BANK OF CANADA
BANQUE DU CANADA

Technical Report No. 109 / Rapport technique nº 109

The Bank of Canada 2015 Retailer Survey on the Cost of Payment Methods: Calibration for Single-Location Retailers

by Heng Chen and Q. Rallye Shen

www.bank-banque-canada.ca

March 2017

The Bank of Canada 2015 Retailer Survey on the Cost of Payment Methods: Calibration for Single-Location Retailers

Heng Chen and Q. Rallye Shen

Currency Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
hchen@bankofcanada.ca

ISSN 1919-689X

Acknowledgements

We thank Jean-François Beaumont, Shelley Edwards, Ben Fung, David Haziza, Kim P. Huynh, May Liu, Alan Roshwalb and Angelika Welte for their useful comments and encouragement. Maren Hansen provided excellent writing assistance.

Abstract

Calibrated weights are created to (a) reduce the nonresponse bias; (b) reduce the coverage error; and (c) make the weighted estimates from the sample consistent with the target population in terms of certain key variables. This technical report details our calibration analysis of single-location retailers for the Retailer Survey on the Cost of Payment Methods. We first compare two types of calibration approaches, consisting of (1) traditional calibration, in which calibration is implemented after explicit nonresponse modelling, and (2) nonresponse-embedded calibration, where the nonresponse correction is automatically built in (Särndal and Lundström, 2005). After carefully selecting auxiliary variables, we find minor differences between these two methods. We also examine the effects of trimming, sample size, smoothing and influential units on the calibrated weights, and show that our calibration is robust in view of these considerations.

Bank topics: Econometric and statistical methods; E-money
JEL codes: C81; C83

Résumé

Des pondérations sont estimées par calage pour a) réduire le biais attribuable aux non-réponses, b) réduire l’erreur de couverture et c) faire concorder les poids d’estimation de l’échantillon avec ceux de la population cible pour certaines variables d’intérêt. Dans ce rapport technique, nous décrivons notre analyse de l’ajustement des poids appliqués aux détaillants indépendants dans l’enquête sur les coûts des différents modes de paiement pour les détaillants. Nous comparons dans un premier temps deux méthodes de calage. Dans l’une, le calage se fait après une modélisation explicite des non-réponses (méthode classique), alors que, dans l’autre, la correction des non-réponses est automatiquement intégrée (calage avec prise en compte préalable des non-réponses selon Särndal et Lundström, 2005). Après avoir soigneusement sélectionné les variables auxiliaires, nous relevons entre les deux approches des différences mineures. Nous examinons également l’incidence que peuvent avoir la délimitation du facteur d’ajustement, la taille de l’échantillon, le lissage et les unités influentes sur les poids d’estimation, et nous montrons que notre calage reste valide.

Sujets : Méthodes économétriques et statistiques; Monnaie électronique
Codes JEL : C81; C83

1 Introduction

In 2015, the Bank of Canada conducted the Retailer Survey on the Cost of Payment Methods to collect information on retailers' costs of accepting certain payment methods at the point of sale (see Kosse, et al. (2017) for a detailed description and key findings). The retailers are divided into two groups: single-location and headquarter/chain retailers. For us to estimate average and total costs, the retailers surveyed must be representative of the population of Canadian retailers. This report produces weights for the constructions of a representative sample of single-location retailers, which involves the following three steps: (1) Welte (2017) computes the inclusion probabilities (IPs) from the sampling design; (2) Hatko (2017) computes the response probabilities (RPs) by modelling nonresponse (NR) behaviour; and (3) Chen and Shen (this report) use either IP or RP-adjusted IP as initial inputs for calibration and produce calibrated weights. Figure 1 provides a visual summary of the production of the weights.

The calibrated weights we create are intended to (a) reduce the NR bias; (b) reduce the coverage error; and (c) make the weighted estimates from the sample consistent with the target population with respect to certain key variables. To achieve these objectives for the single-location sample, we first compare two calibration approaches, consisting of (1) the traditional approach, in which the calibration is implemented after explicit NR modelling, and (2) NR-embedded calibration, where the NR correction is automatically built in (Särndal and Lundström, 2005). After carefully selecting auxiliary variables, we find minor differences between these two methods. We also examine the effects of trimming, sample size, smoothing and influential units on the calibrated weights, and show that our calibration is robust in view of these considerations. Figure 2 shows the structure of the calibration process.

This report is organized as follows: Section 2 provides details on applying calibration to reduce both the NR bias and coverage error; Section 3 features a comprehensive discussion on choosing the auxiliary information for calibration; Section 4 presents the empirical results from various calibration methods, differing by whether the NR bias is corrected within the calibration or not, and whether different types of auxiliary information are used; and Section 5 shows that our calibrated results are robust to different modifications in the calibration approach, such as trimming, smoothing, and accounting for influential units. Section 6 discusses two future projects. Appendix A provides information on two Stata commands used to produce results in this report: *sreweight* (Pacifico, 2014) and *ipfraking* (Kolenikov, 2014).

2 Nonresponse and Coverage Error

In this report, we focus on two key study variables: a continuous variable, Cash at Hand, which is the amount of the cash holding at the start of the typical business day, and a categorical variable, Accepts Credit Card, which has value 1 if a single-location retailer accepted credit card payments in 2014, and 0 otherwise. We consider a finite population U (of Canadian single-location retailers) indexed $k = 1, 2., ..., N$. A probability sample s is drawn from U with a known sampling design $p(s)$. In this cost study, we assume the sampling design is an approximation to probability sampling, where every element k in the population U has a non-zero IP $\pi_k > 0$ of being selected into the sample s, giving rise to design weight $d_k \equiv 1/\pi_k$.

As in most surveys, NR occurs, so that a response set r is realized as a subset of s and we have $r \subseteq s \subseteq U$. The response set r arises when the designated sample s is exposed to an unknown response distribution $q(r|s)$ where q refers to the response mechanism, such that unit k has an unknown RP θ_k, assumed positive. Refusals, out-of-business notifications, and incorrect mailing addresses are also categorized as NRs in this cost study. Then θ_k can be viewed more generally as the probability that the value of the study variable y_k is recorded for the unit $k \in s$. With probability $1 - \theta_k$, the value y_k is unobserved. Therefore, the recorded data include both the value y_k and the outcome of the response with $R_k = 1$ for $k \in s$, for $R_k = 0$ for $k \in s - r$. We also have $E_q(R_k|s) = \theta_k$ for $k \in s$.

If we assume that the only error is sampling error, meaning that all units selected for the sample s provide the desired information (no NR), that they respond correctly and truthfully (no measurement error) and, further, that the frame population agrees with the target population (no coverage error),

Figure 1: Weighting procedure

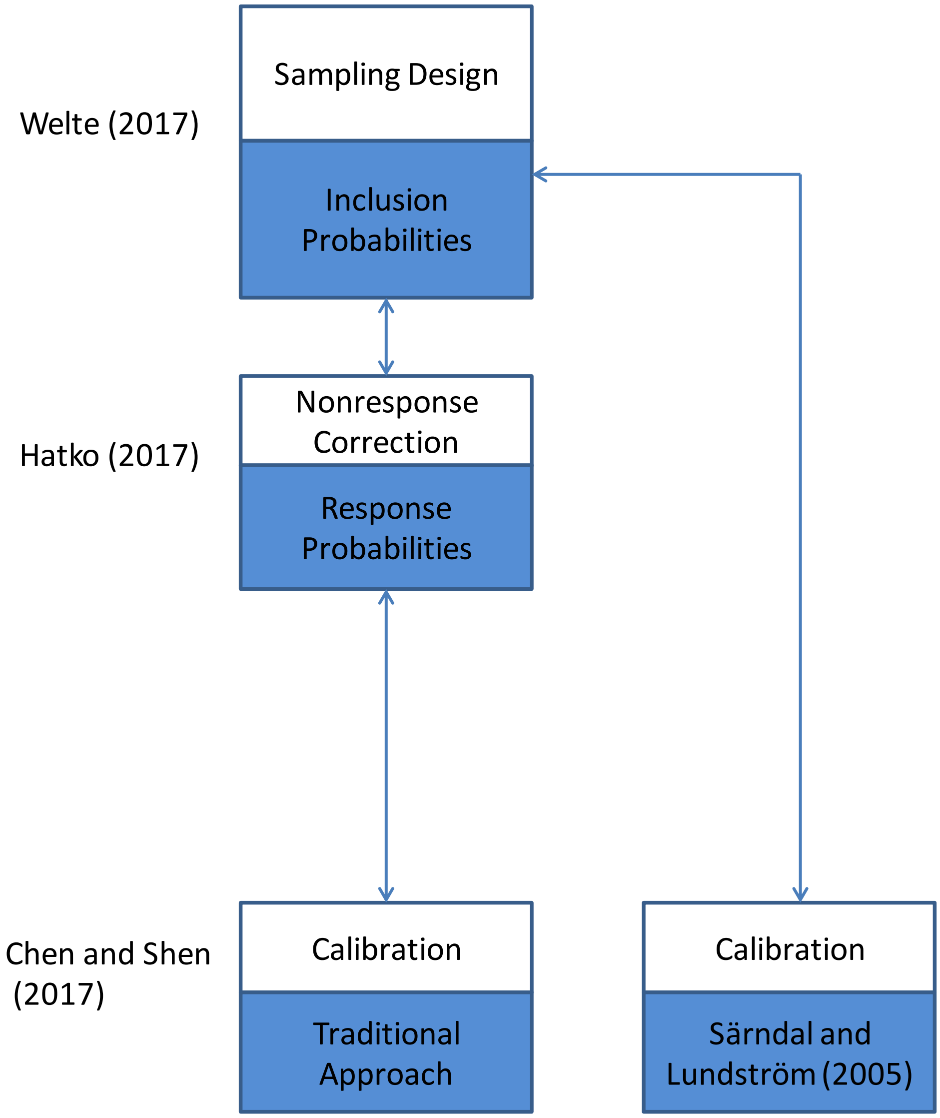
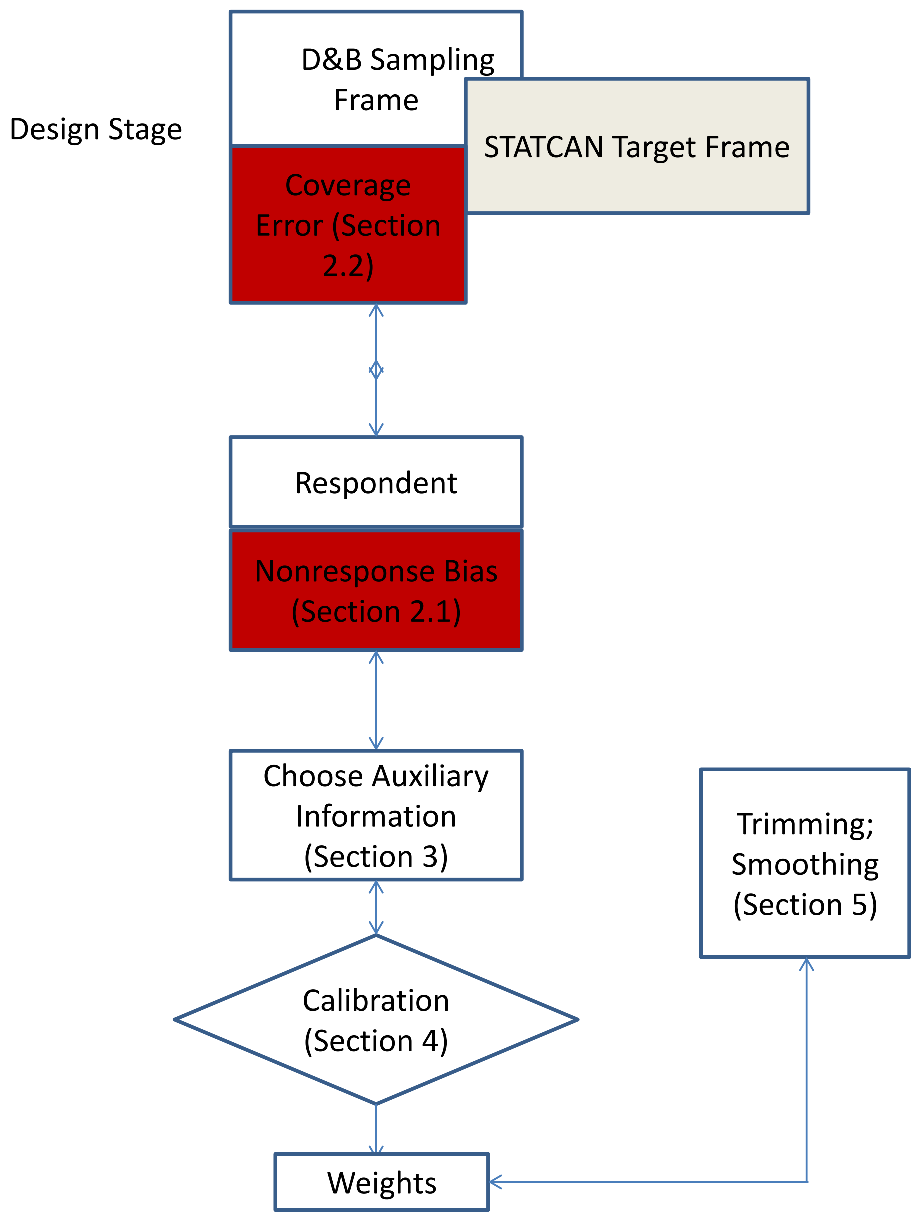


Figure 2: Calibration process for the Single-location sample



then, under these ideal conditions, we can use the Horvitz-Thompson (HT) estimator

$$\widehat{Y}_{HT} \equiv \sum_{s} d_k y_k.$$

Calibration is a systematic approach to using auxiliary variables to improve estimation. For a variable x to qualify as an auxiliary variable, we must know more than x_k for $k \in s$. We need to know the total $\sum_U x_k$. There are two levels of auxiliary information: (1) information at the population level x_k^* (star information), which is usually taken from sampling frame or population registers; and (2) information at the sample level x_k^o (moon information), which is known for every k in s so that $\sum_s d_k x_k^o$ is an unbiased estimate of the population total that is not damaged by NR. The full vector of auxiliary information is

$$X \equiv \begin{pmatrix} \sum_U x_k^* \\ \sum_s d_k x_k^o \end{pmatrix}.$$

In our cost study, the population totals $\sum_U x_k^*$ are obtained from either the Dun & Bradstreet (D&B) sampling frame or the Statistics Canada (STATCAN) Business Register, while the information x_k^o is gathered during the data collection process and includes variables such as *letter type* (type of invitation letter sent) and *CATI recontact* (whether the business was recontacted by phone). In our study we have three information sets, described below as $\{A, B, C\}$:

A Sample (moon) information from the invited sample: *Letter type (none, basic or enhanced), CATI recontact.*

B Population (star) information from the D&B sampling frame: North American Industry Classification System (NAICS) × Region × Size (according to the D&B frame)*, Phone number present, Fax number present, Web address present.* The D&B information is used mainly for correcting for NR bias, and will be discussed in Section 3.

C Population (star) information from the STATCAN target frame: NAICS × Region × Size[1] (according to the STATCAN frame). Here the STATCAN information is used to correct for the coverage error; that is, differences between the D&B sampling frame and STATCAN target frame.

Poststratification to $\{C\}$ is applied using the three-way joint distribution of the variables, while calibration to $\{A\}$ and $\{B\}$ includes marginal distributions. As shown in Table 1, there are no cells with zero or very few respondents across the $\{C\}$ level poststratification cells. This supports our use of poststratification on the joint distribution on $\{C\}$ instead of raking on the three marginals.

2.1 Cost study

Two issues to be addressed on our weighting of the cost study sample are NR bias and coverage error.

Response rates are very low in our cost study (e.g., 2 per cent), as shown in Table 2. High NR has a negative impact on the quality of the statistics produced in a survey, unless powerful adjustment procedures are implemented. The main consequence is that NR bias can become the larger component of the mean squared error (MSE). In addition, NR can also increase the variance because fewer than desired will respond. Here we focus on the unit NR (the selected element does not respond at all), instead of item NR (the selected element responds to some but not all questions on the questionnaire).[2] If the RP θ_k for $k \in r$ is known, the corresponding HT estimator can be revised to

$$\widetilde{Y}_{HT}^{NR} \equiv \sum_{r} d_k \frac{1}{\theta_k} y_k,$$

then NR bias would cease to be a problem and $\widetilde{Y}_{HT}^{NR}$ would be an unbiased estimate. Notice that auxiliary information is not used in the calculation of $\widetilde{Y}_{HT}^{NR}$.

[1]Here Size A in the STATCAN target frame is redefined to be the merged cells of the Indeterminate and Size A business size categories, so that the newly defined business size categories in the STATCAN frame would be closer to that in the D&B frame in terms of the proportion of small businesses.

[2]The nature of NR is treated as non-ignorable in this report.

Table 1: Poststratification adjustment factors from uniform weights, by cell

Poststratification cell	Adjustment factor	Respondents
212	0.222242	10
211	0.224781	19
312	0.245903	24
222	0.28813	13
352	0.379515	55
242	0.388391	13
412	0.402494	11
252	0.448438	12
152	0.490328	56
221	0.500072	21
111	0.511622	18
112	0.539764	17
311	0.556967	9
122	0.598614	25
322	0.599843	23
342	0.615064	25
142	0.645036	33
241	0.647512	17
411	0.679002	23
442	0.740367	18
251	0.833512	17
332	0.841514	34
232	0.848272	11
422	0.885646	10
341	0.917562	13
121	1.026334	19
132	1.124803	32
321	1.145534	10
351	1.184341	15
432	1.203	17
231	1.228478	22
452	1.469169	10
421	1.470459	26
141	1.530681	16
151	1.611546	21
331	1.644274	18
131	2.34154	26
441	2.501821	22
451	2.524679	23
431	3.215469	30

Notes: The first digit of the poststratification cell refers to the NAICS code: 1: 44, 2: 45, 3: 72, 4: 81. The second digit refers to region: 1: Atlantic Region (AT), 2: British Columbia (BC), 3: Ontario (ON), 4: Prairie Region (PR), 5: Quebec (QC). The third digit refers to size of business: 1 - fewer than 5 employees, 2 - 5 to 49 employees. The adjustment factor for each poststratification cell is defined as the ratio of population size to the number of respondents in that cell. For example, the population size is approximately 3.22 times the number of respondents for poststratification cell 431 (81, ON, fewer than 5 employees) is 3.22 times its proportion of respondents.

Table 2: Response rate, by phase

Sample	Response rate
Phase 1	2.5%
Phase 2	2.3%
Phase 1+2	2.5%

Notes: Response rate is defined as the proportion of responding businesses out of all businesses sampled from the sampling frame. Phase 1 and Phase 2 refer to two distinct samples in the cost study, with Phase 2 sampling occurring after Phase 1. The survey methodology differs slightly between the two phases (Welte, 2017).

Coverage errors also occur, since the target frame (STATCAN) is not identical to the sampling frame (D&B). Ideally, the sampling frame should be a perfect match with the target frame; i.e., to the population of all Canadian single-location retailers. This property is essential for the sample to be representative, since it allows for every element in the population to have a non-zero probability of selection, a requirement for unbiased estimation. However, in our cost study, we have two types of coverage errors: *overcoverage* and *undercoverage*. For example, overcoverage might occur as a result of locations in the D&B frame that are in fact out of business, resulting in invalid addresses, while undercoverage might occur because some locations in the STATCAN frame are absent from the D&B frame. If there is only overcoverage and no NR, we can easily modify the HT estimator to be

$$\widehat{Y}_{HT}^{FI} \equiv \sum_{s} a_i d_k y_k,$$

where a_i is a binary variable indicating whether location i is valid ($a_i = 1$) or not ($a_i = 0$). Note the invalid locations are simply excluded from the estimates and the remaining selection probabilities stay unchanged. This leads to unbiased estimates without any model assumptions being made about the probability of sampling a valid element. On the other hand, if there is undercoverage, it becomes difficult to correctly assess the identities of the undercovered businesses. Moreover, when both undercoverage and NR occur, it is impossible to revise the HT estimator without any model assumptions because we cannot determine whether a nonresponding business belongs to the STATCAN frame or to the overcoverage set in the D&B frame.

In this report, we use calibration to reduce as much as possible both NR bias and coverage error. The advantage of the calibration approach is that it brings generality. In the past, a variety of specific estimators were used for surveys, including the Ratio estimator, Weighting Class estimator and Regression estimator. Currently, most of these "conventional techniques" are simply special cases of the calibration approach, differing in the formulation of the auxiliary vector X. It is no longer necessary to consider them separately, since they are all calibration estimators.

In addition to the auxiliary vector X, calibration makes use of a distance function that acts on a set of initial weights. The final calibrated weights are the weights that satisfy the constraints based on X as well as minimize the distance between the initial and final weights as measured by the distance function (Deville et al., 1993).

2.2 Nonresponse correction

In this subsection, we consider two calibration approaches to correcting for NR. The first is traditional calibration, implemented after explicit NR modelling, and the second is NR-embedded calibration, where the NR correction is a part of the calibration itself (Särndal and Lundström, 2005).

Method 1 (traditional calibration) corrects for nonresponse through modelling RPs directly. In principle, such NR adjustments would be appropriate for all the variables of interest, and statisticians can use information on the nonrespondents from the frame to correct for the NR more effectively than information from an external source. On the other hand, Method 2 (Särndal and Lundström, 2005) could also be used to adjust for NR. Indeed, calibration adjustments can be viewed as an inverse RP to some extent: variables related to NR must be chosen along with a distance function so that calibration adjustments are no smaller than 1. In the following sections, we will compare Methods 1 and 2 and show that only minor differences exist between them for our cost study.

2.2.1 Method 1: Traditional calibration

In the traditional approach, the NR is modelled explicitly, and the calibration then serves to correct for the coverage error and reduce the variance of the estimates. Design weights are combined with response weights to construct

$$\widehat{Y}_{HT}^{NR} \equiv \sum_r d_k \frac{1}{\widehat{\theta}_k} y_k,$$

where, in the case of our cost study, θ_k is estimated using a response model from Hatko (2017). The RP θ_k can be computed from either the response homogeneity group (RHG) method or logistic probability model, in which a set of assumptions about the relationship between response indicators and a vector of explanatory variables is made. For the RHG method, the sample s is split into a number of subgroups and then the inverse of the response fraction within a group is used as a weight adjustment to d_k.[3]

Next, if the population totals of the auxiliary variables X are available, we take $d_k/\widehat{\theta}_k$ as initial weights and calibrate to the known population totals X. The estimator from Method 1 is defined as

$$\widehat{Y}_1 \equiv \sum_r d_k \frac{1}{\widehat{\theta}_k} g_{\widehat{\theta}_k} y_k,$$

where $g_{\widehat{\theta}_k}$ is the weight adjustment to satisfy the constraints based on the known totals of X.

In Section 4, we compute weights and their corresponding estimates based on the following three sets of weights:

- $\{IP\} + \{\phi\}$: use only the IP from the sampling design with no NR correction or calibration to any X.
- $\{IP + RP\} + \{\phi\}$: use the IP from the sampling design and then apply the RP for the NR correction, but do not calibrate to any X.
- $\{IP + RP\} + X$: use the IP from the sampling design, then apply the RP for the NR correction, and finally calibrate to the auxiliary information X.

2.2.2 Method 2: Nonresponse-embedded calibration

Adjustment weighting for NR bias with the aid of auxiliary information has been considered by several authors and from diverse angles; e.g., in Bethlehem (1988) and Bethlehem and Schouten (2004); their basic premises mirror Särndal and Lundström (2005). Method 2 does not require explicit modelling of the response, and thus there is no need to estimate θ_k before the calibration. Notice that Method 2 has the same expression as other general calibration estimators. Its main distinguishing feature is the set of chosen auxiliary information X.

The calibration estimator for the auxiliary vector X is defined as

$$\widehat{Y}_2 \equiv \sum_r w_k y_k,$$

where the weight w_k of element k is $d_k \left(1 + \lambda_r^{\intercal} x_k\right)$ and

$$\lambda_r^{\intercal} = \left(X - \sum_r d_k x_k\right)^{\intercal} \left(\sum_r d_k x_k x_k^{\intercal}\right)^{-1}.$$

[3]We thank Jean-François Beaumont for suggesting this alternative RHG method.

When x_k is such that $\mu^\intercal x_k = 1$ for all $k \in U$, then we have a simpler expression

$$\widehat{Y}_2 = \sum_r d_k \left\{ X^\intercal \left(\sum_r d_k x_k x_k^\intercal \right)^{-1} x_k \right\} y_k.$$

After performing calibration, we have $\sum_r w_k x_k = \sum_U x_k$ and $\sum_r w_k = N$. If there is no NR, the bias arising from calibration is very small and decreases faster with sample size than the standard deviation does. However, if NR occurs, $\widehat{Y}_2$ will be more or less biased depending on the strength of the auxiliary vector. In spite of "best efforts" in calibration, some NR bias will always remain in $\widehat{Y}_2$. This NR bias must be dealt with carefully, because the squared bias component often dominates the MSE. Unlike the variance, the NR bias does not approach 0 as the sample size grows to infinity.

In this report, we discuss three estimators based on Method 2 that use the auxiliary variables $x_k^* = \{B, C\}$ and $x_k^o = \{A\}$ and the calibration equations $X \equiv (\sum_U x_k^*, \sum_s d_k x_k^o)^\intercal$ to compute w_k.[4] When implementing Method 2 the calibration could be carried out in one of three ways: One-Step, Multi-Step A or Multi-Step B. In the One-Step approach, the IP is calibrated to $\{A, B, C\}$ totals in one step. In Multi-Step A, the IP is calibrated to $\{A\}, \{A, B\}$ and $\{A, B, C\}$ totals sequentially, with each step including the totals from the previous step as constraints. Finally, in Multi-Step B, the IP is calibrated to $\{A\}, \{B\}$ and $\{C\}$ totals sequentially. In general, these three approaches give different weights, but with only minor differences. Hence, Multi-Step B is used for the following empirical results:

- $\{IP\} + \{A, B\}$: *IP* is calibrated to $\{A, B\}$ totals.
- $\{IP\} + \{A, B, C\}$: *IP* is calibrated to $\{A, B, C\}$ totals.
- $\{IP\} + \{B, C\}$: *IP* is calibrated to $\{B, C\}$ totals.

2.2.3 MSE of Method 2

We will look into the MSE of Method 2 to evaluate the impact of NR on the bias and variance.

First we have

$$\begin{aligned} & \widehat{Y}_2 - Y \\ = & \underbrace{\left(\widehat{Y}_{HT} - Y\right)}_{\text{sampling error}} + \underbrace{\left(\widehat{Y}_2 - \widehat{Y}_{HT}\right)}_{\text{NR error}}. \end{aligned}$$

For the bias of $\widehat{Y}_2$:

$$\begin{aligned} B_{pq}(\widehat{Y}_2) &= E_{pq}(\widehat{Y}_2) - Y \\ &= E_p\left(\widehat{Y}_{HT} - Y\right) + E_{pq}\left(\widehat{Y}_2 - \widehat{Y}_{HT}\right) \\ &= B_{SAM} + B_{NR} \\ &= B_{NR}, \end{aligned}$$

since B_{SAM} is 0 by the unbiasness of the HT estimator in the presence of complete response; moreover, $B_{SAM} \equiv E_p\left(\widehat{Y}_{HT} - Y\right)$ and $B_{NR} \equiv E_{pq}\left(\widehat{Y}_2 - \widehat{Y}_{HT}\right)$.

For the variance of $\widehat{Y}_2$:

$$\begin{aligned} V_{pq}(\widehat{Y}_2) &= E_{pq}\left(\widehat{Y}_2 - E_{pq}\left(\widehat{Y}_2\right)\right)^2 \\ &= E_p\left(\widehat{Y}_{HT} - E_p\left(\widehat{Y}_{HT}\right)\right)^2 + E_p V_q\left(\widehat{Y}_2|s\right) + V_p\left(B_{NR|s}\right) + 2Cov_p\left(\widehat{Y}_{HT}, B_{NR|s}\right) \\ &= V_{SAM} + V_{NR}, \end{aligned}$$

[4]We do not explore the use of model-assisted calibration for binary/discrete y_k as in Wu and Sitter (2001).

where $B_{NR|s} \equiv E_q\left(\widehat{Y}_2 - \widehat{Y}_{HT}|s\right)$ (conditional NR bias), and $V_{SAM} \equiv E_p\left(\widehat{Y}_{HT} - E_p\left(\widehat{Y}_{HT}\right)\right)^2$ and $V_{NR} \equiv E_pV_q\left(\widehat{Y}_2|s\right) + V_p\left(B_{NR|s}\right) + 2Cov_p\left(\widehat{Y}_{HT}, B_{NR|s}\right)$.

We obtain the MSE of $\widehat{Y}_2$ as

$$\begin{aligned} MSE_{pq}(\widehat{Y}_2) &= V_{pq}(\widehat{Y}_2) + \left(B_{pq}(\widehat{Y}_2)\right)^2 \\ &= V_{SAM} + E_pV_q\left(\widehat{Y}_2|s\right) + E_p\left(B^2_{NR|s}\right) + 2Cov_p\left(\widehat{Y}_{HT}, B_{NR|s}\right) \\ &\approx V_{SAM} + E_pV_q\left(\widehat{Y}_2|s\right) + E_p\left(B^2_{NR|s}\right) \text{ if the } Cov \text{ term is small.} \end{aligned}$$

Notice that $E_p\left(B^2_{NR|s}\right)$ can be a very large component of MSE, and Method 2 is designed to reduce the $E_p\left(B^2_{NR|s}\right)$ term by utilizing the auxiliary information X. In addition, note that the bias of $\widehat{Y}_2$ depends jointly on the known sampling design $p(s)$ and unknown response distribution $q(r|s)$.

2.3 Coverage error correction

We also use calibration to correct for coverage error when the sampling frame (D&B) does not completely agree with the target frame (STATCAN). The existing solution is to follow Särndal and Lundström (2005) and Angsved (2006) by estimating either indirectly from the persistor total, or directly from the target population total. However, their method requires identification of the persistors between the sampling and target populations, which would rely on a comparison of every record in the D&B and STATCAN frames — this is impossible in our cost study. Hence, we propose calibrating to the STATCAN frame to make the final total estimates consistent with the STATCAN numbers, so that at the aggregated stratum level, there is no coverage error. However, because of the limitations of the data, we cannot correct for coverage error at the individual level; i.e., for each single-location retailer.

For this purpose we compare the estimates with and without calibration to the target frame (STATCAN): $\{IP\} + \{A, B\}$ versus $\{IP\} + \{A, B, C\}$.

3 Choices of Auxiliary Information

Effective weighting adjustment for NR requires powerful auxiliary information. The desirable features of an auxiliary vector should explain both the response patterns and the study variable(s) in the survey. The weights in the calibration estimator are computed based on information about a specified auxiliary vector. However, even with the "best possible" auxiliary vector, some bias remains in the estimator. A close approximation to the remaining bias, *nearbias*, is presented following Särndal and Lundström (2005). The many potential auxiliary variables allow for a wide variety of possible auxiliary vectors. Therefore, we must compare these vectors to assess their effectiveness for NR bias reduction.

A candidate calibration variable is (a) chosen to be x^* (star information) or x^o (moon information), and then (b) chosen from potential candidates $\{x^*_1, ..., x^*_J\}$ and $\{x^o_1, ..., x^o_M\}$. For a variable x to qualify as star information, $\sum_U x^*_k$ should be known, and in our study star information includes all the variables in information sets $\{B\}$ and $\{C\}$. In order to meet the criterion of being moon information, $\sum_s d_k x^o_k$ needs to be an unbiased estimate and this implies that only the variable *letter type* in $\{A\}$ can be used. This is because the variable *CATI recontact* in $\{A\}$ is adaptively sampled in Phase 2 (Welte 2017) so that its design-weighted total is biased.

Conditional on the available variables $\{A, B, C\}$ selected from the star and moon information, we further select auxiliary variables to reduce the NR bias. First we introduce an ideal case where the "best auxiliary" variables completely remove the NR bias, which sheds light on how auxiliary information can remove the NR bias in extreme cases. However, it is impossible to completely remove the NR bias, given that the RP is genuinely unknown, so we discuss the practical guidelines for maximally reducing the NR bias based on the *nearbias*. Recall from the discussion in Section 2.2.3 that we should choose X based on the NR bias $E_{pq}(\widehat{Y}_2) - \widehat{Y}_{HT}$. This exact bias does not tell us much, but is closely approximated by a much more informative quantity called $nearbias(\widehat{Y}_2)$, for which we have

$$\begin{aligned} E_{pq}(\widehat{Y}_2) - \widehat{Y}_{HT} &\approx nearbias(\widehat{Y}_2) \\ \equiv -\sum_U (1-\theta_k)e_{\theta k}, \end{aligned}$$

with $e_{\theta k} = y_k - x_k^\intercal \left(\sum_U \theta_k x_k x_k^\intercal\right)^{-1} \sum_U \theta_k x_k y_k$. Notice that the expression of $nearbias(\widehat{Y}_2)$ is valid for any sampling design and any auxiliary vector. Furthermore, the *nearbias* formula makes no distinction between the star and moon information. In other words, for bias reduction, the x variable is as important when it carries information at the population level (x_k^*) as when it carries information at the sample level (x_k^o).

3.1 Completely Removing NR *Nearbias*

Principle 1 for completely removing NR bias: The auxiliary vector explains **perfectly** the inverse RP; that is,

$$\begin{aligned} nearbias(\widehat{Y}_2) &= 0 \text{ if for all } k \text{ in } U \\ \frac{1}{\theta_k} &= 1 + \lambda^\intercal x_k \text{ for some constant vector } \lambda. \end{aligned}$$

If Principle 1 is fulfilled, then the nearbias of the calibration estimates is removed for **all** study variables y. To illustrate, suppose the available information follows a classification of the population units or sampling units into J mutually exclusive groups. Then $x_k = (\gamma_{1k}, ..., \gamma_{Jk})^\intercal$, where $\gamma_{jk} = 1$ if k belongs to group j and $\gamma_{jk} = 0$ if it does not. By the above result, the *nearbias* is 0 for this x vector if θ_k are constant within groups, such as groups defined by the NAICS code or region for a population of businesses.

Principle 2 for completely removing NR bias: The auxiliary vector explains **perfectly** the main study variables y, that is,

$$\begin{aligned} nearbias(\widehat{Y}_2) &= 0 \text{ if for all } k \text{ in } U \\ y_k &= \beta^\intercal x_k \text{ for some constant vector } \beta. \end{aligned}$$

If Principle 2 is satisfied, then the *nearbias* is removed in the estimates of the main study variables y, and the variance is also reduced. However, the usual survey involves many y variables. To achieve a 0 *nearbias* for each of those would require the residual $y_k - \beta^\intercal x_k$ to be 0 for all the units as well as for all y variables.

Although Principles 1 and 2 are theoretically attractive, they are difficult to achieve in practice because of the nature of "perfectly linearly related."Hence, in the following subsection, we discuss choosing auxiliary variables to minimize the NR bias, instead of completely removing it.

3.2 Maximally reduce NR *nearbias*

A perfect auxiliary vector would be one that completely eliminates the nearbias. No such vector can be counted on in practice. Even the best of auxiliary vectors leave some bias in a calibration estimator. Nevertheless, if estimates are to be produced at all in the survey, we must ultimately settle for one auxiliary vector and use it in the computation of calibrated weights and survey estimates. Once we stray from either Principle 1 or Principle 2, we need to make sure that x is related to both the response outcome R and study variable y. In this sense, we introduce a new indicator, denoted as H_1, which takes into consideration both Principles:

$$H_1 \equiv q^2 \times f(y, x),$$

where H_1 is a product of q^2 (the variance of the predicted influences of $1/\widehat{\theta}_k$ of the responding units) and a factor $f(y, x)$ depending on the relationship between the study variable y and the auxiliary vector X.

Table 3: Response rate, by key calibration variables

Subsample	Response rate
By NAICS	
45	3.4%
44	3.3%
72	1.6%
81	2.6%
By Region	
AT	1.9%
BC	2.1%
ON	2.4%
PR	2.7%
QC	3.3%
By Business Size	
1–4 employees	2.2%
5–49 employees	3.0%

Note: Response rate is defined as the proportion of responding businesses out of all businesses sampled from the sampling frame.

Table 4: Unweighted mean of cash at hand and credit card acceptance, by key calibration variables

	Cash at Hand	Accepts Credit Card
By NAICS		
44	1,206	0.83
45	806	0.75
72	2,412	0.76
81	275	0.63
By Region		
AT	1,148	0.72
BC	1,436	0.75
ON	805	0.76
PR	1,112	0.80
QC	1,713	0.71
By Business Size		
1–4 employees	470	0.56
5–49 employees	1,899	0.90

Note: Cash at Hand is the amount of cash holding at the start of the typical business day in 2014. Accepts Credit Card is the proportion of businesses in the sample that accepted credit cards in 2014.

See Särndal (2008) for details. In this report, we do not compute the above H_1, but rather the summary statistics shown in Tables 3 and 4, to illustrate the magnitudes of q^2 and $f(y, x)$.

Tables 3 and 4 show that the main variables used in calibration (*NAICS code*, *region* and *business size*) are related to both the response rate and the variables of interest. For example, larger businesses with 5 to 49 employees have a higher response rate, have more Cash at Hand and are more likely to accept credit cards than smaller businesses with 1 to 4 employees. Therefore, there is some evidence (as shown in these tables) that the population-level calibration variables have predictive power for the variables of interest, and that calibration to these variables helps reduce the bias from NR.

However, if we consider the sample-level calibration variables, *letter type* and *CATI recontact*, the relationship is not as clear. The type of letter the respondent received is randomized among the sample as an experiment to test the effect of the letter on participation. Since the letter is randomly distributed, it should have no relationship to Cash at Hand or payment method acceptance, even though it may be related to RP. Meanwhile, the *CATI recontact* variable may introduce bias because businesses are recontacted through CATI only if they did not respond to the first contact.

Hence, the "best" auxiliary variables in the cost study are all variables in $\{B, C\}$ (the star information). No variables in $\{A\}$ should be used, since the variable *letter type* in $\{A\}$ has no explanatory power for the study variables y, and the variable *CATI recontact* is ruled out because of $\sum_s d_k x_k^o$ being biased.

4 Cost Study Weights

To evaluate different methods of correcting for the NR bias (Methods 1 versus 2) and coverage errors (using the star information $\{C\}$ or not), we investigate different weighting methods, as shown in Table 5:

- $\{IP\} + \{\phi\}$: This method does not correct for either coverage error or NR bias. No calibration is implemented, since the auxiliary information is the empty set. In addition, there is no attempt to correct for the NR bias via either the RP model (Method 1) or calibration (Method 2). The results from this method are used as a benchmark.

- $\{IP\} + \{RP\} + \{\phi\}$: This method corrects for NR bias via the RP model (Method 1), but no calibration is implemented, since the auxiliary information is the empty set.

- $\{IP\} + \{A, B\}$: This calibration corrects for the NR bias via calibration (Method 2), instead of the RP model (Method 1). The properties of this calibration estimator depend on the auxiliary vector: the strength of the associations between the auxiliary vector $\{A, B\}$ and study variable y as well as between the auxiliary vector $\{A, B\}$ and the response behaviour R. Notice that although we advocate against using $\{A\}$, as discussed in Section 3, we still compute the results to demonstrate the negative effects of including the auxiliary variables in $\{A\}$.

- $\{IP\} + \{RP\} + \{C\}$: We apply the method of using RP adjustments to correct for the NR bias (Method 1), and then calibrate the weights from the product of $\{IP\}$ and $\{RP\}$ to the STATCAN totals in $\{C\}$. The intention of this approach is to first correct for NR bias using $\{RP\}$ adjustments and then mitigate the coverage error via calibration on the STATCAN totals.

- $\{IP\} + \{A, B, C\}$: This approach corrects for NR bias and coverage error simultaneously via calibration, and therefore does not perform any RP adjustment before calibration. All of the auxiliary information from $\{A, B, C\}$ is used, regardless of the properties of auxiliary variables discussed in Section 3.

- $\{IP\} + \{B, C\}$: This method corrects for both the NR bias and coverage error via calibration on the $\{B, C\}$ totals. In contrast to $\{IP\} + \{A, B, C\}$, the auxiliary information in $\{A\}$ is not included, in accordance with the discussion on the choice of auxiliary variables in Section 3: the variable *letter type* in $\{A\}$ has no explanatory power on the study variables y, and the variable *CATI recontact* is ruled out because $\sum_s d_k x_k^o$ is biased.

Table 5: Summary of six weighting approaches, by different calibration variables and response adjustment

	$\{IP\}+\{\phi\}$	$\{IP\}+\{RP\}+\{\phi\}$	$\{IP\}+\{A,B\}$	$\{IP\}+\{RP\}+\{C\}$	$\{IP\}+\{A,B,C\}$	$\{IP\}+\{B,C\}$
Mean of Weights	447	420	661	661	661	661
Standard Deviation of Weights	342	531	1,067	752	977	847
Min Adjustment Factor	1.0	0.2	0.1	0.2	0.1	0.2
Max Adjustment Factor	1.0	10.9	24.2	13.3	18.4	17.9
Cash at Hand Mean	891	895	880	795	836	783
Cash at Hand SE	87	120	114	94	99	80
Accepts Credit Card	0.73	0.68	0.65	0.67	0.65	0.67
Accepts Credit Card SE	0.019	0.028	0.034	0.027	0.032	0.028

Note: When generating weights from $\{A, B, C\}$, the calibration follows Multi-step B where the calibration occurs in sequential steps over the included levels of auxiliary information—first on the moon information $\{A\}$, then on the star information $\{B\}$ from the D&B frame, and finally on star information $\{C\}$ from the STATCAN frame. Both $\{A\}$ and $\{B\}$ totals are normalized to the size of the STATCAN population. Cash at Hand is the amount of cash holding at the start of the typical business day in 2014. Accepts Credit Card is the proportion of businesses in the sample that accepted credit cards in 2014. Cash at Hand SE and Accepts Credit Card SE are the default standard errors computed by Stata, treating the weights as fixed. They are computed without taking into account the effects of sample design or calibration, and are included only as a rough measure for comparing different sets of weights.

Table 6: Estimates for average hourly wages and annual sales

	$\{IP\}+\{\phi\}$	$\{IP\}+\{RP\}+\{\phi\}$	$\{IP\}+\{A,B\}$	$\{IP\}+\{RP\}+\{C\}$	$\{IP\}+\{A,B,C\}$	$\{IP\}+\{B,C\}$	Population
Hourly Wages	13	11	11	12	12	12	16
MSE (Hourly Wages)	10	22	24	20	17	18	-
Annual Sales	982,087	756,975	787,525	708,616	890,160	772,348	632,300
MSE (Annual Sales)	1.3E+11	2.8E+10	4.1E+10	1.4E+10	8.3E+10	2.8E+10	-

Note: Estimated means are for the retail sector (NAICS code 44/45) only. Population means are from the 2014 Survey of Employment, Payrolls and Hours and the 2013 Statistics Canada Small Business Profiles. The mean sales figures are taken from businesses with sales greater than $30,000 and less than $5 million. The sample differs from the basis for the population estimates because our sample includes only businesses with fewer than 50 employees. The *MSE*s are calculated as the sum of the squared deviation from the population statistic and the default Stata variance. Note that the default Stata variances used in calculating the *MSE* do not incorporate the effects of calibration and are rough measures only.

Table 7: Estimates for average hourly wages by region

	$\{IP\}+\{\phi\}$	$\{IP\}+\{RP\}+\{\phi\}$	$\{IP\}+\{A,B\}$	$\{IP\}+\{RP\}+\{C\}$	$\{IP\}+\{A,B,C\}$	$\{IP\}+\{B,C\}$	Population	Minimum Hourly Wage
BC	13.38	12.36	12.01	12.39	12.54	12.84	16.47	10.25
ON	11.63	10.65	12.42	11.43	12.96	11.97	15.89	11.00
QC	13.58	10.96	8.00	9.99	9.15	9.50	16.01	10.35

Note: Means are for the retail sector (NAICS code 44/45) only. Population means are from the 2014 Survey of Employment, Payrolls and Hours. The sample differs from the basis for the population estimates because our sample includes only businesses with fewer than 50 employees. The minimum hourly wages are as of June 2014.

4.1 Discussion

4.1.1 Summary

We recommend both $\{IP\} + \{RP\} + \{C\}$ (Method 1) and $\{IP\} + \{B, C\}$ (Method 2), as only minor differences exist between them in terms of constructed weights and mean estimates (Table 5) and external validations (Tables 6 and 7). This is expected, because both using the $\{RP\}$ adjustment to directly model the response mechanism (Method 1) and calibrating over the information in $\{B\}$ (Method 2) provide similar information in the context of NR behaviour.[5] Furthermore, since $\{IP\} + \{RP\} + \{C\}$ and $\{IP\} + \{B, C\}$ both calibrate over $\{C\}$, both rely on the same auxiliary information for alleviating coverage errors.

4.1.2 $\{\mathbf{IP}\} + \{\mathbf{RP}\} + \{\mathbf{C}\}$ versus $\{\mathbf{IP}\} + \{\mathbf{B}, \mathbf{C}\}$

We first compare $\{IP\} + \{RP\} + \{C\}$ and $\{IP\} + \{B, C\}$ with respect to both NR and coverage error corrections, and observe that there are many similarities between Methods 1 and 2. As shown in Table 5, $\{IP\} + \{RP\} + \{C\}$ and $\{IP\} + \{B, C\}$ produce lower estimates of Cash at Hand with lower standard errors compared with the other weighting schemes. This suggests that including the STATCAN information $\{C\}$, believed to be more accurate than the D&B information $\{B\}$, reduces both the bias and variance of the estimates. Using the STATCAN information lowers the estimate of Cash at Hand because there is a higher proportion of smaller businesses (with fewer than five employees) in the STATCAN totals than in the D&B totals, and smaller businesses tend to hold less cash at hand. Therefore, calibrating to the STATCAN totals $\{C\}$ reduces the bias from the undercoverage of smaller businesses due to the discrepancies between the sampling and target frame (Section 2.3).

Applying weights from $\{IP\} + \{RP\} + \{C\}$ and $\{IP\} + \{B, C\}$ generates the same estimate for the proportion of businesses accepting credit cards, which is six percentage points lower than using $\{IP\} + \{\phi\}$ alone. In fact, all of the weighting approaches that either use RP or apply calibration reduce the estimated proportion compared with $\{IP\} + \{\phi\}$ alone. This effect is probably due to lower response rates for businesses that do not accept credit cards, so the NR correction is crucial to improving estimates compared with using only $\{IP\} + \{\phi\}$.

For external validation, we compare our estimates of the average hourly wage and annual sales to the estimates (which are labelled "Population") from Statistics Canada and Industry Canada (Tables 6 to 7). The second and fourth rows of Table 6 show the MSEs of the estimates from the different weighting approaches with respect to average hourly wage and annual sales. In terms of annual sales, the $\{IP\} + \{RP\} + \{C\}$ method produces the closest estimate to the population estimate. However, the population estimate of the average hourly wages is higher than all of our estimates, possibly because the basis for estimation is different: the population figure includes businesses with more than 50 employees, while the sample estimates do not. Specifically, the probable reason is that our weighted estimates are based on small and medium-sized businesses (with 0 to 49 employees), while the population numbers from external sources are the aggregate of all businesses (unfortunately, the external sources do not allow for disaggregation by number of employees[6]), and the larger businesses tend to pay higher hourly wages. Examining the MSEs for both hourly wages and annual sales, the $\{IP\} + \{RP\} + \{C\}$ and $\{IP\} + \{B, C\}$ approaches seem to generate lower MSEs overall than the other weighting approaches.

Table 7 shows comparisons of mean hourly wage by region against population estimates and the minimum hourly wage in the province. Only BC, ON and QC are shown, since the AT and PR consist of multiple provinces. Our regional hourly wage estimates are lower than the population hourly wage but are still plausible because they are close to the minimum hourly wage in each province, given that our sample focuses on small and medium-sized businesses. Based on the above considerations, we recommend both $\{IP\} + \{RP\} + \{C\}$ and $\{IP\} + \{B, C\}$.

[5]We thank Jean-François Beaumont for providing this rationale.

[6]Note that the ranges of employee counts are based mostly on data derived from payroll remittances. As such, they should be viewed solely as a business stratification variable. The primary purpose is to improve the efficiency of samples selected to conduct statistical surveys. They should not be used in any manner to compile industry employment estimates.

4.1.3 $\{\mathbf{IP}\}+\{\phi\}$ versus $\{\mathbf{IP}\}+\{\mathbf{RP}\}+\{\phi\}$, $\{\mathbf{IP}\}+\{\mathbf{A},\mathbf{B}\}$

Secondly, we compare $\{IP\}+\{\phi\}$ with $\{IP\}+\{RP\}+\{\phi\}$ and $\{IP\}+\{A,B\}$ in order to observe the impact of NR bias correction, since $\{IP\}+\{\phi\}$ involves no correction for NR bias while both $\{IP\}+\{RP\}+\{\phi\}$ and $\{IP\}+\{A,B\}$ correct for the NR bias but not undercoverage, because they do not include $\{C\}$ information in calibration. To highlight some of the effects of correcting for NR bias, we show that the mean estimates of Cash at Hand are in roughly the same range for $\{IP\}+\{\phi\}$, $\{IP\}+\{RP\}+\{\phi\}$ and $\{IP\}+\{A,B\}$, while the estimated acceptances of credit cards show significant differences: 73 per cent from $\{IP\}+\{\phi\}$, compared with 68 per cent from $\{IP\}+\{RP\}+\{\phi\}$ and 65 per cent from $\{IP\}+\{A,B\}$. These results suggest a strong association between response and credit card acceptance and a weaker one between response and cash at hand.

4.1.4 $\{\mathbf{IP}\}+\{\mathbf{RP}\}+\{\phi\}$ versus $\{\mathbf{IP}\}+\{\mathbf{A},\mathbf{B}\}$

Third, we compare $\{IP\}+\{RP\}+\{\phi\}$ and $\{IP\}+\{A,B\}$ to emphasize the similarities between Methods 1 and 2 when correcting for the NR bias. There are only small differences between $\{IP\}+\{RP\}+\{\phi\}$ and $\{IP\}+\{A,B\}$ in the estimated mean Cash at Hand and Accepts Credit Card, and this again supports the claim that both Method 1 and Method 2 perform well for correcting NR bias and are interchangeable in most respects.

4.1.5 $\{\mathbf{IP}\}+\{\mathbf{RP}\}+\{\phi\}$, $\{\mathbf{IP}\}+\{\mathbf{A},\mathbf{B}\}$ versus $\{\mathbf{IP}\}+\{\mathbf{RP}\}+\{\mathbf{C}\}$, $\{\mathbf{IP}\}+\{\mathbf{A},\mathbf{B},\mathbf{C}\}$

We compare $\{IP\}+\{RP\}+\{\phi\}$ and $\{IP\}+\{A,B\}$ with $\{IP\}+\{RP\}+\{C\}$ and $\{IP\}+\{A,B,C\}$ to demonstrate the importance of correcting for the coverage error. Notice that $\{IP\}+\{RP\}+\{\phi\}$ and $\{IP\}+\{A,B\}$ correct only for the NR bias, while $\{IP\}+\{RP\}+\{C\}$ and $\{IP\}+\{A,B,C\}$ correct for NR bias and coverage errors simultaneously. Further, adding $\{C\}$ information to the calibration reduces the estimated means of Cash at Hand by around \$100, although the estimated proportions of credit card acceptance are similar. Given that smaller businesses usually hold less cash than their medium-sized counterparts, the downward estimate from using the STATCAN information $\{C\}$ reflects an adjustment for the underrepresentation of smaller businesses in the D&B frame.

4.1.6 $\{\mathbf{IP}\}+\{\mathbf{A},\mathbf{B},\mathbf{C}\}$ versus $\{\mathbf{IP}\}+\{\mathbf{B},\mathbf{C}\}$

We compare $\{IP\}+\{A,B,C\}$ and $\{IP\}+\{B,C\}$ to analyze the importance of the choice of auxiliary variables. Recall that in Section 3, we discuss the choice of auxiliary information and conclude that $\{A\}$ information should not be included in the calibration. From the last two columns of Table 5, we observe that incorrectly adding $\{A\}$ to the calibration will increase our estimate of Cash at Hand by \$53 compared with not calibrating to $\{A\}$. In addition, another negative impact of including $\{A\}$ is that the standard error becomes larger: 99 versus 80 for Cash at Hand and 0.032 versus 0.028 for Accepts Credit Card.

4.1.7 $\{\mathbf{IP}\}+\{\mathbf{A},\mathbf{B}\}$ versus $\{\mathbf{IP}\}+\{\mathbf{B},\mathbf{C}\}$

Finally, we compare $\{IP\}+\{A,B\}$ with $\{IP\}+\{B,C\}$ to differentiate between the roles of $\{A\},\{B\}$ and $\{C\}$. As the results in Table 5 show, we find that calibrating over $\{B\}$ essentially corrects for the NR bias, while calibrating over $\{C\}$ mainly corrects for the coverage error. However, $\{A\}$ has a negative impact on the calibration estimates, so the moon information contained in $\{A\}$ should not be used for calibration.

5 Extensions

In this section, we analyze the robustness of $\{IP\}+\{B,C\}$ with respect to trimming, sample size, smoothing and influential units. The results below indicate that the calibrated weights from $\{IP\}+\{B,C\}$ are still the best choice when these modifications are considered.

5.1 Trimming

Here we analyze the impacts of trimming, or imposing bounds on the adjustment factor of the calibrated weights. Table 8 shows that imposing a bound on the weights has very little impact on the estimates for Cash at Hand and Accepts Credit Card, and increases the standard error slightly compared with estimates using the unbounded weights. Therefore, there does not seem to be any benefit to trimming or imposing bounds.

Table 8: Unbounded versus bounded weights

	$\{IP\} + \{B, C\}$	Bounded $\{IP\} + \{B, C\}$
Mean	661	661
Standard Deviation	847	869
Min Adjustment Factor	0.2	0.1
Max Adjustment Factor	17.9	10.9
Cash at Hand Mean	783	786
Cash at Hand SE	80	85
Accepts Credit Card	0.67	0.66
Accepts Credit Card SE	0.028	0.029

Note: The trimming weights are created by imposing a bound of 10 on the adjustment factor in both steps of the two-step calibration to $\{B, C\}$. As a result of convergence issues, the distance function is chosen to be the modified chi-square. Cash at Hand is the amount of cash holding at the start of the typical business day in 2014. Accepts Credit Card is the proportion of businesses in the sample that accepted credit cards in 2014.

5.2 Sample size

Increasing the sample size will reduce the variance. Hence, we compare the estimates based on either Phase 1 Only or Phase 1 and 2 Combined, to investigate the effects of the sample size and the two phases on variance reduction. Using Phase 1 only does not appear to have a large impact on estimates for credit card acceptance, but it increases the estimate of Cash at Hand by $85 and doubles the standard error (Table 9). This impact on the size of the standard error supports our use of the combined two phases rather than Phase 1 only.

5.3 Smoothing

Using smoothed initial weights will reduce the variability of calibrated weights, but this may result in misspecification of either the IP or RP.[7] Moreover, smoothing may introduce bias into the estimates and the resulting standard error estimate is also an underestimation of the true standard error (if the smoothed weight is treated as fixed). In a proper smoothing method (Beaumont, 2008), it is necessary to model the weight as a function of the variables of interest and then generate the smoothed weight as the predicted weight from the model. The accuracy of this method depends on how successful we are at modelling the weights. However, efficiency gains are obtained at the expense of a higher risk of increasing bias.

Table 10 shows that starting with smoothed design weights slightly reduces the standard deviation of the final calibrated weights, but not the standard errors of the estimates of Cash at Hand or Accepts Credit Card. Furthermore, the maximum adjustment factor is considerably larger for the smoothed IP, which indicates that smoothing the IP increases the deviation from population totals for some cells. Therefore, there is no evidence, as shown in the table, that smoothing the design weights brings about any improvement in efficiency or bias reduction. The mean estimates are quite similar between the

[7] We thank Jean-François Beaumont for suggesting the pros and cons of smoothing.

Table 9: Phases 1 and 2 combined versus Phase 1 only

	$\{IP\} + \{B, C\}$	
	Combined phases	Phase 1 only
Mean	661	812
Standard Deviation	847	993
Min Adjustment Factor	0.2	0.2
Max Adjustment Factor	17.9	18.9
Cash at Hand Mean	783	868
Cash at Hand SE	80	160
Accepts Credit Card	0.67	0.66
Accepts Credit Card SE	0.028	0.031

Note: Cash at Hand is the amount of cash holding at the start of the typical business day in 2014. Accepts Credit Card is the proportion of businesses in the sample that accepted credit cards in 2014. The *SE*s are the default standard errors generated by Stata. *Combined phases* refers to applying the weighting procedure $\{IP\} + \{B, C\}$ to Phases 1 and 2 combined, with initial weights based on the IPs for both phases (thus, the *Combined phases* column repeats the results from the last column of Table 5). *Phase 1 only* refers to using only Phase 1 observations for calibration and estimation, with initial weights based on the IPs for Phase 1 only.

two methods, which may suggest that our $\{IP\} + \{B, C\}$ calibration is robust to misspecification of the IP (the complexity of the sampling design is outlined in Welte (2017)). Hence, we recommend not smoothing the design weights.

Table 10: Using unsmoothed versus smoothed design weights in calibration

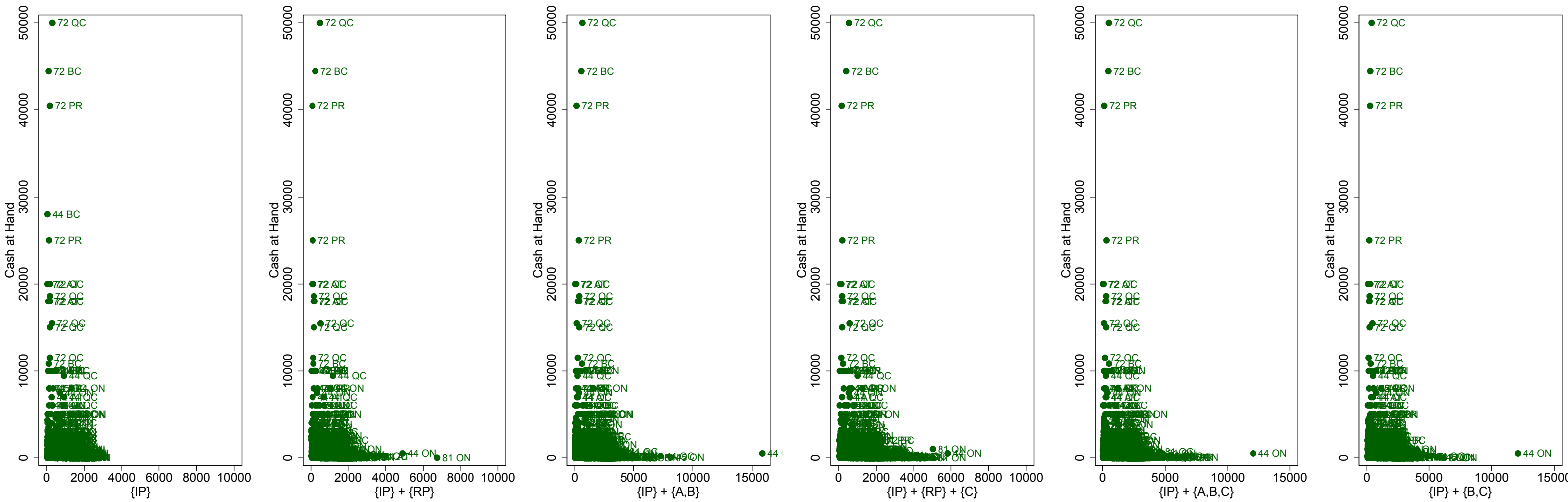
	$\{IP\} + \{B, C\}$	$\{SIP\} + \{B, C\}$
Mean	661	661
Standard Deviation	847	842
Min Adjustment Factor	0.2	0.2
Max Adjustment Factor	17.9	25.2
Cash at Hand Mean	783	818
Cash at Hand SE	80	81
Accepts Credit Card	0.67	0.67
Accepts Credit Card SE	0.028	0.028

Note: Cash at Hand is the amount of cash holding at the start of the typical business day in 2014. Accepts Credit Card is the proportion of businesses in the sample that accepted credit cards in 2014. The *SE*s are the default standard errors generated by Stata. $\{SIP\} + \{B, C\}$ is produced by first averaging the design weights within each poststratification cell to generate "smoothed" design weights SIP, and then using SIP as initial weights in the two-step calibration toward controlling for totals B and C.

5.4 Influential units

Figure 3 shows scatter plots of the six sets of weights considered versus values of Cash at Hand. With regard to influential units (defined as having both a large amount of Cash at Hand and large final weights), we find none lying in the right-top corner of the plots.

Figure 3: Cash at hand versus six different sets of weights



Note: Cash at Hand is the amount of cash holding at the start of the typical business day in 2014. Each data point represents one business that responded to the survey. The points are labelled with the two-digit NAICS code and the region of the respondent.

6 Conclusion

Effective weighting adjustment for NR requires powerful auxiliary information. However, even with the "best possible" auxiliary vector, some bias remains in the estimator. A close approximation to this bias, *nearbias* (Särndal and Lundström, 2005), should be computed and used as a guideline for choosing among a wide variety of possible auxiliary vectors. Thus, a more rigorous quantification of *nearbias* in Section 3.2 is desirable and implemented in future.

There are two other projects worth investigating in the future. First, it will be important to understand how the NR bias is corrected under the NR-embedded calibration. Based on Haziza and Lesage (2016), when auxiliary variables are discrete and used for the poststratification in the NR-embedded method, the NR correction is modelled nonparametrically. On the other hand, if the calibration is performed to match the margins instead of the poststratified cells, the NR-embedded approach implicitly imposes a parametric response model depending on the calibration function. Then the resulting estimators would not be robust to the misspecification of the response model. Hence, empirical studies are needed to investigate these arguments by including more auxiliary variables available from D&B and STATCAN to improve the RP model for the traditional approach, and more poststratified cells for the NR-embedded.

Another future project will be computing the variance of estimates obtained from two calibration approaches, but such computation is a challenging problem given the complicated sampling design and unknown response behaviour. Thus, some simplifying assumptions must be made. For the traditional calibration method, we approximate both sampling design and response behaviour by two separate Poisson distributions, and then use the bootstrap resampling method (Beaumont and Patak, 2012). However, for the NR-embedded calibration, the variance estimation is an open research question.

7 References

Angsved, M. 2006. "Estimating the Finite Population Total under Frame Imperfections and Nonresponse." *Orebro University Working Paper* NO. 4.

Beaumont, J. F., 2008. "A New Approach to Weighting and Inference in Sample Surveys." *Biometrika* 95(3): 539–553.

Beaumont, J. F., & Patak, Z. 2012. "On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling." *International Statistical Review* 80(1): 127–148.

Bethlehem, J. G. 1988. "Reduction of Nonresponse Bias through Regression Estimation." *Journal of Official Statistics* 4(3): 251–260.

Bethlehem, J., & Schouten, B. 2004. "Nonresponse Analysis of the Integrated Survey on Living Conditions (POLS)." *Statistics Netherlands Discussion Paper* 04004.

Deville, J. C., Särndal, C. E., & Sautory, O. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American statistical Association* 88(423): 1013–1020.

Hatko, S. 2017. "The Bank of Canada 2015 Retailer Survey on the Cost of Payment Methods: Nonresponse." *Bank of Canada Technical Report* No. 107.

Haziza, D., & Lesage, E. 2016. "A Discussion of Weighting Procedures for Unit Nonresponse." *Journal of Official Statistics* 32(1): 129–145.

Kolenikov, S. 2014. "Calibrating Survey Data Using Iterative Proportional Fitting (Raking)." *Stata Journal* 14(1): 22–59.

Kosse, A., & Chen, H., & Felt, M. H., & Jiongo, V. J., & Nield, K., & Welte, A. 2017. "The Costs of Point-of-sale Payments in Canada." *Bank of Canada Staff Discussion Paper* No. 2017-4.

Pacifico, D. 2014. "SREWEIGHT: STATA Module for Survey Reweighting." *Statistical Software Components.*

Särndal, C. E., & Lundström, S. 2005. *Estimation in Surveys with Nonresponse.* John Wiley & Sons.

Särndal, C. E. 2008. "Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator." *Journal of Official Statistics* 24(2): 167.

Welte, A. 2017. "The Bank of Canada 2015 Retailer Survey on the Cost of Payment Methods: Sampling." *Bank of Canada Technical Report* No. 108.

Wu, C., & Sitter, R. R. 2001. "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data." *Journal of the American Statistical Association* 96(453): 185–193.

Appendix A Software

Two different **Stata** calibration commands:

1. *sreweight*: Pacifico, D. 2014. “SREWEIGHT: STATA Module for Survey Reweighting.” *Statistical Software Components.*

2. *ipfraking*: Kolenikov, S. 2014. “Calibrating Survey Data Using Iterative Proportional Fitting (Raking).” *Stata Journal* 14(1): 22–59.

Remarks:

1. Both commands are very flexible in accommodating different options for weight trimming, detecting weight outliers, and specifying the auxiliary information equations and the number of iterations for achieving convergence. Both commands allow for generating replicate weights for variance estimation.

2. “sreweight” has more options that can be adjusted to ensure convergence of the calibration procedure (e.g., repeated trials with randomized starting points).

3. “sreweight” has more choices with respect to objective distance functions available (e.g., chi-square, empirical likelihood, exponential tilting, or Deville and Särndal distance function).

4. Both commands generate identical calibrated weights when calibrating to a three-way poststratification table (by *NAICS Code*, *Region* and *Business Size*), with initial weights equal to the product of design weights (the inverse of the IP) and response adjustment factors (the inverse of the estimated RP).

5. From a preliminary simulation, we find that convergence in both commands is determined by the number of auxiliary equations and the range of trimming, as well as the convergence criteria. For example, when there is no trimming, both commands converge when the number of unknown weights is larger than the rank of auxiliary information equations, and thus they are linearly dependent. On the other hand, when trimming weights, both commands sometimes will not converge because the range of trimming is incompatible with auxiliary information equations. Notice that it is necessary to avoid near-collinearity by excluding unnecessary auxiliary variables. Hence, in order to guarantee the convergence of the calibration, we have to carefully specify both auxiliary information equations and the trimming range: we can either reduce the number of auxiliary information equations (e.g., collapsing some region cells) or enlarge the trimming range (e.g., only trimming the 99th percentile of weights instead of the 95th percentile).