of post-stratification, an analogous strategy may be applied.

In our particular example survey, the sampling frame contained the respondents of the first rotation group of the 12 months of the Finnish LFS. Each monthly sample was drawn randomly. The LFS is based on simple random sampling, but due to nonresponse these weights were adjusted by a standard calibration technique (Deville, Särndal and Sautory 1993) using *gender*, *age group* (six categories) and *region* (five categories) as auxiliary variables. Later, we refer to these as design weights. The basic weights for the first-phase respondents were constructed correspondingly adding variable *season* ($4 \times 2 = 8$ categories over two years) to the pattern of auxiliary variables. The '*season*' variable is rarely used in Finnish surveys but was here considered to be necessary due to the 'seasonal' nature of the survey (see Section 2). The present paper does not consider this aspect in detail. The first three variables are usual in Finnish human surveys because such information can be validated well from updates in the population register. This being the case, we now presume

that we will have the best possible estimates for the first-phase respondents when using these adjusted calibrated weights. In any case, we have no further access to other possible useful information from external sources.

It is possible to use estimates obtained from the first-phase respondents as benchmarking information to calibrate the second-phase weights. This strategy is not difficult as such, but all variables X and as many of the variables Y_1 as possible should be included in this process. Moreover, as precise aggregate or domain levels as possible should be found for this strategy, which is not an easy job and hence not attempted in this study. Nevertheless, it is not guaranteed that the estimates for other aggregates will be unbiased enough (results in Laaksonen 1999 give some evidence for this conclusion).

My proposed strategy is more straightforward and works without technical problems for all domains although it is not of course guaranteed that a possible bias will be substantially reduced for all domains. Thus, I have not tried any advanced calibration strategy, although this could be workable. I hope that other authors will show its possible benefits. A useful reference for them is the paper by Dupont (1995) that considers calibration of two-phase survey data, however without empirical evidence. It should be noted that I use calibration, but not a very advanced one (see Section 3).

The proposed methodology of this paper is largely based on a response propensity modelling that has been successfully used in other types of situations, see *e.g.*, Ekholm and Laaksonen (1991), Laaksonen (1999), Duncan and Stasny (2001), and Laaksonen and Chambers (2006). The situation of Rizzo, Kalton and Brick (1996) is fairly close to the two-phase case of this paper, although it is concerned with a panel. Their methodology also has some similar features. In addition, a major difference concerns the response mechanism that here occurs in two steps, that is, both due to voluntariness and due to response in the second phase. We analyse these steps separately, too. Naturally, we compare the results obtained with alternative techniques. In Section 2, we briefly further describe our surveys and datasets, and Section 3 details the principles of our methods. Section 4 presents comparison results, and Section 5 draws a conclusion.

2. Principles of the datasets

The data are from a special survey conducted among Finnish citizens aged 15-74 years old (for more information, see Virtanen, Pouta, Sievänen and Laaksonen 2001). The topic concerns their leisure time activities, especially relating to outdoor hobbies and activities. First, a CAPI (computer-assisted personal interview) survey was conducted, covering various leisure time and hobby questions such as cycling, motorcycling, walking, jogging, sailing, swimming, hunting, fishing, nature photography, skiing, skating and riding. In all cases, the reference period was the previous year. Second, at the end of this survey the respondents were asked whether they would be willing to receive a special postal survey questionnaire in which more detailed questions would be asked about some of these activities. This survey would be conducted in a few weeks' time.

The survey was conducted over two years (1998-2000) in order to reduce response and interviewing burden. Another reason for this was that since these activities are seasonal to some extent, the responses were expected to be seasonally influenced (*e.g.*, responses to questions about skiing might be different in summer and winter). The initial sample size after the removal of overcoverage (104 units of over-coverage) was 12,554 individuals.

We chose the following binary variables for our analysis presented in Section 4: *Outdoors* (person has performed regularly some outdoor activities in the nature), *Health* (is health good enough for outdoor activities?), *Skiing*, *Fishing*, *Skating*, *Boating*, *Cycling* and *Jogging*. In all cases, value = 1 means that a person has engaged in the activity during the preceding year, and value = 0, respectively, the opposite. All these variables were included in the first-phase questionnaire and we hence knew what to expect after the two consecutive phases. Note, that there are more complex variables in the data set but this simpler choice was made in order to interpret results more easily. The main conclusions are the same in the case of another choice.

In Section 4, we present two types of comparison, (i) those based on known information from the first phase, and (ii) those not based on known information. In both cases, we can fortunately check how well we have succeeded in the reduction of bias since we actually know the 'true' (or best possible) estimates. In addition, we analyse some variables only included in the second questionnaire, but we cannot say definitely how well each method works in these cases. We do not present the latest results in detail but these were observed to behave similarly to those of the second approach.

3. Response propensity modelling method and calibration

This study comprises three steps with the following weighting specifications:

First, well-designed calibrated sampling weights for the first phase respondents were created using the variables *Region*, *Gender*, *Age group* and *Season* (see also Section 1). Let us use symbol w_k for these sampling weights for

respondent k. These weights thus are based on calibration, and also called '*Basic*'. Note that before this we have, before naturally, constructed *design weights* for the dataset, based on the stratified random sampling design. These are thus available for the non-respondents of the first phase, too.

Next, we model voluntariness/response probabilities using the most common link function (Logit is not necessarily the best link function as learned from (2006a), but this is what we use here.), that is, $\text{logit} = \log(\pi/(1-\pi))$, in which π is the binary response probability (either $1 =$ volunteer *vs.* $0 =$ non-volunteer or $1 =$ respondent *vs.* $0 =$ non-respondent) and the explanatory variables consist of variables X and some variables Y_1 that have been considered to be 'good.' The model gives the predicted response probabilities p_k that are now used in the following way when constructing each particular adjusted sampling weight:

$$w_k(\text{res}) = \frac{w_k}{p_k} g_c.$$

Here $g_c = a$ scaling factor which benchmarks the weights to certain known aggregates at level c. There are several alternatives for this benchmarking, but some type of calibration could be considered as a standard way. In this study we use post-stratified aggregates h (this being the cross-classified cell of all three X variables = *Age group*, *Gender* and *Region*, the whole number of cells = 6*2*5 = 60) using the following straightforward technique

$$g_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}.$$

As already pointed out, the quality is high in Finland for these kinds of post-stratified aggregates but not necessarily for any other aggregates.

Because we have two steps for the second phase, we have the following three model options that were all also used in Section 4:

(a) Model for voluntariness
(b) Model for the response given that the person volunteered (called also '*TwoStep'*).
(c) Model for the response as one step (called also '*Direct'* and *'OneStep'*).

Note that steps (a) and (b) together give the weights for file B. This leads to the following formulation (vol = volunteer; $p_1 =$ estimated response probability at step 1, $p_2 =$ estimated response probability at step 2; respectively for the scaling factors g_1 and g_2):

Step 1: $w_k(\text{vol}) = \frac{w_k}{p_{1k}} g_{1h}$,

Step 2: $w_k(\text{res}) = \frac{w_k(\text{vol})}{p_{2k}} g_{2h}$.

The correct sampling weights had to be used in each modelling task. For models (a) and (c) this meant weights w_k, but for model (b) weights $w_k(\text{vol})$. In our comparison tests we also modelled the first-phase response and here we used design weights. The use of weights in the modelling gives more correct estimates, since we are trying to make our analysis representative for the target population. In some cases, the influence of weights is substantial, like in business surveys where weights often vary more than in standard household surveys. In this case, the results between weighted and unweighted models were not highly different, although the weighted ones should be used (this is well justified in Laaksonen and Chambers 2006 in which the influence of weights is substantial; Rizzo *et al*. 1996 also use weights). The empirical results (estimates, their standard errors and response probabilities) for the weighted solutions are presented later in Section 4.

In addition to our above key techniques, in the next Section we also use weights w_k when providing our '*best possible*' estimates for such parameters that are known, thus based on variables Y_1.

Moreover, we compare our specific results using post-stratified calibration only without modelling (we use also symbol '*cal*' in the remaining sections). The latter could be interpreted as a very standard way of approaching the weighting problem (this was a house style prior the methodology proposed here). Note, however, that if a response model only includes the variables (and the same categories) used in post-stratification, the response propensity-based weights are exactly the same as obtained by post-stratification.

4. Empirical results

This Section presents results from different methods. First, we give results from different response models and then go on to compare different weights with each other, and at the end of the Section compare some parameter estimates based on different techniques.

4.1 Models for voluntariness and response

In order to fully understand the behaviour of missingness (due to non-response and voluntariness) in all three phases of the survey, we present in Table 2 results that are based on such auxiliary variables X that are available in each step (in practice, we also used the variable 'season' but do not include its effects in this analysis since it is not a key issue in this paper).

Table 2
Logistic regressions using the three common explanatory variables in the three phases, that is, for the first phase respondents, for the voluntary respondents in the second phase and for the real respondents in the second phase. The estimates are odds ratios; their 95% confidence intervals are presented in parenthesis

Explanatory variables and other statistics	Model 1 First phase response	Model 2a Voluntariness	Model 3a Response for volunteers	Model 4a Second phase response
Gender (ref. Female)				
Male	0.71	0.84	0.75	0.77
	(0.65, 0.78)	(0.76, 0.93)	(0.68, 0.83)	(0.71, 0.83)
Age group (65+)				
24 and under	1.00	5.57	0.51	1.49
	(0.83, 1.21)	(4.65, 6.68)	(0.41, 0.64)	(1.28, 1.73)
25-34	0.96	4.76	0.65	1.73
	(0.79, 1.15)	(4.00, 4.81)	(0.52, 0.81)	(1.49, 2.00)
35-44	0.85	4.08	0.64	1.62
	(0.71, 1.02)	(3.46, 4.81)	(0.52, 0.80)	(1.40, 1.88)
45-54	0.89	3.16	0.86	1.82
	(0.74, 1.07)	(2.71, 3.69)	(0.69, 1.06)	(1.58, 2.10)
55-64	1.18	2.05	1.15	1.75
	(0.96, 1.45)	(1.74, 2.41)	(0.90, 1.47)	(1.49, 2.04)
Region (North)				
South-East	0.55	2.12	0.96	1.35
	(0.46, 0.66)	(1.79, 2.50)	(0.79, 1.16)	(1.17, 1.55)
South-West	0.76	1.83	1.04	1.35
	(0.64, 0.91)	(1.57, 2.14)	(0.86, 1.25)	(1.18, 1.55)
Mid-West	1.14	2.14	1.16	1.56
	(0.91, 1.42)	(1.77, 2.59)	(0.93 1.43)	(1.33, 1.83)
Mid-East	0.96	1.20	1.15	1.19
	(0.78, 1.18)	(1.01, 1.44)	(0.92, 1.43)	(1.02, 1.40)
Number of observations	12,554	10,666	8,481	10,666
-2 Log L	10,904	10,296	8,569	14,618

There are many interesting outcomes in these consecutive missingness behaviour models. The results of the first survey are fairly ordinary, for example, men respond more poorly than women in both phases. The response propensities are also lower in the South than in the rest of the country. The differences between age groups are somewhat surprising since the middle-age groups respond most poorly.

The voluntariness estimates are different. People in the Mid-East and North are the least willing to participate in the second survey, but the response premia given that a person is voluntary do not differ much. By age, it seems that younger people are more willing to participate but do not, nevertheless, respond very well. Older people, thus, seem in this sense to be more prepared to make a commitment than young people. However, we see clearly that the oldest ones will be under-represented without adjustments.

When considering the two first internal auxiliary variables (Table 3), it is observed that the people who are not relatively healthy (variable *Health*) and who do not actively pursue recreation in nature (*Outdoor*), are not willing to receive any new questionnaire, either. This is seen from the very high odds ratios. Interestingly, the respective odds ratio for the variable *Health* is close to the one for the volunteers. This, thus, means that a non-healthy person is not very likely to volunteer, but if he/she does, she/he responds as well as a healthy one. The tendency is similar with the variable *Outdoor*. It should be noted that the non-healthy and non-outdoor domains are not very large and although their roles in the response propensity modelling are important, their impacts on the final estimates are not very dramatic (Section 4.3).

When adding the other two internal auxiliary variables, that is, *Skiing* and *Fishing*, the same selectiveness continues although not as substantial. As a conclusion, we see clearly that the response mechanism of the second survey does not seem to be very non-informative. Consequently, it is expected that this leads to some effects on reweights and on survey estimates. These are considered in the next two sub-sections.

Table 3
Logistic regressions using some auxiliary variables from the first phase respondents in addition to those used in Table 2. The model numbers in this table and in Table 2 correspond to each other so that the response variable and the datasets are the same

Explanatory variables and other statistics	Model 2b Voluntariness	Model 3b Response for volunteers	Model 4b Second phase response	Model 4c Second phase response
Gender (ref. Female)				
Male	0.94	0.77	0.82	0.75
	(0.85, 1.04)	(0.69, 0.85)	(0.75, 0.88)	(0.68, 0.83)
Age group (65+)				
24 and under	4.92	0.52	1.30	0.51
	(4.07, 5.97)	(0.41, 0.65)	(1.12, 1.52)	(0.41, 0.64)
25-34	3.83	0.65	1.46	0.65
	(3.18, 4.60)	(0.52, 0.81)	(1.25, 1.70)	(0.52, 0.81)
35-44	3.26	0.64	1.37	0.64
	(2.74, 3.88)	(0.51, 0.80)	(1.18, 1.58)	(0.52, 0.80)
45-54	2.59	0.85	1.56	0.86
	(2.20, 3.05)	(0.68, 1.06)	(1.34, 1.81)	(0.69, 1.06)
55-64	1.73	1.18	1.55	1.15
	(1.45, 2.05)	(0.89, 1.46)	(1.32, 1.81)	(0.90, 1.47)
Region (North)				
South-East	2.15	0.96	1.34	0.96
	(1.81, 2.55)	(0.79, 1.16)	(1.16, 1.54)	(0.79, 1.16)
South-West	1.92	1.04	1.36	1.04
	(1.64, 2.26)	(0.86, 1.25)	(1.19, 1.56)	(0.86, 1.25)
Mid-West	2.09	1.15	1.52	1.16
	(1.71, 2.54)	(0.93, 1.43)	(1.29, 1.78)	(0.93 1.43)
Mid-East	1.17	1.15	1.18	1.15
	(0.98, 1.41)	(0.91, 1.42)	(1.00, 1.38)	(0.92, 1.43)
Outdoor	3.04	1.24	1.93	1.77
	(3.43, 2.71)	(1.43, 1.07)	(2.15, 1.74)	(1.97, 1.59)
Health	3.61	1.02	2.71	2.49
	(4.61, 2.82)	(1.58, 0.66)	(3.51, 2.09)	(3.24, 1.92)
Skiing				1.36
				(1.47, 1.25)
Fishing				1.27
				(1.38, 1.17)
Number of observations	10,666	8,481	10,666	10,666
-2 Log L	9,721	8,560	14,342	14,244

4.2 Comparison of different weights

As already explained, we provided several weights. Table 4 gives a summary of these with descriptive statistics in order to explain the changes that occur after each adjustment operation. The design weights cannot be used in our comparisons since no data on variables Y are available for the initial sample. It is, however, illustrative to see that it has the lowest relative variation measured here with $1 + cv^2$ in which cv is the coefficient of variation. This formula is also used as an approximation of the design effect (DEFF). Rizzo *et al*. (1996) also use this indicator when comparing their weights.

The changes are not dramatic in the first step, that is, from design weights to first-phase basic weights (except for the average that is related to decreasing counts), but in the following two steps the DEFFs are higher. We also see that the variation for both calibrated weights is lower than that for the respective response propensity-based weights. The distribution for each weight is skewed to the right, least for the design weights, naturally. It is somewhat surprising that the skewness is the highest for the volunteer weights. More details about the weight distributions and the differences between the weights are presented in Figures 1 to 3.

Figure 1 illustrates well how some weights have increased substantially due to the response propensity modelling (Model 2b). It is possible to look in detail to see which types of units are under the plots with high weight increase. For example, behind the separate left-side plots with *RP* weights higher than 700 are persons who are not healthy and do not engage much in outdoor activities but are, nevertheless, still in the volunteer data file. Similarly, we can find other interesting groups by using the results from the model estimations. However, the majority of the plots are in the same area and, consequently, less changes can be expected in the estimates than in the area with more substantial weight changes.

Table 4 Descriptive statistics for different sampling weights. *RP* = Response Propensity

Weight	Phase	Unit size	Average	Skewness	$1 + cv^2$
Design Weight	Zero	12,658	308	0.94	1.30
(Calibrated) Basic Weight	First	10,666	365	1.30	1.39
Calibrated Weight	Volunteers	8,481	460	2.52	1.63
RP Weight, Model 2b	Volunteers	8,481	460	4.60	1.82
Calibrated Weight	Second	5,480	712	1.64	1.62
TwoStep RP Weight, Models 2b and 3b	Second	5,480	712	3.60	1.84
OneStep RP Weight, Model 4b	Second	5,480	712	2.56	1.80

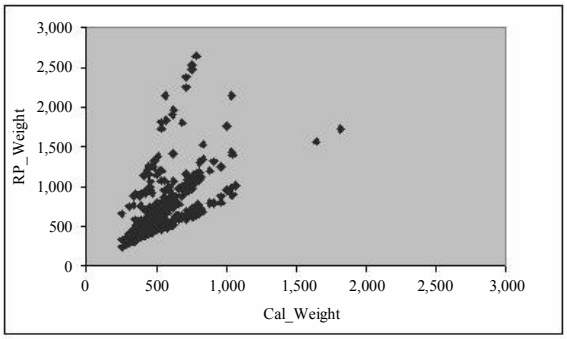

Figure 1 Scatter plot between the two volunteer weights

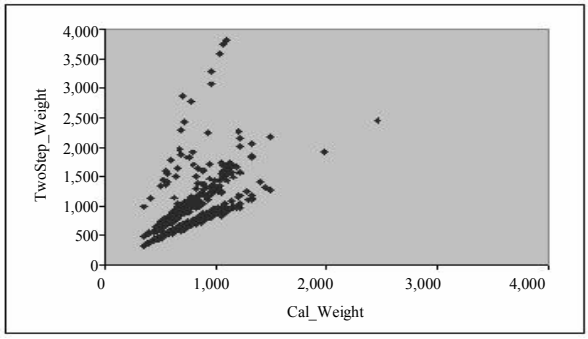

Figure 2 Scatter plot for second phase respondents between the calibrated weight and the two-step response propensity based weight

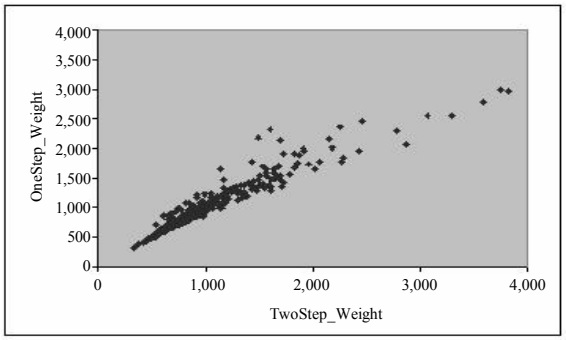

Figure 3 Scatter plot for second phase respondents between the two alternative response propensity modelling weights

The dispersion in the Figure 2 scatter is somewhat stronger than that in Figure 1 but the profile is similar. Consequently, interesting sub-groups can be found behind distinct plots.

Finally, Figure 3 compares the two alternative second phase weights with each other. This scatter differs considerably from the previous two, since the relationship is rather linear. The maximum values of the two-step weights are higher than those of the one-step weights, but the weights of many one-step weights are, however, clearly higher. For example, non-healthy people who, however, engage in outdoor activities receive relatively high one-step weights but there is no clear age effect. On the other hand, people with little outdoor activities in older age groups receive relatively high two-step weights but health does not relate to them. Nevertheless, it is not expected that there will be big differences in respective estimates although one of these two alternatives should have been introduced into use. If this choice were a simpler one, that is, one-step weighting, it would still be useful to analyse both steps and their response propensities separately in order to understand better the reasons for both types of missingness.

4.3 Comparison of parameter estimates

We have not been able to make complete simulation studies with different assumptions in order to analyse which type of method would be best in each particular case. Fortunately, we can get quite close to this by comparing the effects on the estimates from three different perspectives. First, we have prepared the response/voluntariness models by using both X and some Y_1 variables. Consequently, we know the 'best' parameter estimates based on these Y_1 values from the first survey. Second, we add auxiliary variables Y_1 in the model but exclude some Y_1 values from them. However, we know the 'best' values in these cases and thus can make exact comparisons. Third, we can compare some estimates that are not known in any way. In

this last case, we can only deduce which values might be the best.

We present our explicit results based on the variables described in Section 2. Note that we do not consider it important to present standard errors for each estimate because we are concentrating on the biases in these estimates. However, it is good to notice that the standard errors are around 0.2-0.4 percentage points for the first-phase data set and around 0.3-0.5 percentage points for the second-phase data set (lowest always in *Health*, second in *Jogging*, and highest in *Outdoor* and *Skiing*).

Figure 4 gives the results based on the weights without using any adjustment (that is often the case in practice, unfortunately). We see that the bias is substantial in most estimates, lowest in *Jogging,* which was not very actively practised when compared to *Outdoor,* for example. In general, most users are unhappy with such big biases that are statistically significant and highly significant except for *Jogging* in the second phase (*e.g.*, the 95% confidence interval of the bias for *Health* is from 1.7 points to 2.3 points). Here, as in later results, the bias means *over-estimation* so that while missingness increases the estimate becomes too high. The results without good adjustments will be too optimistic, that is, people seem to do 'too much' of all exercises. Note that the same tendency is obviously also in the first-phase estimates but we cannot justify this. There are surprising differences between those two estimates, sometimes the 'volunteer' data give a more biased result, sometimes it takes the second-phase respondents data. We do not interpret these in detail but naturally they reflect differences in missingness, and can be considered to be warnings for a user.

For comparison, we show again in Figure 5 the same unadjusted results for volunteers as in Figure 4 but we have added the corresponding estimates based on post-stratified calibration and response propensity modelling. This graph clearly shows that post-stratification gives some benefit compared to the unadjusted solution. However, the response propensity method is the best in each case, and extremely good for *Health* and *Outdoor* that have been used as auxiliary variables in the supported models.

Figures 6 and 7 concern the final-step estimates and are thus the most important. Figure 6 shows the same conclusion as Figure 5 in the sense that the response propensity technique is superior to post-stratified calibration although all differences are not statistically highly significant (especially *Jogging*). The difference between the one-step method and the two-step method is fairly small and the bias varies from one variable to the next. Hence, basing on this study, we cannot say which of these two specifications is better.

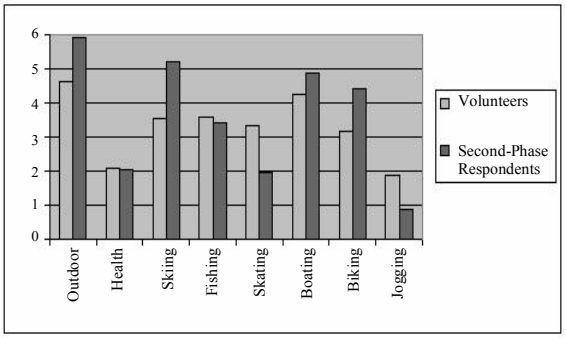

Figure 4 Bias in estimates in percentage points based on unadjusted sampling weights for second-phase respondents and for volunteers

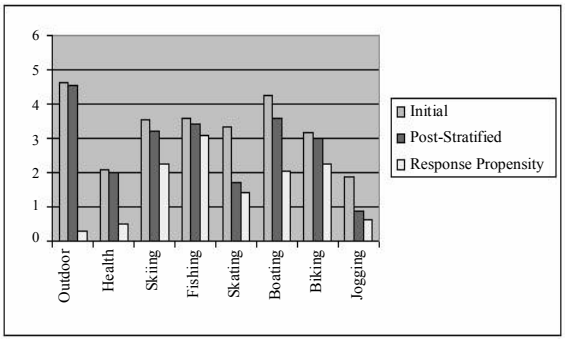

Figure 5 Bias in estimates in percentage points for volunteers based on unadjusted sampling weights (symbol = 'Initial'), post-stratified calibration and response propensity method in which variables *Outdoor* and *Health* have been used as auxiliary variables (Model 2b in Table 3)

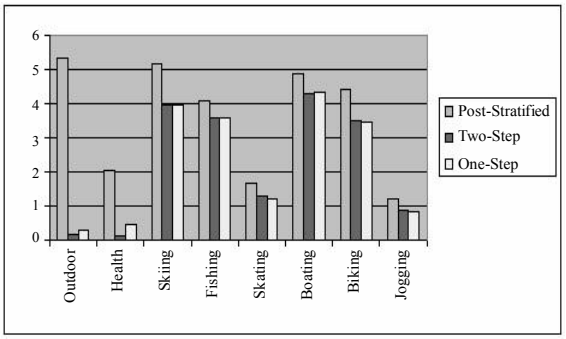

Figure 6 Bias in estimates in percentage points for respondents after both steps based on post-stratification, and the two response propensity methods, *i.e.*, 'Two-step' and 'One-step' so that variables *Outdoor* and *Health* have been used as auxiliary variables. The two-step method is based on the two consecutive models (Models 2b and 3b in Table 3) whereas the one-step method has been constructed direct to the second-phase respondents (Model 4b in Table 3)

Figure 7 presents some comparisons when the two new variables have been added to the response propensity model. The results are quite predictable since this reduces the bias in these estimates and in all other estimates to some extent as well. The bias is still too large in *Boating* and *Biking* in the opinion of many users, I suppose. We can reduce this bias, naturally, by adding new auxiliary variables to the model. How far could we go in this? This has not been examined further in this study. On the other hand, we have worse tools for reducing bias in such variables that have been based on the second survey only. We tested several such estimates and observed some changes in corresponding estimates, being of the same level as in the cases of *Boating* and *Biking* in Figure 7. In this case, however, we cannot check the bias. We can only believe basing on our previous exercises that these results are less biased than those based on more poorly adjusted ones.

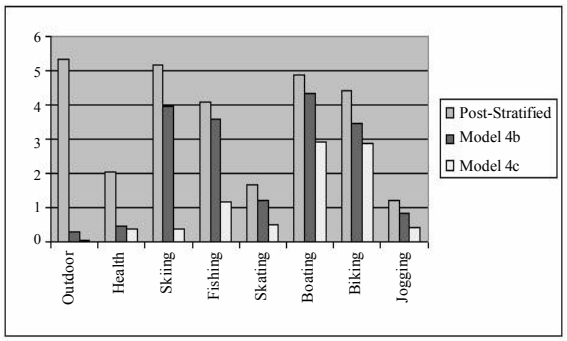

Figure 7 Bias in estimates in percentage points for respondents after both steps based on post-stratification, and the two 'One-step' response propensity methods so that variables *Skiing* and *Fishing* have also been used as auxiliary variables (Model 4c in Table 3). These are compared to those based on Model 4b

5. Discussion

The problem discussed in this paper is common in surveys. There are many surveys which are conducted in more than one step, and some inconsistencies have occurred between these surveys due to missingness and other discrepancies. An internationally well-known example is the European Social Survey (ESS) that includes two questionnaires, a core one and a supplementary one. The number of respondents is naturally smaller for the latter than for the former. This leads to some selectiveness, for example, responding to the second questionnaire being positively associated to political activity. This is awkward from the user's point of view because an estimate based on a larger dataset differs from that based on a smaller one, although both concern the same variable and time period.

Similarly to the ESS, this study concerns two-phase surveyed data. The response rate in the second survey was substantially lower than in the ESS. The effect from selectiveness is also higher. Using the response propensity models we predicted this selectiveness and exploited the results in weighting adjustments, and as the final step we calibrated the sums of the weights to correspond to certain known population aggregates. This strategy aims at making the most of all available auxiliary information, derived both from registers and other external sources, and also of the previous phase of the survey at the micro level.

In our example, the second phase of the survey comprised two different steps but only one data collection. The first step concerned willingness to participate voluntarily in the second phase of the survey, and the second step the actual survey participation of these volunteers, respectively. We examined both steps separately and found interesting information on their response mechanisms. Moreover, we used the results from this analysis for reweighting adjustments. For the sake of comparison, we looked at these both steps in one occasion and built a respective model, and continued the reweighting analogously. Finally, we compared the estimates. It was somewhat surprising that the two results differed quite little in our examples. This is, on the other hand, a good point, since it is easier to work with one step, and hence this could be introduced into use.

We thus propose a certain methodology for two-phase sampling weighting, but cannot say definitely which specification would be the best in each particular case. Our methodology is quite easy to exploit, but the advantages from it depend naturally largely on the availability of good external and internal auxiliary data. If no direct auxiliary variables are available, it will not be clear how good the adjusted estimates will be. Our examples show that these will be easily less biased than the initial ones. However, our recommended technique seems to be somewhat conservative so that all the best adjusted estimates in our analysis are slightly overestimated although not statistically significantly. This is an interesting question for future research that is still needed especially because this problem is becoming more common in the survey world. Another interesting topic for future research is how to make an optimal choice of auxiliary variables in the two-phase survey setting.

Acknowledgements

The author would like to thank the editor and the anonymous referees for their precise and helpful comments.

References

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

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

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21, 125-135.

Ekholm, A., and Laaksonen, S. (1991). Weighting via response modelling in the Finnish household budget survey. *Journal of Official Statistics*, 7, 2, 325-337.

Fuller, W.A., Loughin, M.M. and Baker, H. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Laaksonen, S. (1999). Weighting and auxiliary variables in sample surveys. In "Enquêtes et Sondages. Méthodes, modèles, applications, nouvelles approches" (Eds., G. Brossier and A.-M. Dussaix). Dunod, Paris. 168-180.

Laaksonen, S. (2006a). Does the choice of link function matter in response propensity modelling? *Model Assisted Statistics and Applications*, *An International Journal*. Publisher: IOS Press. 1, 95-100.

Laaksonen, S. (2006b). Need for high quality auxiliary data service for improving the quality of editing and imputation. In *United Nations Statistical Commission*, "*Statistical Data Editing*", 3, 334-344.

Laaksonen, S., and Chambers, R. (2006). Survey estimation under informative non-response with follow-up. *Journal of Official Statistics*, 81-95.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

Lundström, S., and Särndal, C.-E. (2001). Estimation in the presence of nonresponse and frame imperfections. Statistics Sweden.

Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.

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

Virtanen, V., Pouta, E., Sievänen, T. and Laaksonen, S. (2001). Luonnon virkistyskäytön kysyntätutkimuksen aineistot ja menetelmät. (Data and methods of survey on recreational use of nature). In Luonnon Virkistyskäyttö (Recreational use of nature). Finnish Forest Research Institute, 802.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.