BANK OF CANADA
BANQUE DU CANADA

Technical Report/Rapport technique — 118

Last updated: August 11, 2020

Sample Calibration of the Online CFM Survey

by Marie-Hélène Felt[1] and David Laferrière[2]

[1]Marie-Hélène Felt
Currency Department
Bank of Canada, Ottawa, Ontario, Canada K1A 0G9
mfelt@bankofcanada.ca,

[2]David Laferrière
Statistics Canada, Ottawa, Ontario, Canada K1A 0T6
david.laferriere2@canada.ca

ISSN 1919-689X

Acknowledgements

We would like to thank Heng Chen, Kim P. Huynh and Jean-François Beaumont for their comments and suggestions. We acknowledge Statistics Canada's Methodology Branch Micromission Program that made this collaboration possible, and in particular Susie Fortier and Pierre Caron (Statistics Canada), and Michele Sura and Alison Layng (Bank of Canada) for actively supporting this initiative. We acknowledge Michael Hsu and the Ipsos Reid team for their collaboration. We thank Maren Hansen for careful proofreading.

Abstract

The Bank of Canada's Currency Department has used the Canadian Financial Monitor (CFM) survey since 2009 to track Canadians' cash usage, payment card ownership and usage, and the adoption of payment innovations. A new online CFM survey was launched in 2018. Because it uses non-probability sampling for data collection, selection bias is very likely. We outline various methods for obtaining survey weights and discuss the associated conditions necessary for these weights to eliminate selection bias. In the end, we obtain calibration weights for the 2018 and 2019 online CFM samples. Our final weights improve upon the default weights provided by the survey company in several ways: (i) we choose the calibration variables based on a fully documented selection procedure that employs machine learning techniques; (ii) we use very up-to-date calibration totals; (iii) for each survey year we obtain two sets of weights, one for the full yearly sample of CFM respondents, the other for the sub-sample of CFM respondents who also filled in the methods-of-payment module of the survey.

Topics: Econometric and statistical methods

JEL codes: C81, C83

Résumé

Depuis 2009, le département de la Monnaie de la Banque du Canada se sert de l'enquête Canadian Financial Monitor (CFM) pour évaluer l'utilisation de l'argent comptant, la détention et l'utilisation de cartes de paiement ainsi que l'adoption d'innovations en matière de paiement par les Canadiens. Une nouvelle enquête CFM en ligne a été lancée en 2018. Le mode de collecte des données de cette enquête repose sur l'échantillonnage non probabiliste, qui induit très probablement un biais de sélection. Nous présentons plusieurs méthodes de pondération et analysons les conditions requises pour que ces pondérations puissent éliminer le biais de sélection. Nous obtenons ainsi des poids de calage pour les échantillons des enquêtes en ligne de 2018 et 2019. Nos pondérations définitives constituent une amélioration par-rapport à celles fournies par la société de sondage, et ce, de plusieurs façons : 1) nous choisissons les variables de calage selon une méthode de sélection entièrement étayée qui fait appel à des techniques d'apprentissage automatique; 2) nous utilisons des totaux de calage tout à fait à jour; 3) pour chaque année de l'enquête CFM, nous obtenons deux séries de pondérations – l'une pour l'échantillon annuel complet de répondants, et l'autre pour le sous-échantillon des répondants qui ont également rempli le module sur les modes de paiement.

Sujets : Méthodes économétriques et statistiques

Codes JEL : C81, C83

1 Introduction

The Canadian Financial Monitor (CFM) survey is a wealth survey conducted by Ipsos, a market research firm, since 1999. It is a syndicated survey with multiple stakeholders, including some large financial institutions and the Bank of Canada. The Currency Department of the Bank of Canada has utilized the data since 2009, when a section on methods of payment was added to the paper-based survey. It has served to track cash holdings and usage, payment card ownership and usage, and the adoption of payment innovations by Canadians.

The CFM survey was initially a household survey with a paper-based questionnaire. This offline CFM was discontinued at the end of 2018. In January of the same year a new online, individual-based survey was launched. A methods-of-payment (MP) module was added to the questionnaire in April 2018.

This paper's primary goal is to propose an adequate weighting procedure for the yearly online CFM survey sample, given the non-probability nature of its sampling process. The main motivation is that, when weighting the full yearly sample instead of quarterly subsamples, a richer set of calibration variables can be used in the weighting procedure.[1] Using a richer set of calibration variables allows for the derivation of survey weights that are more likely to reduce the potential selection bias in such a non-probability survey. In the end, we obtain final sets of weights for the 2018 and 2019 online CFM samples and weights for the 2018 and 2019 MP modules.[2] The MP series of weights are specifically for calculating estimates for the responses to the MP module, and they are not intended to be used for other modules.

Our contribution is threefold: (i) we fully justify the calibration variables chosen and base our selection procedure on machine learning techniques; (ii) we improve upon the timeliness of the population totals used as targets; (iii) we calibrate not only the full yearly sample of

[1] The weights provided by Ipsos are quarterly weights calibrated on four basic demographic characteristics; see details in Section 2.

[2] We focus on the weighting of the 2018 online CFM survey in the main body of this paper. That of the 2019 online CFM is presented in Appendix B.

CFM respondents, but also the subsample of respondents that filled in the MP module of the questionnaire.

This report is organized as follows. We provide additional details about the online CFM survey in Section 2. Next, we discuss potential approaches for the weighting of non-probability surveys and provide a detailed description of the calibration approach we adopt for the online CFM survey in Section 3. In Section 4, we outline the procedure implemented for the selection of calibration variables. Finally, we conclude with a description and analysis of the final sets of weights in Section 5.

2 The new online CFM survey

Figure 1 presents a timeline of the developments regarding the paper-based and online CFM surveys in 2018 and 2019. The reasons stated by the survey company for a conversion of the CFM to an online survey were the opportunity of larger sample sizes and faster data availability, as well as greater flexibility and expandability offered by an online survey (Ipsos 2018).[3] The change of survey mode was also accompanied by a change in the sampling and observation units. The sampling unit and main unit of observation went from the household in the paper-based CFM to the individual respondent in the online CFM.[4]

The initial version of the online CFM did not collect any information on methods of payment. However, the Currency Department of the Bank of Canada added the MP module to the questionnaire in April 2018. This module contains questions on cash management and the use of different payment methods. Although the main unit of observation in the 2018 online survey is the individual respondent ("How many chequing account(s) do you currently hold?"; "In the past month, have you personally withdrawn cash?"), most questions in the

[3]The paper-based CFM data has an interesting panel dimension due to respondents participating several years in a row; see Chen et al. (2017). However, the panel dimension of the online CFM data is unknown due to the lack of information for ensuring General Data Protection Regulation (GDPR) compliance.

[4]This change implies that meaningful direct comparisons between the two will likely not be possible. However, as the online CFM continues, it will be possible to observe if past trends observed in the offline survey continue in the online version; see Appendix C. Also, the overlap in 2018 between the household and individual CFMs could be exploited to analyze intra-household behaviours, following Felt (2018).

MP module collect information at the household level, e.g., "How many times in the past month did your household use cash to make purchases?" This was modified in February 2019 when the unit of observation was aligned with the sampling unit (the individual respondent) in the whole MP section. Only a subsample of CFM respondents are requested to fill in the MP module. Of the 18,005 respondents to the 2018 online CFM, 12,004 also completed the MP module.

The 2018 online CFM respondents were selected using a quota sample, where the quotas are proportions determined by household income range (five categories), respondent age range (six categories) and respondent gender within each region (Atlantic, Quebec, Ontario, Prairies and British Columbia). The proportion targets Ipsos uses for age and gender are taken from the 2016 Census, and the target for household income is taken from the 2011 National Household Survey. These quotas are in contrast with those used for the offline CFM, which had sample quotas based on household income and size, home ownership, age and employment status of household head, province and city size.

The CFM data sets received by the Bank of Canada include survey weights, which we refer to as the *default weights*. The default weights are calculated by calibrating (via iterative raking) the quarterly subsamples on respondent age, gender and household income, within five geographic regions.[5] The default weights are calculated on proportions (rather than overall population totals), and each unit starts with an initial weight of 1, so the default weights average to 1. Due to the quota sampling methodology, the composition of the raw sample of respondents is very close to the calibration proportions from the population. Consequently, the default weights are very close to 1 for most units (90% of units have weights between 0.97 and 1.03). Finally, the default weights are provided for 18,005 CFM respondents, and there is not a second set of weights for only those 12,004 individuals who were selected to complete the MP module. Note, however, that the demographic distribution for the MP respondents—both raw and weighted—still matches quite closely that of their

[5] The use of such a reduced set of calibration variables may be motivated by the fact that, in the case of small sample size, calibration on many variables is likely to result in extreme weights.

calibration proportions.

3 Weighting methods for non-probabilistic surveys

Traditionally, survey weights for the non-probability surveys used by the Economic Research and Analysis group in the Bank of Canada's Currency Department are created using the raking ratio method (iterative proportional fitting). They are usually implemented in STATA using the *ipfraking* command (Kolenikov 2014, 2019); see Chen et al. (2018) and Henry et al. (2019), for example. The change in survey design and collection instrument for the 2018 CFM, as well as recently implemented features in statistical software packages, has provided an opportunity to review the current research on non-probability survey weighting. In this review, we consider the most appropriate way to create a new series of survey weights.

This section relies heavily on Beaumont (2018), and we also borrow his notation. We refer the reader to the original paper for further details and additional considerations.[6]

In part due to declining response rates and high costs of collection, non-probability surveys have become more popular as a method of collecting data (Elliott and Valliant 2017). Typically, a non-probability survey is conducted by selecting respondents from a pre-existing panel of volunteer respondents (Mercer et al. 2017). However, there are specific challenges associated with non-probability surveys; see Couper (2000), Baker et al. (2013) and Elliott and Valliant (2017). Most notably, self-selection—the decision of an individual to participate in a panel from which survey respondents are selected—can result in substantial bias.

Beaumont (2018) describes two groups of strategies by which data from non-probability sources might be used to replace a probability survey: design-based approaches and model-based approaches. The main feature of design-based approaches is that they aim to produce design-consistent estimators, even when there is significant selection bias from the non-

[6] At the time of publishing this technical report, an updated English version of the paper is also available in Beaumont (2020).

probability data source. However, design-based methods typically require that the user have access to the values of the variable of interest y from a probability survey. We do not have access to probability survey values for most of the CFM variables of interest, in particular those variables in the MP module. Therefore, design-based approaches are not appropriate for our purposes.

Model-based approaches aim to reduce or eliminate selection bias from a non-probabilistic source by establishing a model between the variables of interest and auxiliary data. These methods can be used to obtain statistical inferences, given that certain conditions, necessary for the model to be valid, are satisfied. In his paper, Beaumont (2018) describes four model-based approaches: calibration of a non-probability sample, statistical matching, weighting by the inverse of response propensity and small area estimation. We give a brief description of the first three of these methods and explain why we choose calibration as the method of reducing selection bias when creating estimates from the online CFM surveys.

3.1 Definitions and notation

Let U denote the target population, y denote the variable of interest and y_k denote the value of y for unit k in U. In practice there will typically be several variables of interest, but for the sake of simplicity, we assume there is only one, and that the only estimate of interest is the total of the variable y for the population U, which we denote by $\theta = \sum_{k \in U} y_k$. Finally let $\mathbf{Y}$ represent the vector containing values of y_k for $k \in U$.

Consider the non-probability sample s_{NP} taken from the population U, and let δ_k denote the indicator variable for inclusion in the sample s_{NP}, i.e., let δ_k=1 if unit $k \in s_{NP}$, and let $\delta_k = 0$, otherwise. Denote by $\boldsymbol{\delta}$ the vector containing δ_k for $k \in U$.

From a non-probability sample s_{NP} with n_{NP} units, a naïve estimator of the total θ is given by $\hat{\theta}^{NP} = N \sum_{k \in s_{NP}} y_k / n^{NP}$, where N is the size of the population U. The naïve estimator is known to suffer from significant selection bias (Bethlehem 2016), and the model-based approaches we describe all use a vector of auxiliary variables, $\mathbf{x}_k$, to derive an estimator

that is less biased than the naïve estimator, when certain conditions are satisfied.

Let $\mathbf{X}$ denote the matrix containing the vectors $\mathbf{x}_k$ for all $k \in U$. For $k \in U$, let I_k be its inclusion indicator for s_P, and denote by $\mathbf{I}$ the vector containing all the indicators I_k. Finally, denote by $\boldsymbol{\Omega}$ the collection of all information used to make inferences on y. In the revised version of his paper, Beaumont (2020) includes in $\boldsymbol{\Omega}$ the design information, $\mathbf{Z}$, whether or not a probability sample is used, and potentially other auxiliary variables. He further states that the inclusion indicator δ_k can be used as an auxiliary variable for calibration, and that the vector $\boldsymbol{\delta}$ can thus be included in $\mathbf{Z}$ and $\boldsymbol{\Omega}$.

We can now state the following conditions, of which some combination must be satisfied for each of the model-based approaches.

Condition 1: $\mathbf{I}$ is independent of $\boldsymbol{\Omega}$ and $\mathbf{Y}$ after having conditioned on $\mathbf{Z}$.

Condition 2: $\boldsymbol{\delta}$ and $\mathbf{I}$ are independent after having conditioned on $\boldsymbol{\Omega}$ and $\mathbf{Y}$.

Condition 3: $\mathbf{Y}$ and $\boldsymbol{\delta}$ are independent after having conditioned on $\mathbf{X}$.

According to Beaumont (2018), one can expect that Conditions 1 and 2 are satisfied in the majority of cases, and he gives other implications of these conditions being met. The third condition, called the property of *exchangeability* by Mercer et al. (2017), is critical for eliminating selection bias via the methods in this section. Unfortunately, one cannot formally test whether the exchangeability requirement is met. Conditions 1 and 2 are only required for some of the model-based methods.

3.2 Statistical matching

Statistical matching is a method of combining two different data sources. A typical application is to pair a probability survey and a non-probability survey, where the variable of interest y is known for the units in the non-probability survey; see D'Orazio et al. (2006) or Rässler (2002), for example. This method relies on developing a model that predicts a variable of interest y based on auxiliary variables $\mathbf{x}_k$. This model is then used to impute y for the units in the probability survey, s_P, and the imputed values y_k^{imp} are then used with

the weights w_k for the probability survey to estimate θ as $\hat{\theta}^{SM} = \sum_{k \in s_P} w_k y_k^{imp}$. If the model used to calculate y_k^{imp} is linear, and Conditions 1 to 3 are satisfied, then $\hat{\theta}^{SM}$ is an unbiased estimator. One advantage of statistical matching is that non-parametric models can be used to predict y. However, statistical matching is not a viable approach for weighting the CFM because we do not have access to an appropriate probability survey with which to match it.

3.3 Weighting by the inverse of the propensity score

In contrast to statistical matching and calibration, which both model the relationship between y_k and $\mathbf{x}_k$, one can instead create a model for δ_k as a function of $\mathbf{x}_k$. In other words, one creates a model that predicts the likelihood that a unit k will be in the non-probability sample, given $\mathbf{x}_k$, for every unit $k \in U$. This model is then used to create weights for the units in s_{NP}. For each unit k, the probability of participation $p_k = \Pr(\delta_k = 1|\mathbf{X})$ is estimated as $\hat{p}_k$, and the estimator for θ is given by $\hat{\theta}^{PS} = \sum_{k \in s_{NP}} w_k^{PS} y_k$, where $w_k^{PS} = 1/\hat{p}_k$.

For this method, it is typical to assume that Condition 3 holds in such a way that $\Pr(\delta_k = 1|\mathbf{Y}, \mathbf{X}) = \Pr(\delta_k = 1|\mathbf{X})$. In other words, after conditioning on $\mathbf{X}$, whether or not a unit is in the non-probability sample is independent of $\mathbf{Y}$. Furthermore, one must assume that $p_k > 0, k \in U$. This assumption is called the *positivity* condition by Mercer et al. (2017). Whether or not this assumption is likely to hold is a critical question; see also Chen et al. (2019). The main advantage of this method is its simplicity: there is only one model rather than several models when there is more than one variable of interest.

Ideally, one would know the auxiliary variables $\mathbf{x}_k$ for every unit $k \in U$, but in reality this is typically unlikely to be the case. If $\mathbf{x}_k$ is not known for every unit, an alternative method exists for which it is only necessary to know the sums of $\mathbf{x}_k$ for the population; see Iannacchione et al. (1991). For other variations, see Chen et al. (2019), Lesage (2017), and Kim and Wang (2018). Among the various methods, most require Conditions 1 and 3 to be satisfied, and some also require Condition 2 to be met (see Kim and Wang (2018), for example). Beaumont (2018) describes several plausible situations for which the estimators

for weighting by the inverse of propensity might not have solutions, or the solutions might give propensity scores greater than 1.

For the CFM, U is the population of Canadian adult individuals living in the provinces, so we clearly do not have $\mathbf{x}_k$ for every unit $k \in U$. There is a method of weighting by the inverse of the propensity score that is applicable when $\mathbf{x}_k$ is not known for every unit $k \in U$, but this method can be shown to be equivalent to doing calibration as described in Section 3.4 (Iannacchione et al. 1991). Given that the methods can be made equivalent, and fully developed calibration methods are already implemented in the available statistical software, we choose to proceed with calibration for the CFM.

3.4 Calibration

The creation of survey weights via calibration is commonly used by polling companies that use volunteer panels for their survey samples (Vehovar et al. 2016). The idea is to model the relationship between the variable of interest y_k and the auxiliary variables $\mathbf{x}_k$ for each unit k in the non-probability sample, an approach which is described in Royall (1970) and generalized in Royall (1976); see also Elliott and Valliant (2017) and Valliant et al. (2000). With the calibration approach, the inferences made are conditional on $\boldsymbol{\delta}$ and $\mathbf{X}$. Finally, this method requires that the mechanism by which the units in s_{NP} are chosen is not informative. More formally put, $\mathbf{Y}$ and $\boldsymbol{\delta}$ must be independent after having conditioned on $\mathbf{X}$, i.e., Condition 3 must be satisfied. Beaumont (2018) states, "The richer $\mathbf{X}$ is, the more the conditional independence between $\mathbf{Y}$ and $\boldsymbol{\delta}$ becomes a realistic condition."

One common approach for estimating θ is to consider a linear model for which we assume that the observations y_k are mutually independent with $E(y_k|\mathbf{X}) = \mathbf{x}_k'\boldsymbol{\beta}$ and $\text{var}(y_k|\mathbf{X}) \propto v_k$, where $\boldsymbol{\beta}$ is a vector of unknown model parameters, and v_k is a known function of the variables in $\mathbf{x}_k$; see Valliant et al. (2000). The best linear unbiased predictor (BLUP) of θ is given by

$$\hat{\theta}^{BLUP} = \sum_{k \in s_{NP}} y_k + \sum_{k \in U - s_{NP}} \mathbf{x}_k'\hat{\boldsymbol{\beta}} = \mathbf{T}_{\mathbf{x}}'\hat{\boldsymbol{\beta}} + \sum_{k \in s_{NP}} (y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}} = (\sum_{k \in s_{NP}} v_k^{-1} \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{k \in s_{NP}} v_k^{-1} \mathbf{x}_k y_k$.

The estimator $\hat{\theta}^{BLUP}$ can also be rewritten in the weighted form

$$\hat{\theta}^{BLUP} = \sum_{k \in s_{NP}} w_k^C y_k, \tag{1}$$

where

$$w_k^C = 1 + v_k^{-1} \mathbf{x}_k (\sum_{k \in s_{NP}} v_k^{-1} \mathbf{x}_k \mathbf{x}_k')^{-1} (\mathbf{T}_{\mathbf{x}} - \sum_{k \in s_{NP}} \mathbf{x}_k).$$

It is straightforward to prove that w_k^C is a calibrated weight, i.e., that w_k^C satisfies the calibration equation

$$\sum_{k \in s_{NP}} w_k^C \mathbf{x}_k = \mathbf{T}_{\mathbf{x}}. \tag{2}$$

In other words, this method of estimating θ is equivalent to calculating calibrated weights for the units in s_{NP} under the assumption that the relationship between y_k and $\mathbf{x}_k$ is linear. To use this approach, the vector of control totals $\mathbf{T}_{\mathbf{x}}$ must be known.[7]

The 2018 CFM data contains several demographic and economic variables for which timely calibration totals can be obtained from Statistics Canada, so it is possible to use the calibration estimator for $\hat{\theta}$. Indeed, of all the methods described by Beaumont (2018), we find that calibration is the most appropriate for the CFM. Furthermore, the calibration methods described in Deville and Särndal (1992) are implemented in STATA 16.0, which makes calibration as described above easy to apply for the CFM.

[7]However, if the control totals for the population are not known, one alternative approach is to replace $\mathbf{T}_{\mathbf{x}}$ with estimated values $\hat{\mathbf{T}}_{\mathbf{X}} = \sum_{k \in s_P} w_k \mathbf{x}_k$, where the estimated values come from a probability sample denoted by s_P (Elliott and Valliant 2017).

4 Calibration variables and linearity

As discussed in Section 3.4, calibration can be effective for reducing the impact of the selection bias associated with non-probability surveys when estimating a population total θ, but its effectiveness relies on the assumption of a linear relationship between θ and the calibration variables $\mathbf{X}$. In this section, we detail how we use random forest models to identify the calibration variables that best predict several variables of interest from the MP module of the 2018 CFM. We then describe how we use Lasso and traditional linear regression to examine whether or not the linearity condition required for the calibration estimator is satisfied.

4.1 Selection of calibration variables

To determine on which variables to calibrate, we use a non-parametric model to identify which of the auxiliary data are best at predicting a subset of the continuous/ordinal variables from the MP module; see Table 1. More specifically, we use Breiman's random forest algorithm for classification and regression (Breiman 2001). There are many implementations of this algorithm available, and we choose the `randomForest 4.6-14` package for R.

4.1.1 CART and random forests

The random forest algorithm is an ensemble learning method for classification and regression. It is an extension of the Classification and Regression Tree (CART) class of machine learning algorithms; see Breiman et al. (1984). CART is an umbrella term that covers two main types of decision tree algorithms used in machine learning applications: (i) classification trees are used to predict discrete (categorical) values, and (ii) regression trees are used to predict continuous values.

Although there are slight differences between the two types of trees, the CART algorithm used to predict a variable y using a set of auxiliary variables $\mathbf{X}$ can essentially be summarized

as follows. A binary decision tree is constructed where each interior node represents a split of the data into two subsets. Each split is done according to the values of a single variable $\mathbf{x} \in \mathbf{X}$, and each leaf in the tree is labeled either with a class or with a probability distribution over the classes.

The tree is constructed recursively, i.e., below a split, a subtree is created for each of the two resulting subsets of the data. The splits are chosen based on a predefined set of rules to maximize the accuracy of the model (according to a chosen measure of accuracy, e.g., the Gini coefficient). The algorithm stops splitting the data when its stopping conditions are satisfied. To predict y for a new observation, its corresponding $\mathbf{x}$ values are used with the decision tree to determine which leaf y belongs in, and y is predicted based on the other values in the leaf.

One known issue with classification and regression trees is their tendency to *overfit* the model to the training data (Lewis 2000). In other words, a classification tree might be excellent at categorizing the data in the training set, but it fares poorly when making predictions for new data. This is due to the model essentially inventing relationships that appear to be present in the training data but are in fact a random artifact not reflected in future data. One method (among several) of dealing with overfitting in CART is the random forest algorithm.

The random forest algorithm is an extension of CART where, rather than relying on a single classification/regression tree, it creates a random collection of trees and then aggregates the results of these trees to make predictions. The primary advantage of random forests over single trees is that they are less likely to overfit to the training data, and they are more stable. At a high level, the model determined by the random forest algorithm is created using the following steps (Breiman 2001):

1. Select (with replacement) a random subset of the training data on which to train a tree.

2. At each node in the tree, choose at random a subset of the variables on which to split.

3. Grow the tree to its maximum size using the CART methodology.

4. Repeat steps 1-3 n times to create a forest of n trees.

To make a prediction for new data, use each of the n trees to make n predictions, and return either the average of the predictions (for regression) or the majority vote (for classification). The size of the random sample of the training data used to train each tree, the number of variables chosen at each node and the number of trees are all hyperparameters that are chosen by the user, where the default values may differ depending on the implementation of the algorithm chosen.

The random forest algorithm can be used to reduce the dimensionality of a model's independent variables by calculating an *importance* score for each variable, which can identify the most significant variables for the model. Importance is calculated as follows: for each tree in the forest, the mean squared error (MSE) is calculated for the out-of-bag portion of the data, i.e., the data not in the subsample used to train the tree. Then the same is done after permuting each predictor variable (i.e., re-creating the tree with each predictor variable removed, one at a time). The difference between the two MSEs is then averaged over all trees and normalized by the standard deviation of the differences. The more positive the importance measure is for a variable, the more important it is for predicting the variable of interest.

Table 3 shows an example of the importance values output by the `randomForest` algorithm for predicting *cash purchase (value)*, a continuous variable representing the total amount of cash spent by a respondent's household in the past month.

One common application of the importance function is to reduce the dimension of the auxiliary variables for a model, which for our purposes is equivalent to identifying the most effective variables on which to calibrate.[8]

For the continuous and ordinal variables of interest from the 2018 MP module, the 10

[8]There are other models that can also be used for feature selection, such as XGBoost (Chen and Guestrin 2016) and Lasso (Tibshirani 1996).

potential covariates that we test as predictors are presented in Table 2. Before analyzing the covariates, we collapse categories for several variables. The four Atlantic provinces are collapsed into a single region due to the relatively small number of respondents from each of them. The household size variable on the CFM has 12 categories (1, 2, . . . , 12+), which we collapse to 5 categories (1,2, . . . 5+) to align with available calibration totals from Statistics Canada. Similarly, employment status is collapsed from 10 categories down to 3 (employed, unemployed, not in the labour force) to match the categories of the calibration totals from the Labour Force Survey. Finally, we collapse the 25 categories of personal income down to 5 to match the personal income categories used in other Bank of Canada analyses. We choose to include personal individual income and exclude household income as potential covariates because the latter is more likely to suffer from measurement error, and more timely population totals for personal income are available from Statistics Canada. Also, we find that personal and household income are quite correlated, so there is limited utility in including both as calibration variables.

For each of the variables in Table 1, we run the random forest algorithm using the specified auxiliary variables as predictors. Then we use the importance values for each model to create a ranking of the auxiliary variables in terms of their overall importance. Specifically, a frequency table is constructed for the auxiliary variables counting the number of times each variable is the first, second, third, fourth or fifth most important variable in a given model. These frequencies are then used to calculate a score for each auxiliary variable, where 5 points are assigned for each time the variable is most important, 4 for each time it is second most important, 3 for each time it is third most important and so on.

Table 4 summarizes the variable rankings calculated for all the predictors when used to model the 2018 MP module variables of interest, as described in the previous paragraph. The total score for each potential predictor is shown. From the table, it is clear that age is overwhelmingly the most important variable overall, followed by marital status, personal income, household size, home ownership and employment status. The remaining variables

are relatively unimportant to the models.

Based on these importance rankings, the availability of calibration totals and the demographic variables most important for analysis, we calibrate on the following totals for the 2018 CFM: age range, gender, home ownership, personal income range, employment status, marital status and household size. Furthermore, the variables are nested[9] where possible, i.e., where calibration totals are available and there are a sufficient number of respondents for the resulting strata. Too few respondents in a stratum would result in unstable weights, or the calibration algorithm might not converge to a solution at all. For each variable, calibration is done at the region level, with other variables nested as possible. Table 5 gives the final calibration groups and their descriptions.

4.2 Examination of linearity assumption for calibration

In order to investigate the validity of the assumption of a linear relationship between the MP variables and the calibration variables, we create three models for each MP variable of interest: a random forest, a Lasso model and a stepwise linear regression model.[10] The purpose of these models is twofold: First, we check to see if Lasso identifies significant variables that are not found to be important by the random forest algorithm. Second, we compare the R-squared values for each of these models as a way to assess if the assumption of a linear relationship between the variable of interest and the calibration variables seems reasonable. If the random forest model performs substantially better than the linear models, that would indicate that the relationship between the auxiliary variables and the variable of interest is non-linear.

For each of the eight MP variables of interest, the following process is used to test for linearity. Observations with complete auxiliary data are split into two sets: a training set

[9]We do not test the importance of the nested covariates due to the amount of computational time required and the fact that the possible nestings were essentially already determined by stratum size and data availability considerations.

[10]The Lasso model is implemented in R using `glmnet`. The stepwise linear regression model is implemented in R using the `MASS` package.

(80 percent of the records) and a test set (remaining 20 percent). Then, a random forest, a Lasso model and a stepwise linear regression model are each trained on the same training set. Finally, each of the three models is used to predict the variable of interest using the test set, and the R-squared value for these predictions is calculated. Table 6 presents the resulting R-squared values.

Based on this goodness-of-fit criteria, all three models perform relatively poorly for each of the eight variables of interest. R-squared values for all 24 models range from 0.0002 to 0.1158. However, it is worth noting that the Lasso and stepwise linear models perform similar to each other, and these models almost always outperform the corresponding random forest, i.e., we do not have evidence of a strong non-linear relationship between the calibration variables and the variables of interest. We also perform Ramsey's Regression Equation Specification Error Test or RESET test (Ramsey 1969) that looks for evidence of omitted variables by fitting the original model augmented by the second, third and fourth power of the fitted values from the original model. Under the assumption of no misspecification, the coefficients on the powers of the fitted values will be zero. This would be evidence against non-linearity in the included regressors. We can't reject the null of no misspecification, hence linearity, at the 5 percent significance level only for three of the variables of interest considered; see the last column of Table 6.

The low R-squared obtained with all three models signals that important predictors of the variables of interest are missing, which calls into question whether or not Condition 3 (the exchangeability property) is satisfied for the CFM data. Recall that the plausibility of the exchangeability condition relies on auxiliary variables being able to effectively predict both the variables of interest and the propensity of a respondent to participate in a survey. The low goodness-of-fit measures obtained with the available calibration variables (CFM variables for which population totals are known) are, to this extent, concerning. It is clear that the linearity and exchangeability assumptions would be more strongly supported if there were questions on the survey that provided variables that were good linear predictors

of the variables of interest, and for which calibration totals from either census data or a probability survey were available. If such data were available for the CFM, the methods described by Beaumont (2018) for reducing bias could be applied with greater confidence in their effectiveness.

5 Construction of the final weights

5.1 Obtaining final weights

There are three main steps followed to create the final weights for the 2018 online CFM. First, any calibration variable with missing observations are imputed. Next, two sets of initial weights are calculated via post-stratification by region, age group and gender. The first set is for all 18,005 CFM 2018 respondents, and the second set is for the 12,004 CFM respondents that also filled in the MP module. These post-stratified weights are analogous to sampling weights in a probabilistic survey with a stratified sample design. Finally, calibration on the variables identified in Section 4.1 is performed to create a set of final CFM weights and a set of final CFM-MP weights.

5.1.1 Imputation of missing calibration variables

Two of the variables selected for calibration have missing values in the CFM data. Of 18,005 records, 220 records have missing values for personal income (`PERSONAL_INCOME`), 56 are missing employment status (`EMP01`) and 4 are missing both. In order to run the calibration algorithm to create weights, we must first deal with the missing values. Because the number of missing records is quite small, we choose to impute the collapsed variables `P_INC_CAT` (five categories) and `EMPSTAT` (three categories) using random forest classifiers implemented with the `randomForest` package in R. Details on the imputation process are provided in Appendix A.

5.1.2 Initial weights and trimming

We calibrate using the `svycal` procedure in STATA 16.0 (Valliant and Dever 2018). Specifically, we use its implementation of the truncated linear method described in Deville and Särndal (1992), which is a modified version of the GREG estimator equivalent to the expression in Equation (1). One desirable property of this calibration method is that it allows for setting upper and lower bounds on the calibration weights to be specified. This allows us to bound the weights during the calibration step without any post-calibration trimming, whereas raking gives the most asymmetric weight ratio distribution among the four methods outlined in Deville and Särndal (1992) with a particularly heavy right tail, which leads to more extreme large weights.

The primary purpose of trimming very large weights is to reduce instability in estimators (Särndal 2007). Battaglia et al. (2004) and DeBell and Krosnick (2009) suggest trimming weights during the calibration process to ensure that no units have a weight greater than five times their mean. Such trimming is applied, e.g., in Vincent (2015) and Chen et al. (2018) for the calibration of the Bank of Canada Methods-of-Payment Survey. The truncated linear method allows for similarly bounded weights without post-hoc manual adjustments.

One straightforward way to enforce the desired upper and lower bounds on the final weights is to first derive two sets of initial weights by post-stratifying on region by age group by gender: one set for the CFM respondents and one for the subset of respondents that completed the MP module. These initial weights are then used as input weights for the `svycal` calibration procedure in STATA 16.0, which allows users to specify an upper and lower bound on the ratio of the final weight to the initial weight for each respondent.

5.1.3 Final calibration weights

We limit weights to be between one fifth and five times their mean *within each stratum* defined by the region, age group and gender.[11] However, with these upper and lower limits, `svycal` does not converge for the MP sample, i.e., there is no solution to (2) satisfying the specified bounds. Consequently, we collapse the five household size categories into two categories for the MP calibration: households of one, and households of size greater than one. With this new household size variable, the algorithm converges for the MP sample.

5.2 Final weights description

Figure 2a compares the default weights provided by Ipsos to the initial and final CFM weights.[12] Ipsos weights and the initial CFM weights are quite close, since they are both derived using similar calibration totals. Both weights are calibrated on region, gender and age, but the default weights are, in addition, calibrated on household income estimates from the 2011 National Household Survey.

Figure 2b illustrates the correlation between initial post-stratification weights and final calibrated weights, for both the CFM sample and the MP subsample. As is typical when calibrating on many variables, the initial and final weights differ greatly, although for any given record, the final weight is between one fifth and five times the initial weight. Because the MP sample is only two-thirds the size of the CFM sample, the CFM-MP weights are approximately 50 per cent greater, on average. Consequently, the CFM-MP weights have a wider distribution.

In Tables 7 and 8, we compare unweighted and weighted demographic variables in the 2018 online CFM survey sample and its MP subsample, respectively, to the population

[11] This is the reason for using post-stratified weights (on region by age group by gender) as initial weights in the `svycal` command. If we were to instead use uniform initial weights of 1, we would have to specify a different upper and lower bound for each stratum—which would require running `svycal` once for each stratum—to achieve that goal.

[12] For the analysis in this section the default weights, which average to 1, are multiplied by a scaling factor equal to the sum of the final CFM weights. This is necessary to make meaningful comparisons between these weights.

targets used in calibration. Table 5 describes these targets and their sources.

As expected, the unweighted CFM sample is very close to population counts with respect to gender, age and region since the CFM sampling quotas closely match the distribution of those variables in the 2016 Census. However, the unweighted sample is biased mostly in terms of individual income, employment status and household size. It is also clear that default weights have very little impact on the sample distribution and do not perform well in rebalancing it toward the population's distribution. But the final calibrated weights w^f_{cfm} match the weighted sample distributions to the population ones perfectly for all calibration variables.

Similarly, in Table 8 we observe that the final calibrated weights for the MP subsample, w^f_{mp}, also match the weighted sample distributions to the population ones perfectly. Interestingly, the unweighted CFM and MP samples have very similar demographic compositions. This is not surprising as the MP module is assigned by Ipsos to respondents, not self-selected into by respondents.

In Table 9 we present descriptive statistics about the MP respondents who are assigned extreme calibrated weights. Compared with the overall MP subsample, they are more likely to be males, earn low individual income and be single. In terms of age, individuals receiving extreme weights belong mostly to either the two lowest or the highest age categories (i.e., they are either below 35 or above 65 years old). This analysis provides some evidence about the type of respondents that are under-represented in the CFM sample.

One could worry that individuals being assigned large weights also demonstrate extreme behaviour (i.e., are outliers), thereby strongly influencing our results. Figure 3 presents quantile-quantile plots of four response variables across two groups, respondents that receive extreme (y-axis) and non-extreme (x-axis) weights. The dots below the 45-degree line imply that, for the response variables considered, the individuals with extreme weights do not also behave like outliers.

5.3 Validation analysis

We perform a validation exercise on some variables from the MP module that are overlapping in the 2018 online CFM survey and the Bank of Canada 2017 Methods-of-Payment (MOP) Survey, as described in Table 10.[13]

Figures 4 and 5 compare the unweighted and weighted CFM-MP estimates to the MOP Survey estimates. Six response variables are considered: cash on hand, precautionary/other cash, withdrawal frequency at automated banking machines (ABMs), withdrawal frequency from bank tellers, typical withdrawal amount at ABMs, typical withdrawal amount from bank tellers.

The graphs in Figure 4 summarize the difference in means and conditional means between the CFM and MOP measures. They are obtained as follows. Let Y be the response variable being compared, such as cash on hand. We first compute the squared difference (in percent) across the CFM-MP and MOP estimates for $\overline{Y}$, the overall mean of Y, and for $\overline{Y}_D$, the mean of Y on a demographic domain D. The squared difference in overall means is shown in the first subset of bars in each graph. The remaining subsets of bars in each graph show, by demographic variables, the average of the squared differences in conditional means $\overline{Y}_D$, where the average is taken over all domains defined by this demographic variable. The following demographic variables are considered separately: gender, age, region, employment status, home owner/renter, marital status, household size.[14] The CFM-MP estimates are obtained (i) without weights, (ii) with Ipsos default weights, (iii) with the final calibrated CFM-MP weights w_{mp}^f. The MOP estimates are obtained with the survey questionnaire sample weights of the 2017 MOP data; see Chen et al. (2018). Finally, to improve clarity, the results in each subset are standardized so that the first bar (based on unweighted CFM-MP estimates)

[13] Recall that although the sampling unit of the 2018 online CFM is the individual, most questions in the MP module collect information at the household level. However, some questions on cash management are at the individual levels. These are the only questions directly comparable with questions in the MOP Survey.

[14] For example, for the demographic variable *gender*, the statistics shown in Figure 4 are $0.5 * \left(\left(\overline{Y}^{MP}_{Male} - \overline{Y}^{MOP}_{Male}\right)/\overline{Y}^{MOP}_{Male}\right)^2 + 0.5 * \left(\left(\overline{Y}^{MP}_{Female} - \overline{Y}^{MOP}_{Female}\right)/\overline{Y}^{MOP}_{Female}\right)^2$, where the superscripts MP and MOP denote their respective estimates.

equals one.

The graphs in Figure 5 are obtained similarly, but instead of $\overline{Y}$ or $\overline{Y}_D$ it is the difference $\overline{Y}_D - \overline{Y}$, hence the distribution across domains relative to the overall mean, that is compared across the two data sources. Overall, Figures 4 and 5 indicate that the w^f_{mp} weights help obtain mean estimates that are closer to the MOP estimates than those obtained with Ipsos default weights or without weights, both overall and by domain.

Figure 6 further shows quantile-quantile plots of two variables of interest, cash on hand and cash threshold, as measured in the 2017 MOP Survey (x-axis) and 2018 CFM survey (y-axis). It illustrates how the final calibrated CFM-MP weights bring the distribution of the variable in the MP subsample closer to that in the MOP sample, by reducing the influence of very large observations (outliers) at the right tail of the distribution.

5.4 Weighted results

Table 11 presents mean estimates for two variables of interest, cash purchases and contactless credit card (CTC) usage, unweighted and weighted with the final CFM-MP weights w^f_{mp}. The first variable is the dollar amount of cash the respondent's household used for purchases in the past month. The second is a binary variable indicating whether anyone in the respondent's household has used the contactless feature of a credit card in the past year, so that its weighted mean corresponds to the proportion of households that have used it.

We can observe that the CFM-MP weights increase the mean cash purchases of the overall sample and in most domains considered in Table 11. A notable exception is older respondents (65+) that see their average cash purchase decline as an effect of the w^f_{mp} weights. This shift of the sample toward higher cash purchase means is in line with the fact that young individuals and males, who tend to receive more of the larger weights (see Table 9), also make more cash purchases than the average individual in the sample (as seen in the first column of Table 11). In contrast, the CFM-MP weights decrease the mean proportion of CTC users in the overall sample and in all domains considered. This is in line with the fact that young

individuals and respondents with low individual income, who tend to receive more of the larger weights (see Table 9), also innovate less in terms of CTC usage (or have less access to credit cards) than the average individual in the sample (as seen in the third column of Table 11).

Results for the 2019 online CFM survey are provided in Appendix B, as are some comparisons between the 2015-2018 paper-based CFM surveys and the 2018 and 2019 online CFM surveys.

6 Summary and discussion

The Canadian Financial Monitor survey has been utilized by the Currency Department of the Bank of Canada since 2009 to track cash usage, payment card ownership and use, and the adoption of payment innovations by Canadians. The CFM survey changed from a paper-based household survey (last iteration in 2018) to an online individual survey in 2018.

The online CFM survey, like its paper-based predecessor, uses non-probability sampling for data collection. In this context, selection bias is very likely. We outline various methods for obtaining survey weights and discuss the associated conditions necessary for these weights to eliminate selection bias. In the end, we obtain calibration weights for the 2018 and 2019 online CFM samples. Our final weights improve upon the default weights provided by the survey company in several ways: (i) we choose the calibration variables based on a fully documented selection procedure that employs machine learning techniques; (ii) we use very up-to-date calibration totals; (iii) for each survey year we obtain two sets of weights, one for the full yearly sample of CFM respondents, the other for the subsample of CFM respondents who also filled in the MP module of the questionnaire.

Several lessons can be learned from this analysis. First, calibrating on more variables than the default weights does have an impact on weighted estimates. Second, although the CFM sample and MP subsample both support calibration on many different marginal

and nested totals simultaneously, the MP subsample requires collapsing categories of one variable. Third, random forests are effective for imputation of CFM variables, at least when there is only a small fraction of the data with missing values.

However, the key lesson learned is the need for relevant calibration variables. In order to be more confident that calibration actually is reducing bias, we need to be able to calibrate on variables that are highly correlated with the variables of interest in the survey—and ideally, there would be a linear relationship. A way of accomplishing this is to include, in the CFM questionnaire, some questions from available probability surveys that could provide important predictors of the CFM variables of interest. By doing so, population totals would be available for important calibration variables beyond basic demographic variables.[15]

[15]In a similar spirit, two questions from the Digital Economy Survey, a probability survey conducted by Statistics Canada, were added to the 2019 Cash Alternative Survey, a recent survey undertaken by the Currency Department of the Bank of Canada; see Huynh et al. (ming). Note however that these overlapping questions are used as a source of cross-validation, to assess the bias of the 2019 CAS, but not in its calibration process, per se.

References

Baker, R., J. M. Brick, N. A. Bates, M. Battaglia, M. P. Couper, J. A. Dever, K. J. Gile and R. Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-probability Sampling." *Journal of Survey Statistics and Methodology 1*(2), 90–143.

Battaglia, M. P., D. Izrael, D. C. Hoaglin and M. R. Frankel. 2004. "Tips and Tricks for Raking Survey Data (aka Sample Balancing)." *Abt Associates 1*, 4740–4744.

Beaumont, J.-F. 2018. "Les Enquêtes Probabilistes Sont-elles Vouées à Dis-paraître pour la Production de Statistiques Officielles?" Paper presented at the Dixième Colloque Francophone sur les Sondages, Lyon, France, October 24. Available at `http://sondages2018.sfds.asso.fr/assets/Documents/Session1_ConferenceOuverture_Beaumont_Article.pdf`.

Beaumont, J.-F. 2020. "Are Probability Surveys Bound to Disappear for the Production of Official Statistics?" *Survey Methodology 46* (1), 1–28.

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

Bethlehem, J. 2016. "Solving the Nonresponse Problem with Sample Matching?" *Social Science Computer Review 34*(1), 59–77.

Breiman, L. 2001. "Random Forests." *Machine Learning 45*(1), 5–32.

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen. 1984. *Classification and Regression Trees*. Belmont, California: Wadsworth.

Chen, H., M.-H. Felt and C. Henry. 2018. "2017 Methods-of-Payment Survey: Sample Calibration and Variance Estimation." Bank of Canada Technical Report No. 114.

Chen, H., M.-H. Felt and K. P. Huynh. 2017. "Retail Payment Innovations and Cash Usage: Accounting for Attrition Using Refreshment Samples." *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*(2), 503–530.

Chen, H. and Q. R. Shen. 2015. "Variance Estimation for Survey-weighted Data using Bootstrap Resampling Methods: 2013 Methods-of-Payment Survey Questionnaire." Bank of Canada Technical Report No. 104.

Chen, T. and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 16, New York, 785–794. Association for Computing Machinery.

Chen, Y., P. Li and C. Wu. 2019. "Doubly Robust Inference With Nonprobability Survey Samples." *Journal of the American Statistical Association,* 1-11.

Couper, M. P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly 64*(4), 464–494.

DeBell, M. and J. A. Krosnick. 2009. "Computing Weights for American National Election Study Survey Data." ANES Technical Report Series, no. nes012427. Ann Arbor, MI, and Palo Alto, CA: American National Election Studies. Available at `http://www.electionstidies.org`.

Deville, J.-C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association 87*(418), 376–382.

D'Orazio, M., M. Di Zio and M. Scanu. 2006. *Statistical Matching: Theory and Practice*. Chichester: John Wiley & Sons.

Elliott, M. R. and R. Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science 32*(2), 249–264.

Felt, M.-H. 2018. "A Look Inside the Box: Combining Aggregate and Marginal Distributions to Identify Joint Distributions." Bank of Canada Staff Working Paper No. 2018-29.

Gelman, A. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science 22*(2), 153–164.

Henry, C., K. P. Huynh, G. Nicholls and M. W. Nicholson. 2019. "2018 Bitcoin Omnibus Survey: Awareness and Usage." Bank of Canada Staff Discussion Paper No. 2019-10.

Henry, C., K. P. Huynh and A. Welte. 2018. "2017 Methods-of-Payment Survey Report." Bank of Canada Staff Discussion Paper No. 2018-17.

Huynh, K. P., G. Nicholls and M. W. Nicholson. Forthcoming. "2019 Cash Alternative Survey: Canadians' Cash Use in an Increasingly Digital Economy." Bank of Canada Staff Discussion Paper.

Iannacchione, V. G., J. G. Milne and R. E. Folsom. 1991. "Response Probability Weight Adjustments Using Logistic Regression." In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 637–642.

Ipsos. 2018. "CFM Update." Presentation to the Bank of Canada, Ottawa, Ontario, December 7.

Kim, J. and Z. Wang. 2018. "Sampling Techniques for Big Data Analysis in Finite Population Inference." Unpublished manuscript.

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

Kolenikov, S. 2019. "Updates to the Ipfraking Ecosystem." *Stata Journal 19*(1), 143–184.

Lesage, É. 2017. "Combiner des Données d'Enquêtes Probabilistes et des Données Massives Non Probabilistes pour Estimer des Paramètres de Population Finie." Unpublished manuscript.

Lewis, R. J. 2000. "An Introduction to Classification and Regression Tree (CART) Analysis." Paper presented at the Annual Meeting of the Society for Academic Emergency Medicine, San Francisco, California, May 22−25.

Mercer, A. W., F. Kreuter, S. Keeter and E. A. Stuart. 2017. "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference." *Public Opinion Quarterly 81*(S1), 250–271.

Ramsey, J. B. 1969. "Tests for Specification Errors in Classical Linear Least-squares Regression Analysis." *Journal of the Royal Statistical Society. Series B (Methodological) 31* (2), 350–371.

Rao, J., C. Wu and K. Yue. 1992. "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology 18*(2), 209–217.

Rässler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer Science & Business Media.

Royall, R. M. 1970. "On Finite Population Sampling Theory Under Certain Linear Regression Models." *Biometrika 57*(2), 377–387.

Royall, R. M. 1976. "The Linear Least-squares Prediction Approach to Two-stage Sampling." *Journal of the American Statistical Association 71*(355), 657–664.

Särndal, C.-E. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology 33*(2), 99–119.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Valliant, R. and J. A. Dever. 2018. *Survey Weights: a Step-by-step Guide to Calculation*. College Station, Texas: Stata Press.

Valliant, R., A. H. Dorfman and R. M. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley.

Vehovar, V., V. Toepoel and S. Steinmetz. 2016. "Non-probability Sampling." In *The SAGE Handbook of Survey Methods*, 329–345. Thousand Oaks, California: SAGE.

Vincent, K. 2015. "2013 Methods-of-Payment Survey: Sample Calibration Analysis." Bank of Canada Technical Report No. 103.

TABLE 1: Variables of interest used for calibration variable selection

Variable	Type	Description
Cash purchase (volume)	Ordinal	Number of times the household used cash to make purchases in past month
Cash purchase (value)	Continuous	Total amount of cash the household used to make purchases in past month
Cash on hand	Continuous	Amount of cash in respondent's purse, wallet or pockets right now
Cash threshold	Continuous	How low respondent typically lets amount of cash in purse, wallet or pockets fall before withdrawing more cash
CC purchase in store (value)	Continuous	Total amount the household spent using a credit card in a store in past month
DC purchase in store (value)	Continuous	Total amount the household spent using a debit card in a store in past month
CC purchase online (value)	Continuous	Total amount the household spent using a credit card online in past month
DC purchase online (value)	Continuous	Total amount the household spent using a debit card online in past month

Notes: All variables are constructed based on variables from the MP module of the 2018 online CFM survey (logical cleaning only). CC and DC stand for credit card and debit card, respectively.

TABLE 2: Potential calibration variables considered

Variable name	Description	Number of categories
RESP_GENDER	Respondent gender	2
Year_Month	Age in years	N/A
PROVCAT, collapsed to REGION	Province	10, collapsed to 7
CITYSIZE	Size of city of residence	4
PERSONAL_INCOME, collapsed to P_INC_CAT	Personal income (pre-tax), collapsed	26, collapsed to 6 (one is "Prefer not to answer")
EMP01, collapsed to EMPSTAT	Employment status	10, collapsed to 3 (one is "Prefer not to answer")
Y4, renamed DWELL_TYPE	Dwelling type	6
Y5, renamed OWN_HOME	Own or rent home	2
USMAR2	Marital status	5
HHCMP10, collapsed to HHSIZE	Number of household members	12, collapsed to 5

Note: Variables stem directly from the online CFM survey, except for those that were collapsed into fewer categories.

TABLE 3: Covariates' importance for predicting *cash purchase (value)*

Variable	% increase in MSE
Year_Month	23844
RESP_GENDER	956
REGION	4404
CITYSIZE	2651
P_INC_CAT	7772
EMPSTAT	14468
DWELL_TYPE	7517
OWN_HOME	9083
USMAR2	16070
HHSIZE	23303

Notes: Importance is calculated as follows: for each tree in the forest, the MSE is calculated for the out-of-bag portion of the data. Then the same is done after permuting each predictor variable. The difference between the two MSEs is then averaged over all trees and normalized by the standard deviation of the differences.

TABLE 4: Summary of covariates' importance rankings for all variables of interest

	1	2	3	4	5	Score
Year_Month	5	2	1			36
USMAR2	1	1	3	2		22
P_INC_CAT	1	2	1	2	1	21
EMPSTAT		1	2	4	1	19
HHSIZE	1	2	1			15
OWN_HOME		1	2	2	1	15
DWELL_TYPE					1	1
RESP_GENDER					1	1
REGION					1	1
CITYSIZE						0

Notes: For each of the eight variables of interest on the CFM-MP, a random forest model was created with the potential covariates, and the importance for each covariate was calculated. These importance ranks are used to calculate an overall score for each covariate, where a higher score indicates greater importance overall.

TABLE 5: Calibration variable combinations used for weighting the 2018 online CFM sample

Calibration description (number of strata)	Source of population targets
Region (7) × gender (2) × age group (6)	December 2018 demographic projections
Region (7) × gender (2) × employment status (3)	2018 Labour Force Survey
Region (7) × gender (2) × personal income category (5)	2017 T1 Family File
Region (7) × gender (2) × marital status (3)	2016 Census
Age group (6) × employment status (3)	2018 Labour Force Survey
Age group (6) × personal income category (5)	2017 T1 Family File
Region (5) × home ownership status (2)	2016 Census
Region (5) × household size (5 or 2)	2016 Census

Notes: The Canadian provinces are collapsed either into seven regions (with the four Atlantic provinces representing a single region) or into five regions (with Alberta, Saskatchewan and Manitoba further collapsed into a single geographic area). We obtain custom tabulations for the 2016 Census and 2017 T1 Family File (T1FF) that include only those individuals aged 18 and older.

TABLE 6: R^2 of random forest, LASSO and stepwise linear regression models, and Ramsey RESET test

	R^2			Ramsey RESET
Dependent variable	Random forest	LASSO	Stepwise regression	test (p-value)
Cash purchase (volume)	0.018	0.016	0.017	0.630
Cash purchase (value)	0.009	0.015	0.013	0.617
Cash on hand	0.053	0.080	0.082	0.037
Cash threshold	0.047	0.058	0.057	0.000
CC purchase in store (value)	0.086	0.116	0.114	0.000
DC purchase in store (value)	0.028	0.032	0.038	0.008
CC purchase online (value)	0.082	0.088	0.088	0.000
DC purchase online (value)	0.000	0.024	0.018	0.402

Notes: The first three columns of this table present R^2 for three different models. For each model, 80% of the data ias used as training data. R^2 is calculated on the remaining 20% of the data set aside for testing. In the last column, P-values of the Ramsey RESET test are shown. This test assesses whether the model is linear in the original variables, by adding powers of the fitted values of the dependent variable in the linear model, and tests their joint significance. Under the assumption of linearity, the coefficients on the powers of the fitted values will be zero.

TABLE 7: Sample vs. population composition–2018 CFM sample

	CFM sample	$w_{cfm}^{default}$	w_{cfm}^{i}	w_{cfm}^{f}	Population
Male	48.56	48.56	49.22	49.22	49.22
Female	51.44	51.44	50.78	50.78	50.78
Age:18-24	10.85	10.94	10.96	10.96	10.96
25-34	16.38	16.40	17.42	17.42	17.42
35-44	16.21	16.15	16.66	16.66	16.66
45-54	17.96	17.91	16.36	16.36	16.36
55-64	17.50	17.47	17.39	17.39	17.39
65+	21.12	21.14	21.22	21.22	21.22
Atlantic	6.80	6.83	6.57	6.57	6.57
Quebec	23.49	23.47	23.11	23.11	23.11
Ontario	38.41	38.41	39.31	39.31	39.31
Prairies	17.72	17.72	17.69	17.69	17.69
B.C.	13.57	13.57	13.32	13.32	13.32
Ind. income: <$25K	25.67	25.65	25.71	36.05	36.05
$25-45K	21.34	21.29	21.69	24.04	24.04
$45-60K	15.11	15.14	15.11	13.10	13.10
$60-100K	24.92	24.92	24.71	17.64	17.64
$100K+	12.96	13.00	12.78	9.17	9.17
Employed	59.29	59.28	59.59	62.99	62.99
Unemployed	3.30	3.30	3.37	3.59	3.59
Not in labour force	37.41	37.42	37.04	33.42	33.42
Own their home	68.93	68.95	69.08	72.99	72.99
Rent their home	31.07	31.05	30.92	27.01	27.01
Single	25.81	25.82	26.36	25.16	25.16
Married/common law	62.85	62.87	61.88	60.94	60.94
Widowed/divorced/separated	11.34	11.31	11.76	13.90	13.90
Hh size: 1	19.68	19.63	20.34	14.43	14.43
2	42.72	42.74	42.16	34.12	34.12
3	17.51	17.52	17.58	18.71	18.71
4	13.34	13.35	13.26	18.56	18.56
5+	6.75	6.76	6.67	14.18	14.18

Notes: This table shows the composition of the CFM sample across the eight demographic variables used as calibration variables. Numbers are percentages. Column 1 shows unweighted proportions for the overall CFM sample. Columns 2 to 4 show weighted results, where $w_{cfm}^{default}$ is for Ipsos default weights, w_{cfm}^{i} is for the poststratified initial CFM weights and w_{cfm}^{f} is for the final calibrated CFM weights. Population distributions are presented in the last column.

TABLE 8: Sample vs. population composition–2018 MP subsample

	MP subsample	w^i_{mp}	w^f_{mp}	Population
Male	48.53	49.22	49.22	49.22
Female	51.47	50.78	50.78	50.78
Age:18-24	10.83	10.96	10.96	10.96
25-34	16.39	17.42	17.42	17.42
35-44	16.24	16.66	16.66	16.66
45-54	17.96	16.36	16.36	16.36
55-64	17.51	17.39	17.39	17.39
65+	21.08	21.22	21.22	21.22
Atlantic	6.80	6.57	6.57	6.57
Quebec	23.50	23.11	23.11	23.11
Ontario	38.39	39.31	39.31	39.31
Prairies	17.74	17.69	17.69	17.69
B.C.	13.58	13.32	13.32	13.32
Ind. income: <$25K	25.64	25.72	36.05	36.05
$25-45K	20.98	21.23	24.04	24.04
$45-60K	14.98	15.00	13.10	13.10
$60-100K	25.24	25.04	17.64	17.64
$100K+	13.16	13.01	9.17	9.17
Employed	58.94	59.30	62.99	62.99
Unemployed	3.15	3.21	3.56	3.59
Not in labour force	37.91	37.49	33.45	33.42
Own their home	69.47	69.51	72.99	72.99
Rent their home	30.53	30.49	27.01	27.01
Single	26.02	26.55	25.16	25.16
Married/common law	62.59	61.59	60.94	60.94
Widowed/divorced/separated	11.40	11.87	13.90	13.90
Hh size: 1	19.91	20.69	14.43	14.43
2+	80.09	79.31	85.57	85.57

Notes: This table shows the composition of the MP subsample across the eight demographic variables used as calibration variables. Numbers are percentages. Column 1 shows unweighted proportions for the MP subsample. Columns 2 and 3 show weighted results, where w^i_{mp} is for the poststratified initial CFM-MP weights, and w^f_{mp} is for the final calibrated CFM-MP weights. Population distributions are presented in the last column.

TABLE 9: Characteristics of respondents with extreme weights

	MP subsample	≥90pct	≥99pct
Male	48.53	60.28	64.67
Female	51.47	39.72	35.33
Age:18-24	10.83	15.07	42.67
25-34	16.39	19.57	20.67
35-44	16.24	8.91	0.00
45-54	17.96	6.16	0.00
55-64	17.51	14.15	0.00
65+	21.08	36.14	36.67
Atlantic	6.80	6.83	6.67
Quebec	23.50	22.56	16.00
Ontario	38.39	41.13	53.33
Prairies	17.74	16.74	15.33
B.C.	13.58	12.74	8.67
Ind. income: <$25K	25.64	71.52	96.00
$25-45K	20.98	21.32	2.67
$45-60K	14.98	5.50	1.33
$60-100K	25.24	1.17	0.00
$100K+	13.16	0.50	0.00
Employed	58.94	53.12	58.67
Unemployed	3.15	5.66	4.00
Not in labour force	37.91	41.22	37.33
Own their home	69.47	70.19	74.00
Rent their home	30.53	29.81	26.00
Single	26.02	29.31	47.33
Married/common law	62.59	47.71	44.67
Widowed/divorced/separated	11.40	22.98	8.00
Hh size: 1	19.91	13.16	1.33
2+	80.09	86.84	98.67
Cash purchase (value)	$321	$387	$346
CTC user	0.55	0.44	0.47

Notes: The first part of this table shows demographic compositions, in proportions. Column 1 shows unweighted estimates for the overall MP subsample. Columns 2 and 3 describe respondents with weights above the 90^{th} and 99^{th} percentile of the final CFM-MP weights distribution, respectively. The last two rows of this table show mean estimates for two variables of interest: *Cash purchase (value)* is the dollar amount of cash the respondent's household used for purchases in the past month; *CTC user* is a binary variable indicating whether anyone in the respondent's household has used the contactless feature of a credit card in the past year, so that its weighted mean corresponds to the proportion of households that have used it.

TABLE 10: Overlapping questions in the 2018 online CFM and 2017 MOP surveys

	2018 online CFM	2017 MOP
Withdrawal frequency	How many times in the past month did you withdraw cash using: an ABM/ a bank teller	In a typical month, how often do you obtain cash in each of the following ways? From an ABM/ATM (bank/cash machine)/ From a bank teller
Withdrawal amount	What was the typical withdrawal amount from.....? an ABM/ a bank teller [Proposed brackets: $1 - $20; $21- $40; $41 - $60; $61 - $80; $81 - $100; $101 - $200; $201 - $500; $501 - $1,000; More than $1,000]	When you withdraw cash from each of the following sources, what amount do you typically withdraw? Please write in dollar amount
Cash on hand	How much cash do you have in your purse, wallet or pockets right now? Write in the amount $	How much cash do you have in your purse, wallet, or pockets right now? Please write in
Cash threshold	How low do you typically let cash in your purse, wallet or pockets fall before obtaining more cash? Write in the amount	With respect to the cash you carry in your wallet, purse, or pockets, how low do you typically let your cash supply fall before obtaining more cash? Please write
Precautionary/ other cash	How much cash on hand does your household hold for emergencies or other precautionary reasons? [Proposed brackets: None; $1 - $49; $50 - $99; $100 - $249; $250 - $499; $500 - $999; $1,000 - $2,999; $3,000 or more]	Does your household currently have any other cash holdings in your home, vehicle, or elsewhere? If yes, what is the total value of this cash? If you answer 'Yes', please write in a dollar amount

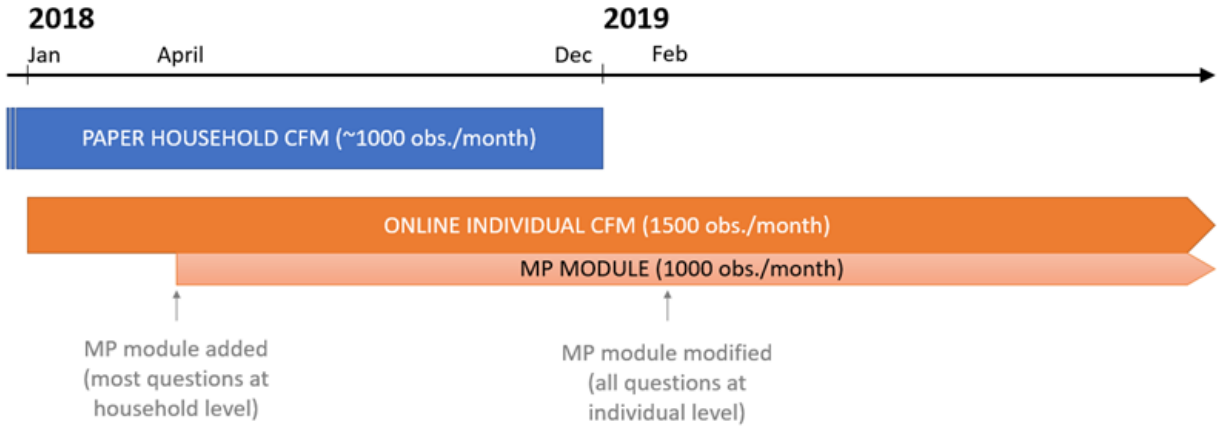
Notes: This table presents questions on cash management at the individual level that overlap in the 2018 online CFM survey and the Bank of Canada 2017 Methods-of-Payment (MOP) Survey.

Table 11: 2018 online CFM mean estimates

	Cash purchase (value)		CTC user (%)	
	MP subsample	w^f_{mp}	MP subsample	w^f_{mp}
Overall	321	335	0.55	0.52
Male	371	395	0.56	0.52
Female	273	277	0.53	0.52
Age: 18-24	341	426	0.52	0.50
25-34	301	323	0.57	0.54
35-44	360	376	0.51	0.50
45-54	356	378	0.52	0.52
55-64	303	298	0.54	0.51
65+	281	262	0.61	0.54
Atlantic	259	245	0.51	0.48
Quebec	279	289	0.53	0.50
Ontario	373	398	0.56	0.53
Prairies	305	297	0.54	0.51
B.C.	299	324	0.59	0.57
Ind. income: <$25K	301	322	0.44	0.44
$25-45K	267	282	0.50	0.50
$45-60K	337	388	0.56	0.55
$60-100K	342	356	0.63	0.62
$100K+	384	404	0.68	0.66
Employed	341	361	0.55	0.53
Unemployed	292	290	0.54	0.51
Own their home	338	339	0.59	0.56
Rent their home	281	322	0.46	0.41
Single	299	358	0.49	0.46
Married/common law	341	342	0.59	0.56
Widowed/divorced/separated	261	261	0.49	0.47
Hh size: 1	243	273	0.47	0.43
2+	340	345	0.57	0.54

Notes: This table shows unweighted and weighted mean estimates for two variables of interest: *Cash purchase (value)* is the dollar amount of cash the respondent's household used for purchases in the past month; *CTC user* is a binary variable indicating whether anyone in the respondent's household has used the contactless feature of a credit card in the past year, so that its weighted mean corresponds to the proportion of households that have used it. Columns 1 and 3 show unweighted estimates for the MP subsample, where these two variables are observed. Columns 2 and 4 show weighted results using the final calibrated CFM-MP weights.

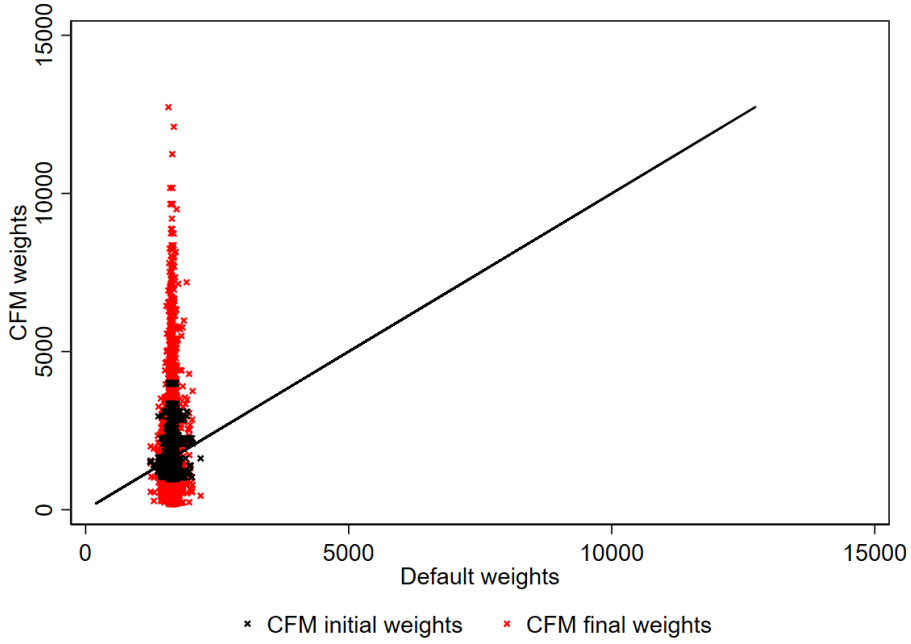
Figure 1: Timeline of developments in the paper-based and online CFM surveys in 2018 and 2019



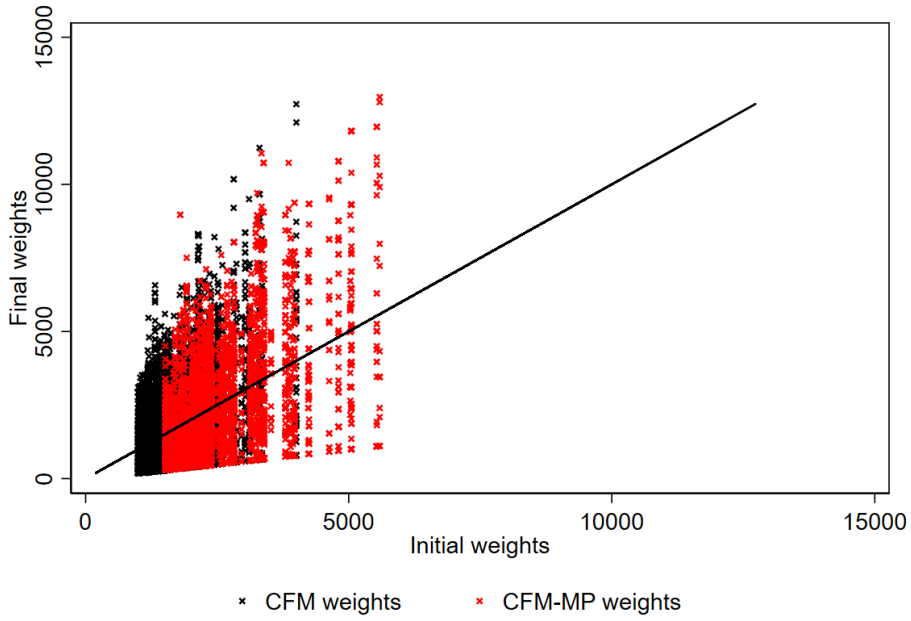
Notes: The new online individual survey was launched in January 2018. It is fielded monthly via online or mobile survey and targets a sample size of 1,500 individuals per month. The old CFM paper-based survey was last run in 2018. Its monthly sample size is 1,000 households. The key difference between the paper-based and online CFM is the level of sampling and reporting. The paper-based CFM reports at the household level while the online CFM is individual based.

Figure 2: Comparing weights distributions

(a) Correlation between default weights and CFM weights



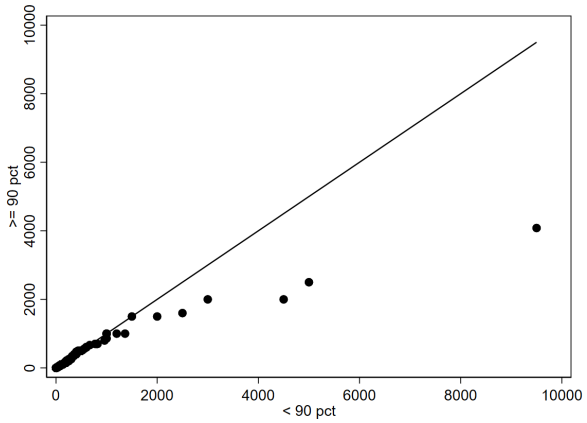
(b) Correlation between initial and final weights



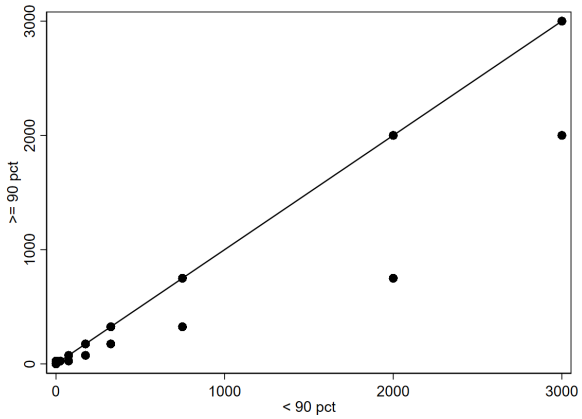
Notes: The top scatter plot shows the association between the default weights $w_{cfm}^{default}$ and (i) the initial post-stratified CFM weights w_{cfm}^{i} (in black); (ii) the final calibrated CFM weights w_{cfm}^{f} (in red). The bottom scatter plot shows the association between the initial post-stratified weights and the final calibrated weights for (i) the CFM sample (w_{cfm}^{i} and w_{cfm}^{f}) in black; (ii) the MP subsample (w_{mp}^{i} and w_{mp}^{f}) in red.

Figure 3: Quantile-quantile plots of response variables for respondents with and without extreme weights

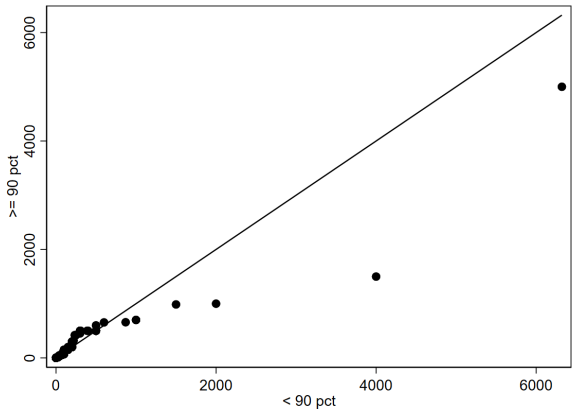
(a) Cash on hand



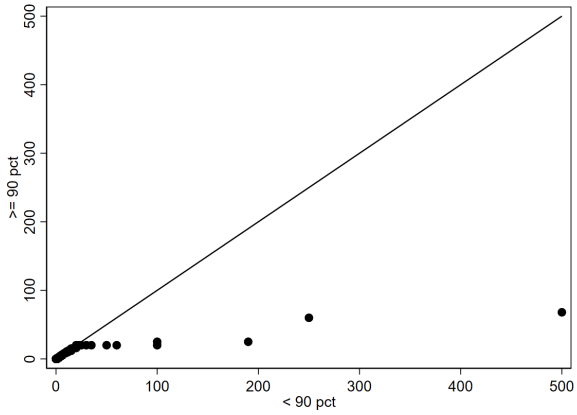
(b) Precautionary/other cash



(c) Cash threshold



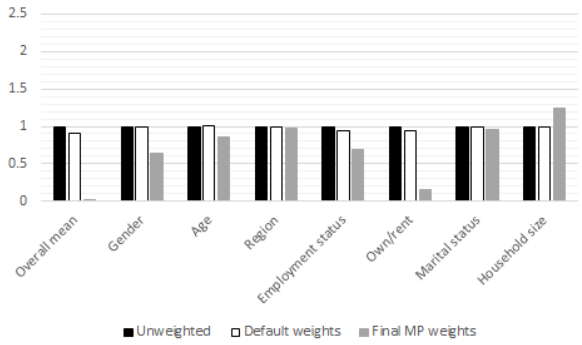
(d) Withdrawal frequency at ABMs



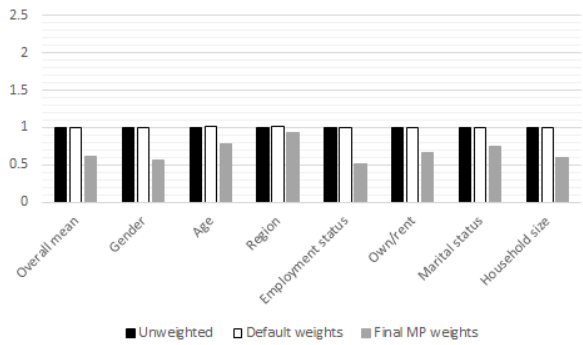
Notes: These graphs show quantile-quantile plots of four variables of interest (unweighted) across two groups, respondents that receive final calibrated CFM-MP weights below (x-axis) and above (y-axis) the 90th percentile of their distribution.

Figure 4: Comparing the CFM and MOP mean estimates, overall and by domain

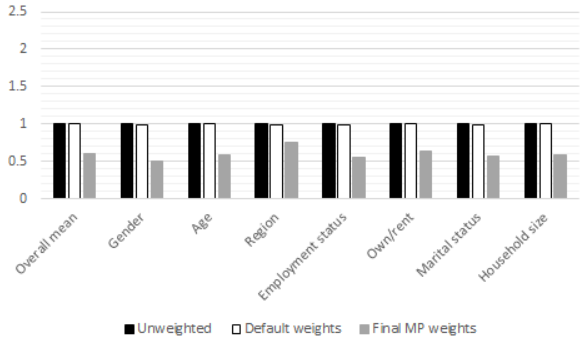
(a) Cash on hand



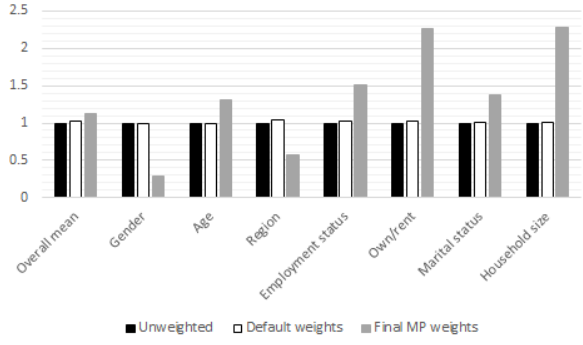
(b) Precautionary/other cash



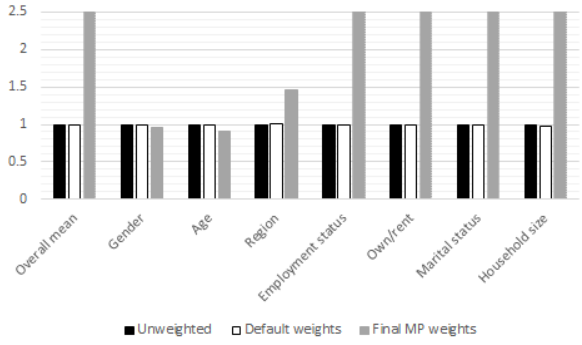
(c) Withdrawal frequency at ABMs



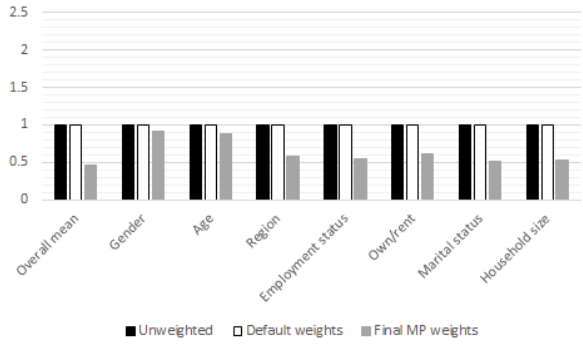
(d) Withdrawal frequency from bank teller



(e) Withdrawal amount at ABMs



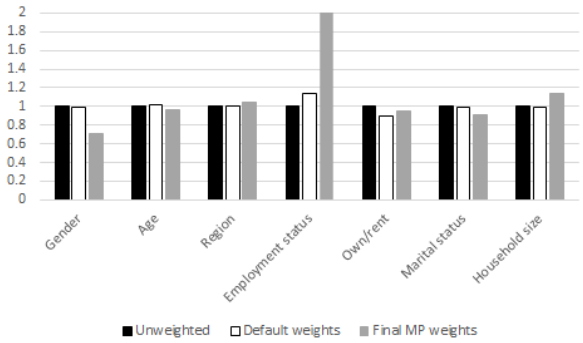
(f) Withdrawal amount from bank teller



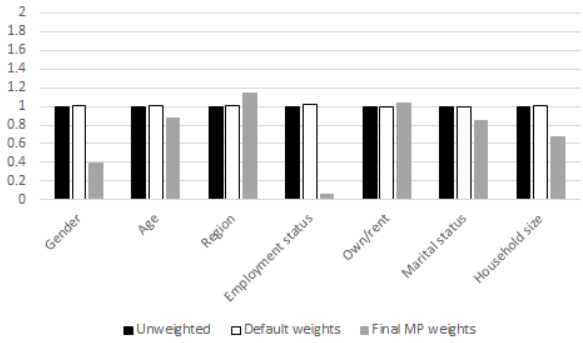
Notes: These graphs summarize the difference in means and conditional means between the CFM and MOP surveys, for six variables of interest. The squared difference in overall means $\left(\left(\overline{Y}^{MP} - \overline{Y}^{MOP}\right)/\overline{Y}^{MOP}\right)^2$ is shown in the first subset of bars in each graph, where the superscripts MP and MOP denote their respective estimates. The remaining subsets of bars in each graph show, by demographic variables, the average of the squared differences in conditional means $\overline{Y}_D$, where the average is taken over all domains defined by this demographic variable. For example, for *gender*, the statistics shown are $0.5 * \left(\left(\overline{Y}^{MP}_{Male} - \overline{Y}^{MOP}_{Male}\right)/\overline{Y}^{MOP}_{Male}\right)^2 + 0.5 * \left(\left(\overline{Y}^{MP}_{Female} - \overline{Y}^{MOP}_{Female}\right)/\overline{Y}^{MOP}_{Female}\right)^2$. The CFM-MP estimates are obtained (i) without weights, (ii) with Ipsos default weights, (iii) with the final calibrated CFM-MP weights w^f_{mp}. To improve clarity, the results in each subset are standardized so that the first bar (based on unweighted CFM-MP estimates) equals one.

Figure 5: Comparing the CFM and MOP deviations between domain and overall mean estimates

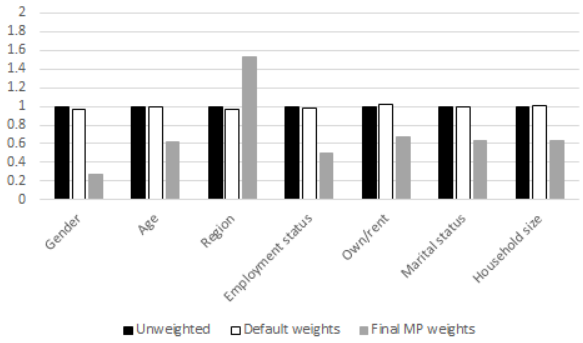
(a) Cash on hand



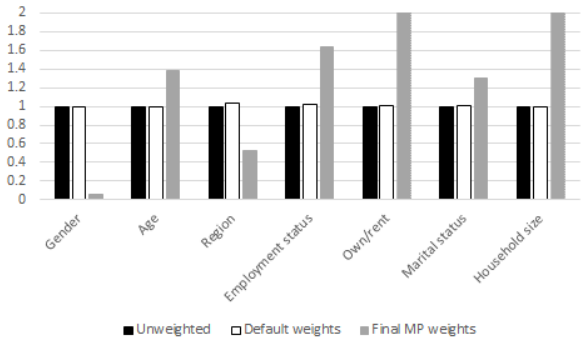
(b) Precautionary/other cash



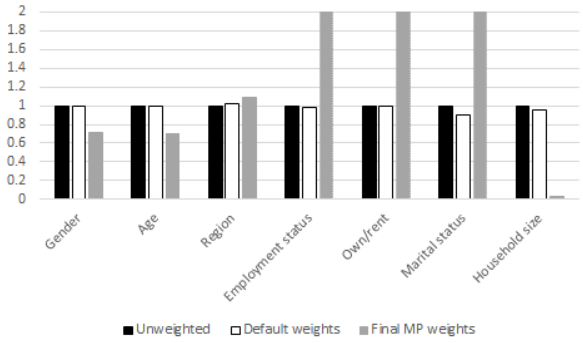
(c) Withdrawal frequency at ABMs



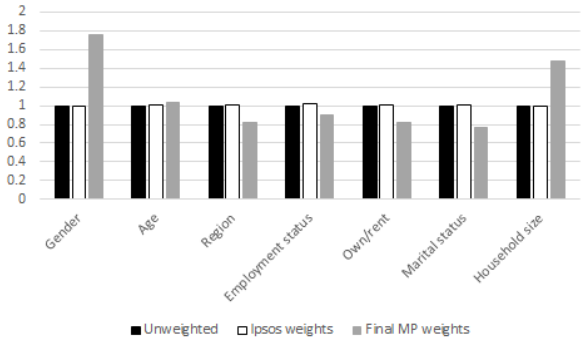
(d) Withdrawal frequency from bank teller



(e) Withdrawal amount at ABMs



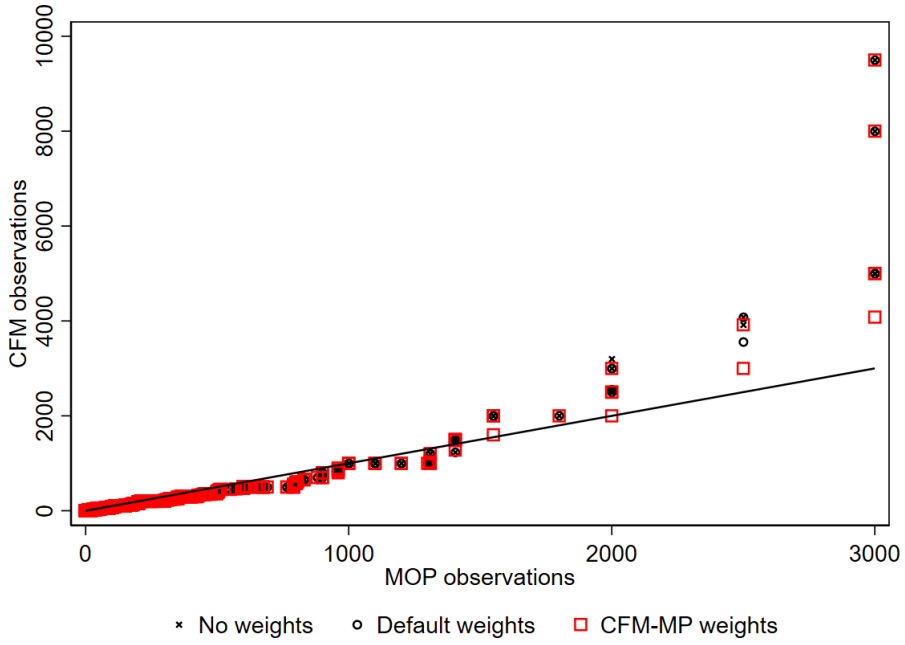
(f) Withdrawal amount from bank teller



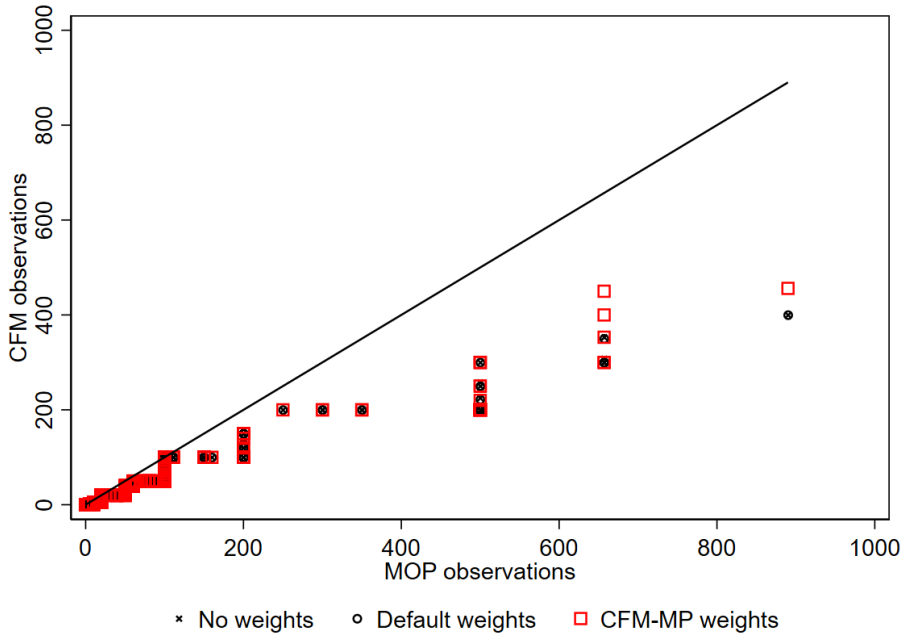
Notes: These graphs are obtained similarly to the graphs in the previous figure, but here it is the difference $\overline{Y}_D - \overline{Y}$ between the domain means and the overall mean that is compared across the CFM and MOP surveys. For example, for *gender*, the statistics shown are $0.5 * \left[\left((\overline{Y}^{MP}_{Male} - \overline{Y}^{MP}) - (\overline{Y}^{MOP}_{Male} - \overline{Y}^{MOP})\right)/(\overline{Y}^{MOP}_{Male} - \overline{Y}^{MOP})\right]^2 + 0.5 * \left[\left((\overline{Y}^{MP}_{Female} - \overline{Y}^{MP}) - (\overline{Y}^{MOP}_{Female} - \overline{Y}^{MOP})\right)/(\overline{Y}^{MOP}_{Female} - \overline{Y}^{MOP})\right]^2$, where the superscripts MP and MOP denote their respective estimates.

Figure 6: Quantile-quantile plots of CFM and MOP response variables

(a) Cash on hand



(b) Cash threshold



Notes: These graphs show quantile-quantile plots of two variables of interest, *cash on hand* and *cash threshold*, as measured in the 2017 MOP Survey (x-axis) and 2018 online CFM survey (y-axis). The CFM estimates are obtained (i) without weights (black stars), (ii) with Ipsos default weights (black circles) and (iii) with the final calibrated CFM-MP weights w_{mp}^{f} (red squares). The MOP estimates are obtained with the survey questionnaire sample weights of the 2017 MOP data; see Chen et al. (2018).

A Details on the implementation of imputation

We impute variables using the following process:

1. Employment status is imputed for the 52 records missing only that variable. The imputation uses Model 1, which includes personal income as a variable.

2. Employment status is imputed for the 4 records missing both that variable and personal income. The imputation uses Model 2, which does not include personal income as a variable.

3. Personal income is imputed for all 220 records for which that variable is missing. The imputation uses Model 3, which includes employment status as a variable.

For the employment status models (Model 1 and Model 2) we take advantage of the `sampsize` option in `randomForest` to force the algorithm to sample a relatively higher proportion of unemployed individuals when randomly selecting a subset of units for each tree in a forest. This oversampling is necessary because for the 17,949 complete cases, there are only 582 unemployed individuals, whereas 10,651 are employed and 6,716 are not in the labour force. As a result, when we train models for predicting `EMPSTAT` by sampling the entire population uniformly, the resulting classifiers are very poor at correctly identifying unemployed individuals.

We find that models are significantly better at identifying unemployed individuals when we set `sampsize` = c(600, 300, 600). What this means is that each classification tree in the forest is trained on a random subset of the data containing 600 employed individuals, 300 unemployed individuals and 600 individuals not in the workforce. However, using the non-uniform sampling does lead to a slight reduction in accuracy when identifying individuals in the other two categories. Note that because the data is quite balanced across the five personal income categories, we do not use the `sampsize` option for Model 3.

The other hyperparameter we endeavour to optimize is `mtry`. This hyperparameter controls how many of the available variables are randomly selected at each node for each tree in

the random forest. For each model, the impact of this hyperparameter is found to be quite minimal, and we settle on `mtry` = 3 for Model 1 and Model 3. For Model 2, we set `mtry` equal to 4.

As with the feature selection models, we use 80 percent of the records to train the model and reserve 20 percent as a test set for estimating each model's effectiveness, which we primarily evaluate using the balanced accuracy score for each class to be predicted.[16] Table A.1 presents balanced accuracy scores for each model. In terms of overall accuracy, the respective scores of Model 1, Model 2 and Model 3 are 77.0 percent, 75.5 percent and 65.5 percent. Considering the relatively small number of records requiring imputation, we are satisfied with the predictive ability of the models.

[16]The balanced accuracy of a classifier is calculated for each class. It is the arithmetic mean of the true positive rate and the true negative rate for that class. The maximum value of the balanced accuracy for a class is 1.

TABLE A.1: Balanced accuracy of models 1, 2 and 3

Model 1 (`EMPSTAT`)

Category	Balanced accuracy
Employed	0.8078
Unemployed	0.6540
Not in labour force	0.7931

Model 2 (`EMPSTAT`)

Category	Balanced accuracy
Employed	0.7786
Unemployed	0.6371
Not in labour force	0.7833

Model 3 (`P_INC_CAT`)

Category	Balanced accuracy
<$25K	0.8486
$25-45K	0.7536
$45-60K	0.7039
$60-100K	0.7821
$100K+	0.7689

Notes: The balanced accuracy scores for each of the final imputation models are calculated for the test sets containing the 20% of the data not used to train the model.

B Calibration of the 2019 online CFM survey

In February 2019, the online CFM survey was modified in several ways. First, the MP module was modified to align with the remainder of the survey: all questions on methods-of-payment usage were changed from household-level to individual-level questions, so as to make the sampling unit and unit of observation coincide; see Table B.1. Second, a question about the education level of the respondent was added to the core module.

Several sets of calibration weights are computed for the 2019 online CFM data, following the same calibration procedure employed for the 2018 data. First, we compute calibration weights for the yearly 2019 CFM sample (w^f_{cfm}) and for the yearly 2019 MP subsample (w^f_{mp}) using the variable combinations used for the 2018 sample and presented in Table 5. Second, we compute weights for the subsample of respondents that filled in the new version of the MP module between February and December 2019: one set of weights (w^f_{mp*}) is obtained with the same combination of nested calibration variables as w^f_{mp}, while the other (w^+_{mp*}) is obtained by calibrating in addition on region (seven categories) $\times$ education.[17]

Table B.2 shows the composition of the 2018 and 2019 online CFM samples, unweighted and weighted using the w^f_{cfm} weights, as well as the population targets in both years. Overall, the compositions of the unweighted samples in the two years are very close. However, we can notice a significant shift of the raw sample composition toward higher income categories between 2018 and 2019. Just as for 2018, the final calibrated weights for the 2019 CFM sample match the weighted sample distributions to the population ones perfectly.

Table B.3 presents means estimates for two variables of interest from the February-December 2019 online CFM survey: *Individual cash purchase (value)* is the dollar amount of cash the *respondent* used for purchases in the past month; *Individual CTC user* is a binary variable indicating whether the *respondent* has used the contactless feature of a credit card in the past year, so that its weighted mean corresponds to the proportion of respondents that

[17]For calibration, the education variable from the CFM is collapsed from eight categories to three: 1 - no higher than high school, 2 - completed college/CEGEP/trade school, 3 - some university or higher.

have used it.[18] Weighted mean estimates are obtained with two different sets of CFM-MP weights, calibrated with and without education as an extra calibration variable. Overall, the estimates obtained with both sets of calibrated weights are very close: adding education to the set of calibration variables does not make a significant impact on the results for the two response variables considered.

As in 2018, we can observe that the CFM-MP weights increase the mean cash purchases of the overall sample and in most domains considered. Again, this shift of the sample toward higher cash purchase means is most important for young and single respondents. Also, as in 2018, the CFM-MP weights decrease the mean proportion of CTC users in the overall sample and in all domains. Finally, we find that more educated respondents use less cash and innovate more (CTC adoption) than less educated respondents, which is in line with what has been observed in other studies; see, e.g., Henry et al. (2018).

[18]Recall that these variables are not directly comparable with the equivalent 2018 variables *Cash purchase (value)* and *CTC user*, which measure cash and CTC use at the respondent's *household* level.

TABLE B.1: Sampling unit and unit of observation in the paper-based and online CFM surveys

	Paper-based CFM	2018 online CFM	2019 online CFM (since February)
Sampling unit	Household	Individual	Individual
Unit of observation in the MP section/module:			
- for methods-of-payment questions	Household (aggregate)	Household (aggregate)	Individual
- for cash management questions	Individual	Individual	Individual
Unit of observation in other sections/modules:	Household (all individuals)	Individual	Individual

Notes: This table presents the changes in the sampling unit and unit of observation over the years in the CFM surveys. The unit of observation is the individual if the question is at the individual level, e.g., 'In the past month, have you personally withdrawn cash?" The unit of observation is the household (aggregate) if the question is at the aggregate household level, e.g., "How many times in the past month did your household use cash to make purchases?" Finally, the unit of observation is all individuals in the household if the question asks to provide disaggregate information for all individuals in the household, e.g., "Does anyone in your household have any chequing account? Please list using one line for each account in the household."

TABLE B.2: Sample vs. population composition–2018 and 2019 online CFM samples

	CFM sample		w^f_{cfm}		Population	
	2018	2019	2018	2019	2018	2019
Male	48.56	48.58	49.22	49.26	49.22	49.26
Female	51.44	51.42	50.78	50.74	50.78	50.74
Age:18-24	10.85	10.83	10.96	10.87	10.96	10.87
25-34	16.38	16.37	17.42	17.49	17.42	17.49
35-44	16.21	16.19	16.66	16.75	16.66	16.75
45-54	17.96	17.91	16.36	15.85	16.36	15.85
55-64	17.50	17.65	17.39	17.32	17.39	17.32
65+	21.12	21.06	21.22	21.72	21.22	21.72
Atlantic	6.80	6.80	6.57	6.52	6.57	6.52
Quebec	23.49	23.47	23.11	22.96	23.11	22.96
Ontario	38.41	38.40	39.31	39.48	39.31	39.48
Prairies	17.72	17.73	17.69	17.67	17.69	17.67
B.C.	13.57	13.60	13.32	13.37	13.32	13.37
Ind. income: <$25K	25.67	23.83	36.05	36.09	36.05	36.09
$25-45K	21.34	20.42	24.04	24.08	24.04	24.08
$45-60K	15.11	14.44	13.10	13.10	13.10	13.10
$60-100K	24.92	25.76	17.64	17.60	17.64	17.60
$100K+	12.96	15.55	9.17	9.12	9.17	9.12
Employed	59.29	59.72	62.99	63.39	62.99	63.39
Unemployed	3.30	3.19	3.59	3.51	3.59	3.51
Not in labour force	37.41	37.08	33.42	33.10	33.42	33.10
Own their home	68.93	67.41	72.99	73.00	72.99	73.00
Rent their home	31.07	32.59	27.01	27.00	27.01	27.00
Single	25.81	26.33	25.16	25.16	25.16	25.16
Married/common law	62.85	61.51	60.94	60.94	60.94	60.94
Widowed/divorced/separated	11.34	12.16	13.90	13.90	13.90	13.90
Hh size: 1	19.68	19.77	14.43	14.42	14.43	14.42
2	42.72	41.21	34.12	34.11	34.12	34.11
3	17.51	17.69	18.71	18.71	18.71	18.71
4	13.34	14.43	18.56	18.57	18.56	18.57
5+	6.75	6.90	14.18	14.19	14.18	14.19

Notes: This table shows the composition of the 2018 and 2019 online CFM samples across the eight demographic variables used as calibration variables. Numbers are percentages. Columns 1 and 2 show unweighted proportions for the overall CFM samples. Columns 3 and 4 show weighted results obtained using the final calibrated CFM weights w^f_{cfm}. Population distributions are presented in the last two columns.

TABLE B.3: 2019 online CFM mean estimates

	Individual cash purchase (value)			Individual CTC user		
	MP Feb-Dec	w^f_{mp*}	w^+_{mp*}	MP Feb-Dec	w^f_{mp*}	w^+_{mp*}
Overall	208	218	223	0.55	0.51	0.50
Male	269	290	296	0.56	0.50	0.50
Female	151	148	152	0.53	0.52	0.51
Age:18-24	232	364	363	0.51	0.47	0.47
25-34	185	189	206	0.52	0.47	0.46
35-44	195	196	196	0.52	0.50	0.49
45-54	211	204	210	0.53	0.52	0.51
55-64	222	229	233	0.54	0.51	0.50
65+	209	186	188	0.62	0.57	0.57
Atlantic	248	228	228	0.51	0.47	0.46
Quebec	206	255	255	0.50	0.47	0.47
Ontario	208	203	210	0.56	0.53	0.52
Prairies	226	245	253	0.56	0.51	0.51
B.C.	168	155	161	0.57	0.54	0.53
Ind. income: <$25K	187	229	231	0.42	0.42	0.42
$25-45K	209	210	211	0.50	0.51	0.50
$45-60K	179	182	189	0.54	0.53	0.52
$60-100K	199	190	195	0.61	0.60	0.59
$100K+	278	299	324	0.68	0.67	0.67
Employed	218	235	241	0.55	0.51	0.51
Unemployed	193	188	191	0.54	0.50	0.50
Own their home	213	223	228	0.59	0.54	0.54
Rent their home	198	205	208	0.46	0.42	0.41
Single	197	250	253	0.47	0.43	0.42
Married/common law	215	214	220	0.58	0.55	0.54
Widowed/divorced/separated	195	177	179	0.52	0.50	0.49
Hh size: 1	171	159	162	0.50	0.46	0.46
2+	217	228	233	0.56	0.52	0.51
High school/college	227	231	231	0.47	0.45	0.45
University	187	200	206	0.63	0.60	0.60

Notes: This table shows unweighted and weighted mean estimates for two variables of interest: *Individual cash purchase (value)* is the dollar amount of cash the respondent used for purchases in the past month; *Individual CTC user* is a binary variable indicating whether the respondent has used the contactless feature of a credit card in the past year, so that its weighted mean corresponds to the proportion of respondents that have used it. Columns 1 and 4 show unweighted estimates for the February to December MP subsample, where these two variables are observed. Columns 2 and 5 show weighted results using the calibrated CFM-MP weights for the February-December subsample obtained without education as a calibration variable (w^f_{mp*}). Columns 3 and 6 show weighted results using the calibrated CFM-MP weights for the February-December subsample obtained with education as a calibration variable (w^+_{mp*}).

C Comparing CFM outcomes across years

In the paper-based CFM survey (run until 2018), the sampling unit as well as the unit of observation in most sections is the household. In the 2018 online CFM survey, these changed to the individual respondent. However, in the MP module, all the questions were kept identical to the 2018 paper-based CFM; hence questions on methods-of-payment usage collected information at the household aggregate level. This was modified in February 2019 when the unit of observation was aligned with the sampling unit (the individual respondent) in the whole MP section. Refer to Table B.1 for a summary of the changes in the sampling unit and unit of observation in the various sections of the CFM surveys over the years.

These changes in the unit of observation make potential comparison across years limited to a handful of variables on cash management that were measured at the individual level in all the successive paper-based and online CFM surveys. They are described in Table C.1. Note that only the question on *cash on hand* remained exactly the same over the years. Questions on cash withdrawal as well as *cash threshold* underwent slight modifications over the years. By contrast, the question on precautionary cash changed significantly from "cash for emergencies" to "cash outside your purse or wallet" formulations, so that comparison across years is impeded. Finally, it is important to keep in mind that even though the unit of observation is uniform over the years for these questions, the paper-based CFM surveys are household surveys with household weights, while the online CFM surveys are individual surveys with individual weights. This further limits potential comparisons of CFM outcomes over the years.

Table C.2 shows unweighted and weighted mean estimates from the 2018 and 2019 online CFM surveys for four questions that are identical or very close across both years. Weighted estimates are obtained using the final calibrated CFM-MP weights w^f_{mp}. In each year, weights have little impact on the overall means. They impact more intensively respondents who are young, who have low income or who are single, which is in line with the fact that these domains receive most of the very large weights; see Table 9. Respondents in these

demographic categories see their average cash on hand and cash threshold increase as an effect of the weights, which is coherent with the previous finding that they also tend to use more cash in a month than the average individual in the sample; see Section 5.4.

Comparing the 2018 and 2019 estimates, we observe a decline in average *cash on hand* overall and across all domains. This is somehow contradictory with the increase observed in the *cash threshold* mean estimates from one year to the next. Note however that the wording of the *cash threshold* question changed slightly between both years. The wording of the question on cash withdrawal frequency also changed from asking about cash "withdrawn" in 2018 to cash "withdrawn or received" in 2019. This could explain the slight increase in the estimated withdrawal frequency between 2018 and 2019. As can be seen in columns 9 to 12 of Table C.2, the mean number of withdrawals at ABMs per month was estimated to be 2.1 in 2018 and 2.4 in 2019. This slight increase contradicts somehow the declining–although not steady–trend in the frequency of withdrawals at ABMs observed in previous years, as presented in Figure C.1c.

Table C.3 provides estimates of the change in overall means of the variables of interest between 2018 and 2019. The estimated differences are obtained using two different methods. Column 3 shows the simple difference in weighted means, while column 4 reports the estimated coefficient of a year dummy in a linear regression also controlling for all the calibration variables used in the weighting. We observe that the regression coefficient and the change in weighted averages always have the same sign and also tend to be quite close in magnitude. Interestingly, this result implies that regression modelling could be an easy alternative to constructing survey weights for estimating time trends in the CFM survey; see Gelman (2007) for details.

Figure C.1 shows weighted means of six variables as measured in the paper-based CFM surveys from 2015 to 2018 and in the 2018 and 2019 online CFM surveys. As discussed above, these six variables stem from questions with similar wording. However, note that paper-based CFM outcomes are weighted with household weights (default weights provided

by Ipsos), while online CFM outcomes are weighted using our person-level calibrated CFM-MP weights w^f_{mp}. This means that even for very similarly worded questions, the weighted means for the paper-based surveys and those for the online surveys are measuring different quantities.

For example, consider the first question from the paper-based CFM in Table C.1: "In the past month, have you withdrawn cash in any of the following ways? If yes, how many times in the past month?" When using household weights, the weighted mean for the second part of that question represents the average number of withdrawals by Canadian heads of household who have withdrawn cash in the past month. Conversely, when using person-level weights, the weighted mean represents the average number of withdrawals by all Canadian adults. This is a subtle but important distinction that makes comparison between the paper-based survey and the online survey very difficult since corresponding questions are measuring similar, but fundamentally different, concepts.

There is a clear break in the time series of variable *Cash threshold*, for which the 2018 paper-based and online results differ significantly. For the remaining variables, the 2018 paper-based and online results are relatively close to each other, especially when the variables are winsorized.

The new trends that are appearing in the online CFM data do not always align well with the trends observed in the paper-based CFM. For example, although the average amount of cash on hand was virtually unchanged over the 2015-2018 period at about $80, the 2019 online CFM measured mean *cash on hand* estimates close to $70. We reiterate that this disparity is at the very least exacerbated by the fact that the paper-based survey weights are *household*-level, and the online weights are *person*-level, which hampers the utility of direct comparisons.

TABLE C.1: Overlapping questions in the paper-based and online CFM surveys

	2015-2017 paper-based CFM	2018 paper-based CFM	2018 online CFM	2019 online CFM
Withdrawal frequency	In the past month, have you withdrawn cash in any of the following ways? If yes, how many times in the past month? Write in the number of times: ABM/ Bank teller/ Cash-back by debit card at retailer/ Cash advance by credit card	In the past month, have you withdrawn cash in any of the following ways? If yes, how many times in the past month? Write in the number of times: ABM/ Bank teller/ Cash-back by debit card at retailer/ Cash advance by credit card	In the past month, have you personally withdrawn cash in any of the following ways? How many times in the past month did you withdraw cash using: an ABM/ a bank teller/ cash-back by debit card at retailer/ a cash advance by credit card/ a cash advance by a prepaid card/ Other ways	In the past month, have you personally withdrawn or received cash in any of the following ways? How many times in the past month did you withdraw or received cash…From an ABM/ From a bank teller/ Cash-back with debit card purchases at a physical store/ Cash advance by credit card/ From other sources (cash received from friends or family, salary paid in cash, tips, etc.)
Withdrawal amount	What was the typical withdrawal amount? [Proposed brackets: $1 - $20; $21- $40; $41 - $60; $61 - $80; $81 - $100; $101 - $200; $201 - $500; $501 - $1,000; More than $1,000]	What was the typical withdrawal amount? [Proposed brackets: $1 - $20; $21- $40; $41 - $60; $61 - $80; $81 - $100; $101 - $200; $201 - $500; $501 - $1,000; More than $1,000]	What was the typical withdrawal amount from…….? [Proposed brackets: $1 - $20; $21- $40; $41 - $60; $61 - $80; $81 - $100; $101 - $200; $201 - $500; $501 - $1,000; More than $1,000]	What was the typical amount withdrawn or received? [Proposed brackets: $1 - $20; $21- $40; $41 - $60; $61 - $80; $81 - $100; $101 - $200; $201 - $500; $501 - $1,000; More than $1,000]
Cash on hand	How much cash do you have in your purse, wallet or pockets right now? Write in	How much cash do you have in your purse, wallet or pockets right now? Write in	How much cash do you have in your purse, wallet or pockets right now? Write in the amount $	How much cash do you have in your purse, wallet or pockets right now? Write in the amount $
Cash threshold	How low do you typically let cash in your purse, wallet or pockets fall before obtaining more cash? Write in	How low do you typically let cash in your purse, wallet or pockets fall before obtaining more cash? Write in	How low do you typically let cash in your purse, wallet or pockets fall before obtaining more cash? Write in the amount	How low do you typically let the amount of cash in your purse, wallet or pockets fall before withdrawing more cash? Write in the amount $
Precautionary/ other cash	How much cash on hand does your household hold for emergencies, or other precautionary reasons? [Proposed brackets: None; $1 - $49; $50 - $99; $100 - $249; $250 - $499; $500 - $999; $1,000 - $2,999; $3,000 or more]	How much cash does your household hold outside of your purse, wallet or pockets right now? Write in	How much cash on hand does your household hold for emergencies or other precautionary reasons? [Proposed brackets: None; $1 - $49; $50 - $99; $100 - $249; $250 - $499; $500 - $999; $1,000 - $2,999; $3,000 or more]	How much cash do you have outside of your purse, wallet or pockets right now? [Proposed brackets: None; $1 - $49; $50 - $99; $100 - $249; $250 - $499; $500 - $999; $1,000 - $2,999; $3,000 or more]

Notes: This table presents questions on cash management at the individual level which overlap in the 2015-2017 paper-based CFM, the 2018 paper-based and online CFM and the 2019 online CFM surveys.

TABLE C.2: Comparing the 2018 and 2019 online CFM outcomes–mean estimates for some variables of interest

	Cash on hand				Cash threshold				Withdrawal freq. at ABMs				Withdrawal amount at ABMs			
	2018		2019		2018		2019		2018		2019		2018		2019	
	MP	w^f_{mp}	MP	w^f_{mp}	MP	w^f_{mp}	MP	w^f_{mp}	MP	w^f_{mp}	MP	w^f_{mp}	MP	w^f_{mp}	MP	w^f_{mp}
Overall	88	85	71	71	32	32	40	40	2.1	2.1	2.5	2.4	138	132	141	138
Male	111	104	87	85	43	43	49	51	2.3	2.3	2.7	2.7	146	137	153	147
Female	67	66	55	56	23	22	30	30	2.0	1.9	2.3	2.2	131	128	129	128
Age:18-34	75	78	56	73	35	38	36	45	1.9	1.9	2.5	2.6	110	109	115	119
35-54	85	85	65	62	32	33	39	39	2.3	2.2	2.7	2.6	140	136	141	139
55+	101	89	86	76	31	27	42	38	2.2	2.1	2.4	2.2	154	144	157	150
Atlantic	78	75	63	68	25	27	29	28	2.4	2.1	2.5	2.3	131	129	131	135
Quebec	91	92	66	68	37	39	48	50	2.4	2.3	3.3	3.1	150	146	155	153
Ontario	94	88	73	72	35	32	39	38	2.2	2.1	2.4	2.4	140	132	137	133
Prairies	82	76	71	71	26	25	37	41	1.9	1.9	2.0	1.9	125	119	136	135
B.C.	82	78	76	72	29	31	36	36	1.8	1.8	2.1	2.3	133	128	139	132
Ind. income: <$25K	64	70	52	64	24	27	31	37	1.9	1.9	2.2	2.3	122	120	129	132
$25-45K	74	77	64	69	26	28	35	37	2.3	2.2	2.6	2.4	128	128	130	130
$45-60K	87	86	66	66	30	30	39	43	2.3	2.3	2.7	2.7	132	127	138	143
$60-100K	110	109	77	74	42	44	42	42	2.2	2.3	2.5	2.4	146	144	143	140
$100K+	119	112	101	99	46	45	54	53	2.0	2.2	2.7	2.7	173	169	169	166
Employed	88	87	70	72	35	36	41	43	2.2	2.1	2.6	2.6	133	130	136	134
Unemployed	89	80	73	68	28	25	38	35	2.1	2.1	2.4	2.2	146	137	149	146
Own their home	98	91	79	77	36	34	43	42	2.1	2.1	2.3	2.2	141	132	146	140
Rent their home	65	67	53	53	23	26	33	35	2.3	2.2	2.9	3.0	133	133	133	134
Single	74	75	59	67	29	32	36	40	2.1	2.0	2.7	2.4	119	118	126	126
Married/c.l.	94	90	75	73	35	34	41	42	2.2	2.1	2.5	2.4	145	138	147	143
Wid./div./sep.	88	81	75	68	27	24	37	34	2.2	2.2	2.6	2.6	142	130	144	140
Hh size: 1	86	81	69	68	30	28	39	38	2.2	2.1	2.6	2.4	135	131	135	133
2+	89	85	71	71	33	33	40	41	2.1	2.1	2.5	2.4	139	133	143	139
High school/college			66	65			39	39			2.6	2.5			139	135
University			74	76			43	44			2.5	2.4			143	141

Notes: This table shows unweighted and weighted mean estimates for four variables of interest that are comparable in the 2018 and 2019 online CFM surveys: *Cash on hand* is the dollar amount of cash in the respondent's purse, wallet, or pockets right now; *Cash threshold* is how low the respondent typically lets the amount of cash in their purse, wallet or pockets fall before withdrawing more cash; *Withdrawal frequency at ABMs* is the number of times the respondent withdrew cash from an ABM in the past month; *Withdrawal amount at ABMs* is the typical withdrawal amount at an ABM; see Table C.1. Variables are winsorized at the 99.5th percentile. Columns headed "MP" show unweighted estimates obtained on the yearly MP subsamples. Columns headed "w^f_{mp}" show weighted results obtained using the final calibrated CFM-MP weights for the yearly MP subsamples.

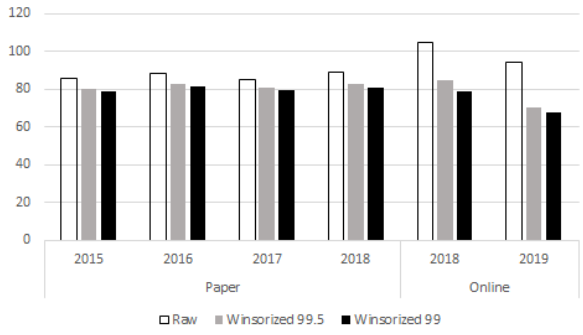
Table C.3: Comparing the 2018 and 2019 online CFM outcomes–trend estimates

	Weighted means		Difference in	Linear regression
	2018	2019	weighted means	coefficient of time
Cash on hand	85	71	-14.13	-14.00
Cash threshold	32	40	8.17	8.75
Withdrawal frequency at ABMs	2.1	2.4	0.35	0.45
Withdrawal amount at ABMs	132	138	5.69	4.37

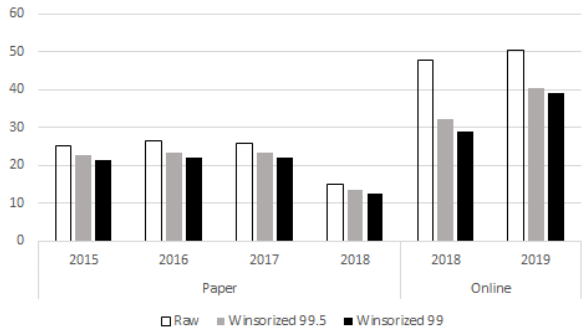
Notes: This table shows the estimated average responses in each year, and the estimated differences obtained using two different methods, as suggested in Gelman (2007). Column 3 shows the simple difference in weighted means, while column 4 reports the estimated coefficient of a year dummy in a linear regression also controlling for all the calibration variables used in the weighting. Response variables are winsorized at the 99.5th percentile before analysis.

Figure C.1: Comparing the paper-based and online CFM outcomes–mean estimates for some variables of interest

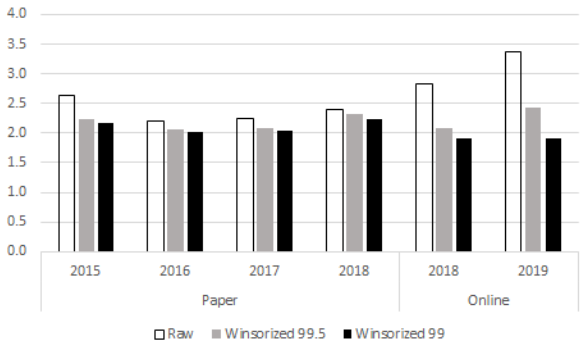
(a) Cash on hand



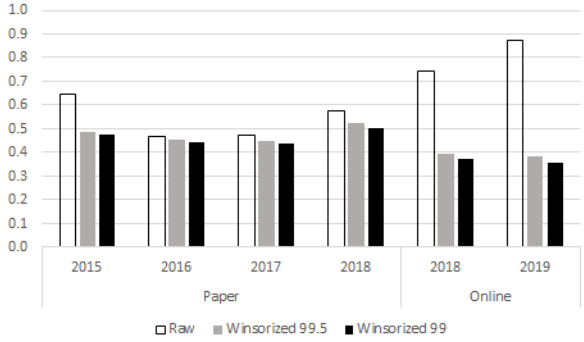
(b) Cash threshold



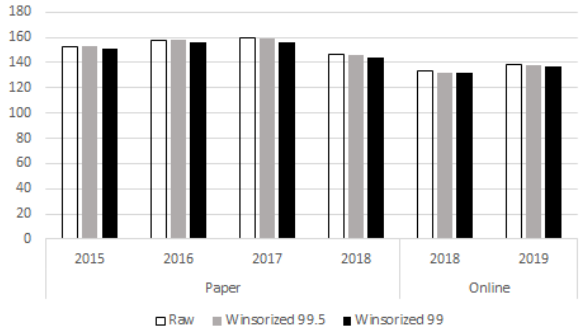
(c) Withdrawal frequency at ABMs



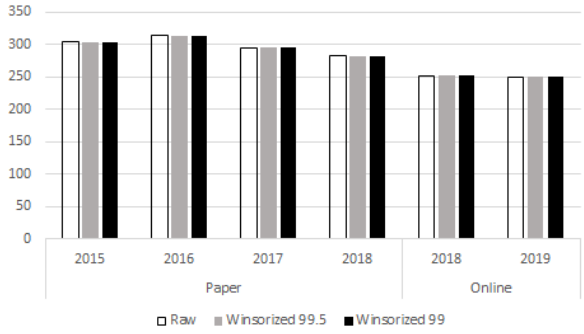
(d) Withdrawal frequency from tellers



(e) Withdrawal amount at ABMs



(f) Withdrawal amount from tellers



Notes: These graphs show weighted means of six variables of interest as measured in the paper CFM surveys from 2015 to 2018, and in the 2018 and 2019 online CFM surveys. *Cash on hand* is the dollar amount of cash in the respondent's purse, wallet or pockets right now; *Cash threshold* is how low the respondent typically lets the amount of cash in their purse, wallet or pockets fall before withdrawing more cash; *Withdrawal frequency at ABMs* (resp. *from tellers*) is the number of times the respondent withdrew cash from an ABM (resp. from bank tellers) in the past month; *Withdrawal amount at ABMs* (resp. *from tellers* is the typical withdrawal amount at an ABM (resp. from a bank teller); see Table C.1. Weighted estimates from the paper-based CFM surveys are obtained using household weights (the default weights provided by Ipsos), while those from the online CFM surveys are computed using the final calibrated CFM-MP weights w^f_{mp}. Mean estimates of raw and winsorized variables at the 99.5th and 99th percentile are shown.

D Considerations on variance estimation

For variance estimation, we propose the following three options for further consideration:[19]

1. Create bootstrap weights as has been done for previous surveys used by the Bank of Canada, e.g., for the 2013 and 2017 Methods-of-Payment surveys; see Chen and Shen (2015) and Chen et al. (2018). This approach would use the same Rao-Wu-Yue bootstrap methodology (Rao et al. 1992) that is typically used for stratified random samples: (i) proceed as if the CFM sample is a random sample stratified by region, age, gender and income and replicate the sample accordingly using the initial weights from Section 5.1.2; (ii) calibrate the weights for each replicate.

2. Group the sample into clusters based on the final weights from Section 5.1.3. Use these clusters as strata and treat these strata as though we have a random stratified sample. Create bootstrap weights from these weights using replicated sampling, then calibrate the weights for each replicate.

3. Create bootstrap replicates from either the initial weights or the final weights using a generalized bootstrap process as described in Beaumont and Patak (2012).

[19] We thank Jean-François Beaumont from Statistics Canada for fruitful discussions regarding this issue.