SSC Annual Meeting, June 2024
Proceedings of the Survey Methods Section

SUBSAMPLING FOR NON-RESPONSE FOLLOW-UP IN SOCIAL SURVEYS: A CASE STUDY

Anne Mather[1], Cilanne Boulet[2], and Andrew Brennan[3]

ABSTRACT

With declining response rates, dealing with non-response has become more costly and time-intensive. One practical approach is to carry out more expensive non-response follow-up on only a portion of the sample, such as by limiting telephone follow-up to a random subsample. This is a generalization of the classical approach described by Hansen and Hurwitz (1946). The Canadian Social Survey experimented with this approach and restricted telephone operations on two iterations of the survey. This article covers the adjustments that were applied to the weights of responding units to adjust for this subsampling. Moreover, it presents alternate estimators that could be used, their theoretical properties with respect to biasedness, and the results of a simulation study comparing these estimators.

KEY WORDS: Non-response, collection effort, subsample

RÉSUMÉ

Avec la baisse des taux de réponse, le suivi de la non-réponse est devenu plus coûteux et prend de plus en plus de temps. Une approche pratique consiste à n'effectuer un suivi plus coûteux que sur une partie de l'échantillon, par exemple en limitant les interviews téléphoniques à un sous-échantillon aléatoire. Ceci est une généralisation de l'approche classique décrite par Hansen et Hurwitz (1946). L'Enquête sociale canadienne a testé cette approche et a restreint ses opérations téléphoniques lors de deux itérations de l'enquête. Cet article décrit les ajustements qui ont été appliqués aux poids des unités répondantes pour tenir compte de ce sous-échantillonnage. En outre, d'autres estimateurs qui pourraient être utilisés sont présentés, avec leurs propriétés théoriques en ce qui concerne le biais et les résultats d'une étude de simulation comparant ces estimateurs.

MOTS CLÉS : Non-réponse, effort de collecte, sous-échantillon

1. INTRODUCTION

1.1 Motivation

The current research is motivated by two important trends in recent years, at Statistics Canada as well as at other statistical organizations. Firstly, response rates to social surveys have continued to experience a decline. Secondly, there has been an increased use of online self-response as a collection mode due to its ease and relatively low cost. The declining response rates point to a need for deeper follow-up with non-respondents; moreover, the increased reliance on self-response can lead to more bias in certain situations. Although they may reduce bias, interviewer-facilitated responses (such as face-to-face or by telephone) are substantially more expensive.

The current work examines an estimator for collection activities that use less expensive non-response follow-up (e.g., mail and email) for the whole sample concurrently with more expensive interviewer-facilitated non-response follow-up for a subset of non-respondents. This is different from previous work in which the non-response follow-up was carried out after a first self-response phase (e.g., Hansen and Hurwitz, 1946 and Neusy et al., 2022) and from the experience of the National Household Survey in Canada in which the initial sampling fraction was large (Beaumont, Bocci, and Hidiroglou, 2014). Initial results indicate that there can be substantial cost savings without unduly compromising the quality of estimates.

[1] Anne Mather, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6, anne.mather@statcan.gc.ca
[2] Cilanne Boulet, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6, cilanne.boulet@statcan.gc.ca
[3] Andrew Brennan, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6, andrew.brennan@statcan.gc.ca

The Canadian Social Survey (CSS) has trialed this approach on two waves. The waves each began with a two-week period of early collection during which units could respond online; this was followed by a four-week period during which online self-response continued, but telephone follow-up was also done. In fall 2022, a 75% subsample of non-respondents received this interviewer-facilitated follow-up, and in early 2024, a 50% subsample. In both, the sampled units were split into two groups: those who would receive *full* follow-up (mail, email, text messages, and telephone follow-up), and those who would receive only *reduced* follow-up (mail, email, and text messages). It is important to stress that rather than, for example, reducing the follow-up telephone calls by some amount across the entire sample, this approach completely eliminated calls for the *reduced* follow-up subsample, but kept *full* follow-up effort on the remainder of the sample. Further, the subsamples were selected at random, and thus controlled by design. However, this subsampling introduced a risk of bias; if the non-response (NR) mechanism underlying response with *reduced* follow-up was different from the NR mechanism underlying response with *full* follow-up, then the subsampling would need to be accounted for at the estimation stage. The current work presents a generalized formulation of the weighting adjustment employed by the CSS in these contexts; work is ongoing to better understand the situations in which this approach could be expanded to other surveys or contexts.

1.2 Organization of the Paper

Section 2 defines the concurrent multi-mode estimator (CMME) used by the CSS and situates it in the context of other similar estimators. Section 3 discusses how required parameters can practically be estimated and with what data. Section 4 presents the results of simulations informed by and expanding on the CSS trials and section 5 draws conclusions from these results.

2. CONCURRENT MULTI-MODE ESTIMATOR

2.1 Development of the CMME

The differential follow-up between the two groups requires an estimator to account for any differences in the NR mechanism underlying response in each. In the waves of the CSS where all units receive *full* follow-up (i.e., waves other than the trials with differential follow-up), the estimator used is based on the Narain-Horwitz-Thompson estimator, with an adjustment for NR. Initial approaches considered in the first wave with differential follow-up were adaptations of this estimator.

One option was simply to account for the type of follow-up (*full* or *reduced*) when estimating the response propensity for each unit. For example, we could have estimated the response propensities separately in the two groups or included an indicator of type of follow-up in the NR model. However, the subsamples had been selected completely at random, without consideration of response propensity; thus, the type of follow-up was completely independent of any survey outcome variables, and accounting for such an indicator in non-response estimation would needlessly increase variance by increasing the dispersion of the weights.

Instead, we drew on the double-expansion estimator (DEE) developed by Hansen and Hurwitz (1946), which subsamples certain units for additional follow-up in a second phase and expands the weights of these units to account for those not selected. However, the DEE was developed for a context where collection effort for the non-selected units ceased in the second phase and in which there was complete response in this second phase. This is quite different from the CSS context, where there are two groups with concurrent differential follow-up, and where there is non-response in both groups. The estimator that we developed, the CMME, builds on the DEE by expanding the weights of the respondents in the subsample that received more (*full*) follow-up in a way that accounts for both NR and concurrent differential follow-up.

2.2 Formulation of the CMME

The CMME accounts for NR by incorporating the probability of response under *full* follow-up collection operations into the double-expansion estimator (DEE). It also accounts for the fact that the CSS does not have distinct collection phases by replacing the term for phase 2 respondents by two separate terms for “late” respondents under (1) *reduced* follow-up and (2) *full* follow-up. There are three groups of respondents in the estimator: those who respond during the initial self-response phase, those who are not selected for *full* follow-up and respond after the initial phase, and those who are selected for *full* follow-up and respond after the initial phase.

The formula for the CMME is:

$$\hat{t}_{y,CMME} = \sum_{k \in s_E} \frac{y_k}{\pi_k f_k} + \sum_{k \in s_{LR}} \frac{y_k}{\pi_k f_k} + \sum_{k \in s_{LF}} \frac{y_k}{\pi_k f_k} \left(\frac{1}{\varphi_k} - \frac{(1-\varphi_k)}{\varphi_k} q_k \right)$$

where

- $\hat{t}_{y,CMME}$ is the CMME estimate of the population total t_y,
- s_E, s_{LR}, and s_{LF} are, respectively, the early survey respondents during the initial self-response phase of collection, the late respondents under *reduced* follow-up collection operations, and the late respondents under *full* follow-up collection operations,
- y_k is the value of variable y for unit k in the population,
- π_k is the probability of selection into the sample for unit k,
- f_k is the probability of response under *full* follow-up collection operations for unit k,
- φ_k is the probability of unit k being selected for *full* follow-up collection operations, like being selected for phase 2 collection in the DEE, and
- q_k is the ratio between a unit's probability of responding late under *reduced* follow-up to their probability of responding late under *full* follow-up.

Figure 1 – Graphical depiction of the CMME

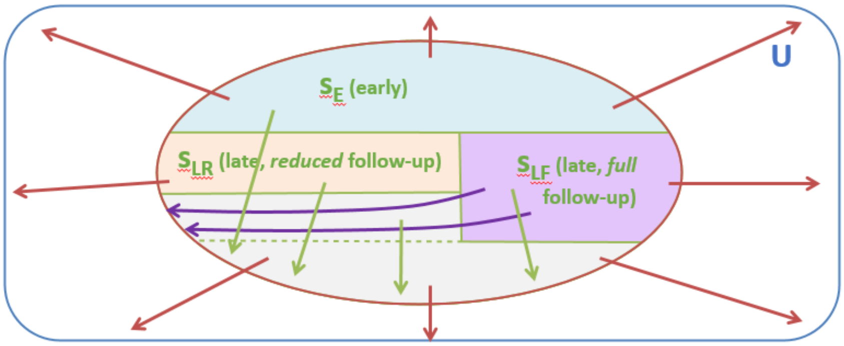


Figure 1 is a graphical depiction of the CMME. It can be broadly understood as first expanding the late respondents who received *full* follow-up (s_{LF}) in order to account for some "missing respondents": those who were not selected for *full* follow-up, but who would have responded if they had been; this expansion is represented by dark purple arrows. With the expansion, the respondents (s_E+s_{LR}+s_{LF}) together represent all units who would have responded if the entire sample had been eligible for *full* follow-up. After this, adjustments are applied to all respondents to account for NR (as represented by the green arrows) and for being selected in the sample (as represented by the red arrows).

The CMME is unbiased under correct specification of f_k and q_k, as demonstrated in Mather, Boulet, and Brennan (2024). The following sections elaborate on its components.

2.3 Ratio of Probabilities of Responding Late Under Different Modes of Follow-up (q_k)

Regardless of whether a population unit k is selected into the *full* or *reduced* follow-up subsample, it can be considered to have a probability of responding early during the initial phase of collection (p_{E_k}), a probability of responding late under each mode of follow-up (p_{LR_k} and p_{LF_k} for *reduced* and *full* follow-up respectively), and a probability of not responding at all under each mode of follow-up (nr_{R_k} and nr_{F_k}). By definition we have p_{E_k}+p_{LR_k}+nr_{R_k}=1 and p_{E_k}+p_{LF_k}+nr_{F_k}=1; that is, once a unit has been selected for a given follow-up mode, they must be either an early respondent, a late respondent, or a non-respondent. The quantity q_k is defined for each unit of the population as the ratio p_{LR_k}/p_{LF_k} for that unit.

2.4 Term for Late Full Follow-up Respondents $\left(1/\varphi_k - {(1-\varphi_k)q_k}/\varphi_k\right)$

Of the three groups of respondents (early, late with *reduced* follow-up, and late with *full* follow-up), the first two are treated in the same way in the CMME: the value y_k is expanded by a factor of $1/\pi_k f_k$ to account for sample selection and for NR under *full* follow-up. The f_k in the denominator is somewhat counter-intuitive: although it is the probability of responding under *full* follow-up collection operations, neither of these groups of respondents actually received *full* follow-up. The former responded early, and did not need telephone calls or repeated other reminders to prompt them to respond, and the latter was not selected to receive the more intensive telephone follow-up efforts. However, f_k is nevertheless defined for each unit, as it is their probability of responding (whether early or late) if they were to receive *full* follow-up.

The third group, respondents who are selected for *full* follow-up and respond late after the more intensive telephone calls have begun (s_{LF}), are further expanded by the factor $\left({}^{1}/_{\varphi_k} - {}^{(1-\varphi_k)q_k}/_{\varphi_k}\right)$. This factor accounts for differences between the two groups of late respondents, and can be thought of as inflating the late *full* follow-up respondents so that they also represent units who were not selected for *full* follow-up, but who would have responded if they had been ("missing respondents"). Looking more deeply at this factor, the following can be noted:

- Units in s_{LF} are first fully expanded by $1/\varphi_k$, like phase 2 respondents in the DEE.
- An amount is then subtracted from this, ${}^{(1-\varphi_k)q_k}/_{\varphi_k}$. The amount subtracted corresponds to the extent to which that unit is already represented in s_{LR}.

The units in s_{LF} were selected for expansion since it is expected that these are the units who are most like the "missing respondents". With a correctly specified value of q_k, each unit is expanded the appropriate amount to account for the "missing respondents". Moreover, as discussed below, the data available to estimate q_k includes questionnaire responses and thus can be quite rich.

3. ESTIMATION OF PROBABILITIES OF RESPONSE

The unbiasedness of $\hat{t}_{y,CMME}$ rests on the correct specification of f_k and q_k. However, these quantities are not known and must be estimated. On the CSS, f_k is estimated using variables that are known on the frame for all units: respondents and non-respondents. Since the primary sampling units are dwellings, only minimal information is available: geography, dwelling type, and some flags regarding expected household composition. The weak auxiliary information limits the accuracy of this adjustment. When response depends on additional factors that also related to the variables of interest, the NR adjustment cannot eliminate all of the NR bias.

On the other hand, q_k can be estimated using much richer respondent data, including demographics and even responses to the survey questionnaire. It is the ratio of two propensities (of responding under *reduced* and under *full* follow-up) that are defined for each unit in the population. Although these propensities are not known, they can be estimated by comparing the attributes of two groups of respondents: s_{LR} and s_{LF}. Moreover, a wealth of information is known about these groups, namely all of their responses to the survey in question. The richness of the information that is available to the model can reduce much of the bias that would be related to differences in captured characteristics between the two groups. Though there are many modelling options, for the CSS we modelled q_k by calibrating the design-weighted distributions of s_{LF} to those of s_{LR}. (More details are available in Mather, Boulet, and Brennan, 2024.)

4. SUBSAMPLING SIMULATIONS

4.1 Set-up

We simulated a fixed sample of 20,000 units, with a variety of auxiliary and outcome variables. The auxiliary variables can be classified into three types: those available for all units on the frame (e.g., geography), those observed for respondents during collection (e.g., educational attainment), and those never observed (e.g., busyness). These variables were combined to build the four outcome variables for which simulation results are shown below: A is a random variable that is independent of the auxiliary information, B is related to information known on the frame, C is related to variables observed during collection, and D is related to unobserved auxiliary information that is itself related to response propensities. Each unit was assigned probabilities of response p_E ($\mu = 0.25, \sigma = 0.13$), p_{LF} ($\mu = 0.31, \sigma = 0.05$), and p_{LR} ($\mu = 0.15, \sigma = 0.09$) related to their auxiliary variables. For example, units with higher levels of educational attainment are more likely to have higher response propensities. These parameters were chosen to be similar to probabilities observed in CSS data. Five hundred Monte Carlo replicates were then drawn to simulate subsampling for non-response follow-up for each of five values of φ_k: 1, 0.75, 0.50, 0.25, and 0. That is, in the first set of replicates, all units were selected for *full* follow-up; in the next set, each unit had a probability of 0.75 of being selected for *full* follow-up, and so on.

The estimators compared, in the order presented in the results that follow, are:

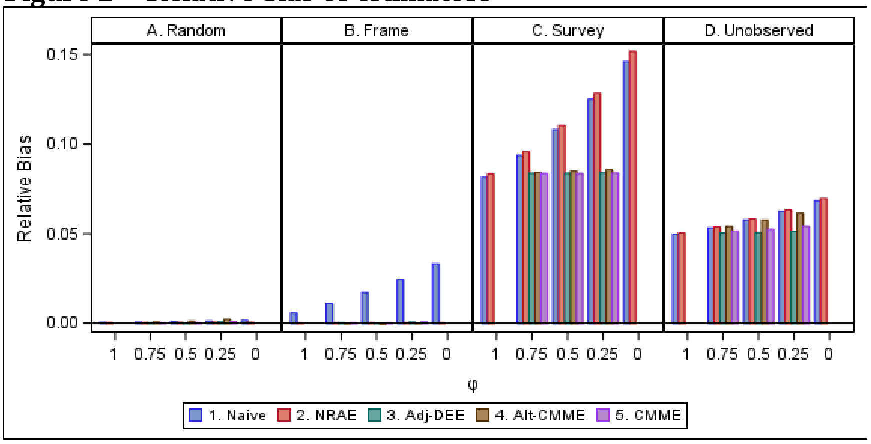
- Naïve: A non-response-adjusted Narain-Horwitz-Thompson estimator that expands respondents' design weights by π_k and by the inverse of the overall response rate.
- NRAE: A second non-response-adjusted Narain-Horwitz-Thompson estimator that expands respondents' design weights by π_k and by the inverse of an estimated probability of response; a logistic regression model with frame-level variables was used to estimate the probability of response, but it did not include an indicator for follow-up group (for reasons discussed in section 2.1).
- Adj-DEE: A non-response-adjusted double-expansion estimator, which uses a logistic regression with frame-level variables to estimate the probability of response; note that this estimator drops s_{LR}, the late respondents who received only *reduced* follow-up.
- Alt-CMME: An "alternative CMME", which expands s_{LR} instead of s_{LF}, and uses calibration on survey-level variables to estimate q_k for s_{LR} and logistic regression with frame-level variables to estimate f_k for all respondents.
- CMME: The CMME, which uses calibration on survey-level variables to estimate q_k for s_{LF} and logistic regression with frame-level variables to estimate f_k for all respondents.

The estimators were compared in terms of their relative bias and their effective sample size (number of respondents/design effect, calculated using the Kish approximation (1965)) for totals of the four outcome variables.

4.2 Results

Figure 2 shows the relative bias of the five estimators at five levels of subsampling (φ_k), for each of the four variables of interest (A, B, C, and D defined above). Only two estimators are available at $\varphi_k = 0$ and $\varphi_k = 1$, since $\widehat{q_k}$ is not relevant without respondents from both *reduced* and *full* follow-up groups for the CMME and alternative CMME, and the adjusted DEE is identical to the NRAE at these levels of subsampling.

Figure 2 – Relative bias of estimators



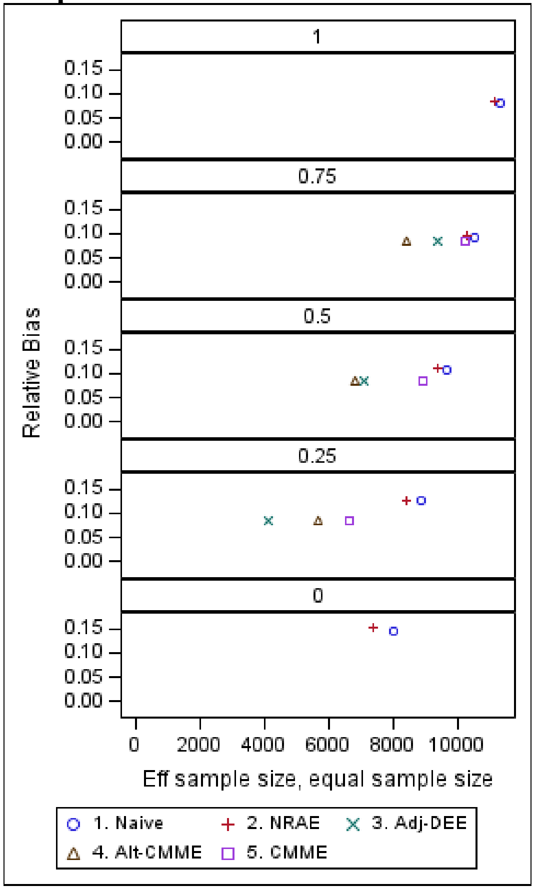
When using the naïve estimator (1. blue) or the NRAE (2. red), the relative bias for variables C and D rises as the subsampling fraction for *full* follow-up drops; when using the naïve estimator, this is also the case for variable B. However, this is not the case for the adjusted DEE (3. green), the alternative CMME (4. brown), or the CMME (5. purple); that is, the bias is not made substantially worse by subsampling for *full* follow-up at a lower rate and using one of these estimators. Additionally, any bias that cannot be accounted for by $\widehat{f_k}$ also cannot be reduced by these three estimators, so the bias will always at least as large as it is for $\varphi_k = 1$. This means that the adjusted DEE corresponds to the least amount of bias that could be achieved in these scenarios.

These simulations also show that the CMME is less biased than the alternative CMME, in particular for variable D. Indeed, by expanding the *full* follow-up respondents (who were more likely to respond than their *reduced* follow-up counterparts) to account for the "missing respondents," the CMME offers better protection against model misspecification than does the alternative CMME that expands *reduced* follow-up respondents. The *full* follow-up respondents are more likely to resemble the "missing respondents" in terms of unmeasured characteristics (such as busyness in this simulation).

Effective sample size can be used as an indication of the variance in the estimators, and of the resulting precision of estimates. A higher effective sample size suggests that estimates will generally have higher precision and/or that reliable estimates can be calculated for more granular subgroups. Figure 3 shows how – for the estimate of variable C with a fixed overall sample size of 20,000 – as the subsampling fraction drops (moving down the figure), the effective number of respondents also drops. However, the naïve and NRAE estimators also increase in bias, whereas the others do not visibly do so. The CMME outperforms the other lower-bias estimators: the DEE has a lower effective sample size since it throws out late respondents with *reduced* follow-up, and the alternative CMME has a lower effective sample size because its variance (and thus design effect) is higher. (We note here that at even lower levels of subsampling for *full* follow-up, it may be that the alternative CMME would have a lower variance than the CMME; this is an avenue for future exploration.)

Using the CMME can be used to lead to cost savings for a given effective sample size or to increased precision for the same cost. The effective sample size using the CMME drops as the subsampling fraction decreases (Figure 3), but it also costs less since fewer units receive more expensive follow-up (e.g., telephone). Similarly, for a fixed budget, decreases in the subsampling fraction lead to higher effective sample sizes (shown in Mather, Boulet, and Brennan (2024)). In between, it is possible to find a point where somewhat less collection budget is spent at lower levels of telephone follow-up, but the sample is also increased somewhat to maintain the effective sample size.

Figure 3 – Relative bias vs. effective sample size for variable C



5. CONCLUSION

The CMME was developed as an estimator for the CSS in order to deal with a subsample of units receiving less collection effort than the remainder; it is unbiased under correct specification of f_k and q_k. In practice, these values must be estimated. Unlike f_k, q_k can be modelled using *respondent data* from the *full* and *reduced* follow-up groups, rather than using only frame-level data that is also available for non-respondents; this opens up the potential of a richer and more accurate model that allows for a reduction of bias. Indeed, the CMME was selected over other estimators for the following desirable properties:

- It leverages the lower bias amongst respondents who receive more (and more expensive) follow-up by taking advantage of richer survey data to model response propensities.
- It has a lower variance than other estimators with similar levels of bias. In other words, it has higher effective sample sizes allowing for more precise or granular estimates.

This work, along with the relative costs of various follow-up activities, suggests that there could be substantial cost savings to intentionally limiting the proportion of the sample who receive more expensive follow-up without unduly compromising the level of bias in estimates. Further, this research suggests that using the CMME could also be used to reduce the overall bias compared to spreading the same collection budget over a larger group, under the assumption that respondents from a subsample who receive deeper follow-up are more representative of the target population than respondents who receive *reduced* follow-up.

REFERENCES

Beaumont, J.-F., Bocci, C. and Hidiroglou, M. (2014). On weighting late respondents when a follow-up subsample of nonrespondents is taken. Paper presented to the Advisory Committee on Statistical Methods, Statistics Canada, May 2014, Ottawa.

Hansen, M. H., and Hurwitz, W. N. (1946). The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association*, 41(236), 517–529.

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

Mather, A., Boulet, C. and Brennan, A. (2024). An estimator for concurrent use of full and reduced collection effort on random subsamples. Paper presented to the Advisory Committee on Statistical Methods, Statistics Canada, May 2024, Ottawa.

Neusy, E., Beaumont, J.-F., Yung, W., Hidiroglou, M. and Haziza, D. (2022). Non-response follow-up for business surveys. *Survey Methodology*, 48(1), 95-117.