WORKING PAPER 66

Working Paper

Calibrated Multilevel Regression with Poststratification for the Analysis of SMS Survey Data

By Jonathan Gellar (Senior Statistician, Mathematica, Washington, DC. JGellar@mathematica-mpr.com (Corresponding author)), Constance Delannoy (Doctoral Candidate, Department of Applied Mathematics, University of Colorado Boulder), Erin Lipman (Data Scientist, Mathematica, Chicago, IL), Shirley Jeoffreys-Leach (Senior Specialist: Data & Analytics, FinMark Trust, Johannesburg, South Africa), , Bobby Berkowitz (Senior Specialist: Data & Analytics, FinMark Trust, Johannesburg, South Africa), Grant J. Robertson (Executive: Data & Analytics, FinMark Trust, Johannesburg, South Africa), and Sarah M. Hughes (Senior Fellow, Mathematica, Chicago, IL)

June 26, 2021

Acknowledgements:

This work was carried out under grant # OPP111490 to the FinMark Trust insight2impact facility (FMT i2i) from the Bill & Melinda Gates Foundation and with support from Mathematica. The Innovative and Sustainable Data: Mixed Modes Data Collection and Modelling project was led by FinMark Trust with the participation of Geopoll, which conducted the SMS surveys. The authors would like to thank Jessica Gao and Erin Lipman of Mathematica for computational support, and Mariel Finucane and Jesse Chandler of Mathematica for helpful comments on the content of this manuscript. Address correspondence on mobile data collection and survey design to Shirley Jeoffreys-Leach: Shirleyj@finmark.org.za. Address correspondence on calibrated MRP to Jonathan Gellar: JGellar@mathematica-mpr.com

Abstract

Face-to-face (FTF) surveys have been the traditional method to gather nationally representative data and remain the dominant data collection mode in resource-poor countries. Conducting these surveys is expensive and time consuming. With the rapid expansion of mobile phone use, Short Message Service (SMS) presents an opportunity to conduct inexpensive, fast, and scalable surveys. However, these samples are typically not representative of the target population. Standard adjustments to correct for nonrepresentative sampling are insufficient, due to two types of bias: residual sampling bias based on unobserved variables, and survey mode effects. We introduce calibrated multilevel regression with poststratification (cMRP), a procedure that corrects for residual bias by incorporating a relatively small sample of FTF data that is known to be unbiased. We apply this method to the problem of estimating financial inclusion (access to formal banking systems) in Uganda. We find that our cMRP approach is effective in replicating estimates from a larger and much more expensive FTF survey. This paper includes a description of our methods as well as results from the financial inclusion study and a discussion of limitations and future areas for research.

Keywords: Multilevel regression with poststratification, SMS survey, calibration, representative data, financial inclusion

I. Introduction

Face-to-face (FTF) surveys have traditionally been the primary method to gather nationally representative data in low and middle-income countries. However, those surveys are expensive and time consuming. In a world increasingly dominated by mobile phones, using Short Message Service (SMS) to conduct surveys is a cheap, fast, and scalable alternative to FTF surveys. Unfortunately, these surveys will only reach the portion of the population that has access to a mobile phone, which, in resource-poor countries, is skewed toward urban and peri-urban younger men; Lau, Lombaard, Baker, Eyerman, & Thalji (2019) found that SMS surveys underrepresent women, older people, those with less education, and those with less technological capability. Therefore, any inference for a national population based on SMS survey data must account for nonrepresentative sampling or they could be subject to sampling bias.

Accounting for nonrepresentative sampling is not a new problem. The standard solution is to adjust the sample results through respondent-level weights so that the weighted sample appears as close as possible to the target population on a set of auxiliary variables, such as demographic characteristics. For example, under quasi-randomization (Valliant, 2020), a binary regression model is used to approximate the inclusion probabilities for each survey respondent, and estimates are calculated using inverse probability weighting. An alternative approach, commonly referred to as raking, uses an algorithm that generates a set of weights such that the weighted sample appears as close as possible to the target population along the margins of each auxiliary variable (Deming & Stephan, 1940).

Another approach to adjust for nonrepresentative sampling is model-based prediction, in which the researcher fits a statistical model to the survey sample and uses that model to estimate outcomes in the target population (Little, 1993; Valliant, Dorfman, & Royall, 2000). Though this model can take many forms, recent work has shown that hierarchical models that borrow strength across strata demonstrate increased efficiency and the ability to account for more auxiliary variables (Gelman & Little, 1997; Park, Gelman, & Bafumi, 2004). The modern implementation of this approach, known as multilevel regression with poststratification (MRP, or "Mister P"), is now widely used across the social sciences (Ghitza and Gelman 2013; Wang et. al. 2015; Trangucci et.al. 2018) and has been called the "gold standard for estimating preferences from national surveys" (Selb & Munzert, 2011).

In this paper, we explore whether MRP can be used to obtain unbiased and efficient estimates for national indicators based on SMS survey data. We find that, while MRP successfully adjusts for sampling bias that is explained by observed characteristics, it fails to adjust for two types of residual bias that are common to SMS surveys: *residual sampling bias* and *survey mode bias*. We define residual sampling bias as systemic errors in estimation induced by differences between the SMS sample and the target population that we cannot control for, and the survey mode bias to be differential response that is directly due to the mode in which the respondent was surveyed. We then propose a novel extension, calibrated MRP, that addresses both types of bias simultaneously. Calibrated MRP adjusts for residual bias in the SMS survey by incorporating a small FTF sample that is assumed to be free of these residual biases. We apply this method to estimate several indicators related to financial inclusion in eight countries in Africa and Asia.

We are not the first researchers to suggest incorporating a second survey mode in order to calibrate estimates from a biased survey. Elliott and Davies (2005) suggested adjusting case weights of the biased survey using a propensity score model, so that the distribution of propensity scores in the two samples are similar. Though this approach addresses observed sampling bias, it ignores the residual biases described above. Elliott and Haviland (2007) proposed an estimator that is a weighted average of separate estimates derived from each survey mode, where the weights are based on the estimated mean squared error of each mode. Raghunathan et. al. (2007) similarly suggest combining mode-specific estimates, but they combine these estimates at a small-area level using a Bayesian hierarchical model. As far as we know we are the first to suggest an MRP-based approach that incorporates multiple modes of data, addressing residual bias through a single unified model.

This study is motivated by work carried out by Finmark Trust's Insight2Impact facility (FMT i2i) to estimate rates of financial inclusion in low and middle-income countries. FMT i2i studies access to financial tools among marginalized groups, including women, the poor, and rural populations, all of which are also typically underrepresented in SMS surveys. FMT i2i strives to improve data collection while reducing costs, and this study therefore has the dual aim of estimating financial inclusion for poor rural women in several countries and of determining the optimal approach to sampling in terms of data collection costs and estimates' accuracy.

II. Context and Survey Design

Financial inclusion, or access to financial products and services, is a key enabler for sustainable development in countries of all income levels. It permits individuals to invest in themselves through education and business ownership, to weather financial shocks, and to increase savings (Demirguc-Kunt, Klapper, Singer, & Oedheusden, 2015). Yet more than one billion individuals worldwide remain excluded from access to basic financial services (World Bank Group, 2017). With this large financially excluded population in mind, the World Bank, policymakers, and private sector partners set a goal of achieving universal financial access by 2020.

Nonetheless, measuring financial inclusion poses challenges in many parts of Africa and Asia. The traditional method to doing so is through detailed face-to-face surveys. Among a few existing surveys is the Financial Inclusion Insights (FII) survey, which has had up to six survey rounds in 14 countries since 2013 (Financial Inclusion Insights, 2020). FII data collection is carried out in person among a nationally representative sample and requires training a team of interviewers who travel nationwide. This FTF mode of data collection is costly in general but can be particularly expensive when it requires interviewing respondents in remote parts of a country to achieve national representativeness.

To explore a lower-cost alternative to the FII, FMT i2i and Mathematica designed an SMS survey that aligned with the FII on key financial inclusion indicators. SMS surveys are conducted using a self-completion survey method in which the questions are delivered to participants via SMS and respondents answer by replying with an SMS. For this study, respondents received a small airtime incentive on completion of the survey to mitigate the cost of responding. The sample comes from a database of validated mobile numbers sourced in partnership with in-country mobile network operators. This database is then classified into two groups, an active and inactive database. The active database includes those people who have responded to a survey in the past, and the inactive includes those who have not. Response rates for new surveys tend to be higher among the active database, and this database contains some profiling information that assists in reaching the desired target audience, for example, by age, gender, and region. Half of each sample was sourced using the inactive database and half was sourced using the active database. Quotas were set on age, gender, region, and urbanicity with the aim of achieving higher coverage than a simple random selection would produce.

The survey was administered in eight countries of varying size and geographic region: Uganda, Tanzania, Kenya, and Nigeria in Africa and Pakistan, India, Indonesia, and Bangladesh in Asia (Jeoffreys-Leach, Grundling, Robertson & Berkowitz, 2020). Because our findings regarding the utility of SMS surveys are consistent across countries, this paper focuses on the results for Uganda, which are based on 1,362 SMS respondents and 3,001 FII respondents.

III. Methods

3.1. Poststratification

Suppose we are interested in the population mean $\theta = \frac{1}{N}\sum_i Y_i = \bar{Y}$ for a particular outcome Y, where N is the size of the total population, and we collect data from a sample S from that population of size $n < N$. When the sample is not representative of the target population, the sample mean $\frac{1}{n}\sum_{i\in S} Y_i$ could be a biased estimate of θ. In poststratification, the strategy to estimate the population mean from a nonrepresentative sample is to divide the population up into mutually exclusive groups, which we call *poststratification cells* (Little, 1993). These cells are typically formed as the unique combinations of a set of *poststratification variables* that describe the ways in which members of the sample differ from the target population. The list of poststratification variables should include any potential confounders, that is, variables that are associated with both the outcome and selection into the sample. If these confounders are continuous variables, they are discretized to form categorical variables (for example, age categories) so that mutually exclusive cells can be formed.

Let j index the cell and C_j indicate the set of individuals in the jth cell. We can express the population mean as a weighted average of the means within each cell:

$$\theta = \frac{1}{N}\sum_i Y_i = \frac{1}{N}\sum_j \sum_{i:i\in C_j} Y_i = \frac{1}{N}\sum_j N_j\bar{Y}_j = \sum_j W_j\bar{Y}_j \tag{1}$$

where N_j is the population size of the jth cell, $\bar{Y}_j = \frac{1}{N_j}\sum_{i:i\in C_j} Y_i$ is the cell mean, and $W_j = \frac{N_j}{N}$ is the relative size of the cell in the target population (the *poststratification weight*). Equation (1) suggests that if the poststratification weights are known, we can obtain an unbiased estimate of θ if we have unbiased estimates of the cell-specific means, $\theta_j \equiv \bar{Y}_j$. For estimates $\{\hat{\theta}_j\}$, the poststratified estimate of the population mean is

$$\hat{\theta} = \sum_j W_j\hat{\theta}_j \tag{2}$$

Poststratification can also be used to estimate outcomes for any subgroup of the population that is defined as a subset of poststratification cells. If S is one such subset, an estimate of the outcome for the subgroup is obtained by averaging cell-specific means in the corresponding poststratification cells: $\hat{\theta}^S = \sum_{j\in S} W_j\hat{\theta}_j \,/\, \sum_{j\in S} W_j$.

The most common choice for $\hat{\theta}_j$ is the sample mean within each cell (Levy & Lemeshow, 2008), but when the number of cells is large, there will be cells with little to no sample and cell-specific means will be either highly unstable or impossible to compute.

3.2. Multilevel regression with poststratification

In MRP, the cell-specific means θ_j are instead estimated using a Bayesian multilevel regression model (Gelman & Hill, 2006). The model borrows strength across similar cells in order to maximize efficiency (Gelman & Little, 1997), especially for cells with little to no sample. This allows poststratification to be performed with more poststratification variables than the traditional implementation of poststratification (based on sample means), thereby reducing bias due to confounding that is unaccounted for.

The specification of the multilevel regression model is flexible and can be tailored to the application. Let $k \in \{1, \dots, K\}$ index the poststratification variables, and $l \in \{1, \dots, L_k\}$ index the level of the k^{th} poststratification variable. Also let y_i be the outcome for respondent i, and x_{ikl} be the (0/1) indicator that poststratification variable k takes the value l for respondent i. For a binary outcome, a simple multilevel logistic regression model could take the following form:

$$\begin{gathered} \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{k=1}^{K}\sum_{l=1}^{L_k} \alpha_l^k x_{ikl} \\ \alpha_l^k \sim N(0, \sigma_k) \qquad \forall k \end{gathered} \tag{3}$$

where $p_i = \Pr(y_i = 1)$. The second row of the equation specifies a shrinkage prior that identifies the model as a multilevel (or hierarchical): the coefficients corresponding to the different levels l of poststratification variable k are assumed to come from a normal distribution; these priors promote borrowing of strength across the different levels of each poststratification variable. Finally, we follow the advice of Gelman (2005) and place a half-normal hyperprior on the variance components, i.e. $\sigma_k \sim N^+(0, \tau)$, to promote identifiability.

The multilevel regression model can be extended to include, for example, interactions between different poststratification variables. The full specification of the model including all two-way interactions can be found in Appendix A of the supplemental material. Appendix B contains sample R code that uses the `brms` package to fit the regression model (Bürkner, 2018).

3.3. Assumptions of MRP in the context of SMS surveys

The key assumption in any poststratification approach is that the cell-specific estimates (which in MRP are based on the multilevel regression model) are unbiased for the cell-specific population means. That is,

$$\hat{\theta}_j^{MRP} \equiv E^{MR}\left[Y \middle| X_j\right] = \theta_j \tag{4}$$

where $E^{MR}[Y|X]$ is the expected value of Y based on the multilevel regression model. For this assumption to hold in MRP, several criteria must be met. Some of these are our typical regression modeling assumptions, in particular that we have specified an appropriate functional form for the model, including any necessary interactions between predictors. Because we are fitting a Bayesian model, the assumptions include the use of an appropriate prior distribution for model parameters—these assumptions are particularly important for estimating cells with little or no survey data because they specify how information is borrowed from neighboring cells in order to produce these estimates. However, the most important assumption required for (4) to hold is that of "no unobserved confounders", in other words, that we are including any variable in the model that is associated with both outcomes and selection into the sample (Si, Trangucci, Gabry, & Gelman, 2019). Stated differently, we assume that within a poststratification cell, the poststratification variables account for all the differences in outcomes between those in the survey (from the SMS sample) and those we are interested in (for the target population).

The “no unobserved confounders” assumption is particularly risky for SMS surveys for two reasons. The first reason is the potential for residual sampling bias—that the SMS sample may differ from the target population even within poststratification cells, in ways that are associated with outcomes. We attempt to mitigate this risk by including all poststratification variables that could account for differences in the populations, but there could be variables associated with both SMS survey participation and outcomes that we either are unable to measure or neglect to include. For example, access to a cellular phone could be a key confounder for some outcomes, but we cannot control for it because an SMS sample does not contain any information about individuals without cellular phone access.

The second reason that SMS surveys are susceptible to residual confounding is survey mode bias. The same exact population could respond differently to the survey questions if asked the question over SMS message, compared to being asked the question in a face-to-face survey. There are several reasons for this. For example, question wording may differ between the two surveys because SMS questions must be shorter to fit within a fixed character limit. SMS respondents may also be more likely to misinterpret a survey item because there is no interviewer who can clarify the question. Additionally, SMS respondents may be more prone to “straight-lining responses” (rushing through a survey and selecting identical responses to each survey item) because of lack of motivation to complete the survey. In each of these cases, the confounding variable is the survey mode itself. Unfortunately, that variable cannot be adjusted for when all survey responses are of the same mode.

Another assumption of MRP (or any poststratification procedure) that we have ignored until now is that the poststratification weights W_j must be known. In practice these weights will rarely be known exactly, and usually this information will come from a large national survey. From this survey the poststratification weight for cell j can be estimated as $\widehat{W}_j = \sum_i w_i I_{ij} / \sum_i w_i$, where w_i is the case weight for observation i in the national survey, and I_{ij} is the 0/1 indicator that observation i falls in poststratification cell j. We explore the effect of having inexact estimates of poststratification weights through simulations in Section 4.

3.4. Calibrated MRP

Since we cannot correct for residual sampling bias or survey mode bias when all the data come from the same mode, we propose a simple solution: we augment the sample with a small amount of data from another survey mode that we can assume to be free of residual bias within cells. In practice this unbiased sample will typically be more expensive to collect per respondent, and have a smaller sample size, than the SMS survey. For example, it can be a smaller version of the FTF survey that we are intending to replace. We then fit a regression model to the combined dataset that borrows strength across the two survey modes, leveraging the larger size of the SMS sample with the unbiased property of the smaller FTF sample to provide efficient estimates of how a FTF participant would respond to the survey question. We can then poststratify these estimates to any target population of interest.

We do not require that the “smaller sample” FTF survey be conducted on a sample that is representative of the target population, though in practice it will often be more representative than the SMS survey. For example, SMS surveys are often subject to sampling biases that do not apply to the FTF survey, as cell phone ownership may be more common for certain segments of the population. Regardless, we do assume that we can adjust for the nonrepresentative sampling of the unbiased survey by applying MRP to the observed auxiliary variables (i.e., no unobserved confounders). Mathematically, we assume that $E\left[\hat{\theta}_j^{FTF,MRP}\right] = \theta_j$, where $\hat{\theta}_j^{FTF,MRP}$ is the MRP-based estimate for each cell j based on the FTF survey. One option for obtaining national estimates would be to ignore the SMS data altogether and simply apply MRP to the smaller sample FTF data. Although this approach will result in unbiased estimates, these estimates could be more uncertain because of the small sample of FTF data that is expected to be available.

Instead, we propose an approach that optimally leverages both the SMS and FTF data in order to achieve more efficient estimates. We refer to the procedure as calibrated MRP (cMRP), and it consists of the following steps:

1. Stack the SMS and FTF data into a single dataset, retaining the information about the survey mode (SMS versus FTF) as a covariate.
2. Fit a multilevel regression model to the stacked dataset. This model estimates outcomes conditional on both the poststratification variables and the survey mode.
3. Estimate the cell-specific means (θ_j) for each poststratification cell under the *unbiased* survey mode: $\hat{\theta}_j^{cMRP} = E^{MR}\left[Y \middle| X_j, mode = FTF\right]$.
4. Poststratify the resulting estimates using equation (2), using the cMRP-based estimates $\hat{\theta}_j^{cMRP}$ for $\hat{\theta}_j$.

The intuition behind the procedure is relatively straightforward. We first build a model that estimates the mean outcome for any combination of auxiliary variables (i.e., poststratification cell) and survey mode. We then use only the estimates we consider to be unbiased, which are those that correspond to the FTF survey mode. Importantly, these estimates should be free of both residual sampling bias and survey mode bias, as long as our assumption holds that the FTF survey is free of these biases after adjusting for observed covariates. In addition, we note that these unbiased estimates are informed by both SMS and FTF data by borrowing strength across survey modes, an essential feature because we expect the sample size of FTF data to be relatively small. At a high level, the larger SMS dataset is used to understand which poststratification cells have higher versus lower rates of the outcome, and the smaller FTF sample is used to adjust those estimates up or down to calibrate them to how FTF survey participants would respond.

We estimate the Bayesian multilevel regression model using Markov Chain Monte Carlo sampling, which provides the posterior distribution for all model parameters. Since our target estimands are linear combinations of these model parameters, the uncertainty around the parameters propagates naturally to the poststratified estimates (Gelman & Hill 2006).

To illustrate, consider a simplified scenario where we have two dichotomous auxiliary variables, gender (male/female) and urbanicity (urban/rural), and two survey modes, SMS and FTF. The two dichotomous auxiliary variables create four poststratification cells, and we wish to estimate the mean of the outcome in each of these cells. Following the above procedure, we first combine the two datasets into a single dataset that contains variables for the outcome, gender, urbanicity, and survey mode. Next, we fit a multilevel regression model to estimate $E[Y|X, M]$, where the covariates X_i include gender and urbanicity and M_i represents the survey mode. We then use this model to calculate $\hat{\theta}_j = E^{MR}\left[Y \middle| X = x_j, M = FTF\right]$ for each of the four poststratification cells, where x_j is the combination of auxiliary variables that corresponds to cell j. Finally, we poststratify by applying equation (1) to the cell-specific estimates: $\hat{\theta} = \sum_{j=1}^{4} W_j \hat{\theta}_j$, where the poststratification weights W_j come from an external source and $\sum_{j=1}^{4} W_j = 1$.

As in the standard implementation of MRP, the procedure is flexible in the form of the multilevel regression model. However, we recommend including interactions between survey mode and the poststratification variables, thereby allowing the calibration for sampling mode to vary for each poststratification cell. Thus, a reasonable version of the regression model for the case with a binary outcome would be

$$\begin{gathered} \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta M_i + \sum_{k=1}^{K} \sum_{l=1}^{L_k} \left(\alpha_l^k x_{ikl} + \gamma_l^k x_{ikl} M_i\right) \\ \alpha_l^k \sim N(0, \sigma_k) \qquad \forall k \\ \gamma_l^k \sim N\left(0, \sigma_k^M\right) \qquad \forall k \end{gathered} \tag{5}$$

The full specification (including priors) of this model as well as an extension that includes interactions between poststratification variables can be found in Appendix A in the online supplement, along with sample `brms` code in Appendix B.

IV. Simulation Study

We conducted a simulation study in order to assess the performance of cMRP compared to alternative approaches, under a range of conditions. We consider a total of 36 simulation scenarios, which constitute all combinations of the following four parameters:

1. $\alpha \in \{0,0.5,1\}$: the level of observable sampling bias for the SMS survey
2. $\gamma \in \{0,0.5\}$: the level of residual bias for the SMS survey
3. $N^{FTF} \in \{150,300\}$: the size of the face-to-face sample
4. $N^{ref} \in \{3000,10000,\infty\}$: the size of a reference dataset containing poststratification variables in the target population. A size of ∞ indicates that the joint distribution of the poststratification variables is known exactly.

Note that these simulations do not explicitly distinguish between residual sampling bias and survey mode bias, because mechanistically their effects on outcomes are identical. Instead, the parameter γ controls the combined effect of both types of residual bias. We run each scenario for 100 iterations, for a total of 3600 simulated datasets. The next section describes more specifically how the four parameters are used to generate data.

1.2. Data generating process

4.2.1. Sampling poststratification variables

We consider 5 poststratification variables for the purposes of this simulation: age category (5 levels), gender (2 levels), rural status (2 levels), region (4 levels), and education (5 levels). In order to generate realistic distributions of these variables, we sample respondents from the Uganda financial inclusion data. As previously mentioned, this data consists of two samples: an SMS survey, and a face-to-face survey (the FII). For the purposes of the simulation, we also require two samples (an SMS survey and a face-to-face survey). To generate the face-to-face survey, we sample N^{FTF} respondents from the Uganda FII data, with replacement. To generate poststratification variables for the SMS sample, we sample αN^{SMS} respondents from the Uganda SMS data, and $(1-\alpha)N^{SMS}$ respondents from the Uganda FII data. In this way, α controls the representativeness of the resulting SMS sample. When $\alpha = 0$, the SMS sample will be defined based entirely on the Uganda FII data; when $\alpha = 1$, it will be defined based on the less representative Uganda SMS data. We set $N^{SMS} = 1500$ for all scenarios, as this is the approximate size of the SMS sample in the Uganda data. The two samples (SMS and face-to-face) are combined into a single dataset, retaining the survey mode as a variable M_i ($M_i = 0$ for face-to-face, $M_i = 1$ for SMS).

In addition to the SMS and face-to-face samples, we assume a reference dataset of size N^{ref} is available, which contains information on the joint distribution of the poststratification variables in the target population. We define the true joint distribution of the poststratification variables to be the distribution implied by the weighted Uganda FII data (these weights were generated through raking by the FII survey administrators and provided along with the data). For finite values of N^{ref}, we generate a reference dataset by drawing a weighted sample (using the raked weights) of size N^{ref}, with replacement, from the FII data. When $N^{ref} = \infty$, we assume the entire weighted FII dataset is available.

4.2.2. Generating outcomes

For each observation, we generate outcomes using a gradient-boosted model (GBM) that is fit to the Uganda FII data. Gradient-boosting is an ensemble, nonparametric machine learning technique built by combining many weak learners. Our GBM implementation uses the `gbm` package in R to estimate the probability of a respondent having an active mobile money account in Uganda, conditional on the five poststratification variables. This model allows for two-way interactions between all predictors, with the optimal number of boosting iterations estimated using 5-fold cross-validation. Note that as this model is fit to the Uganda FII data (not the sampled data for the simulation), the model only needs to be fit one time.

Let X_i be the vector of poststratification variables for sampled respondent i, and $\hat{f}^{GBM}(X_i)$ be the predicted log odds based on the gradient-boosted model. We generate outcomes $\{Y_i\}$ for all respondents in our sample (SMS or face-to-face) according to the following model:

$$\log\left(\frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = 1)}\right) = \hat{f}^{GBM}(X_i) + \gamma M_i \quad (6)$$

Thus, the parameter γ controls the residual bias: it is the increase in the log odds that $Y_i = 1$ when an individual is sampled via SMS. We assume that this bias is the same for all respondents on the log odds scale.

1.2. Estimation and performance

For each iteration of each of the 36 simulation scenarios, we generate 26 estimates. The 26 estimates differ from one another in three ways:

a. **Target population.** We consider two different target populations. The first is the full Uganda population, as defined by the joint distribution implied by the weighted Uganda FII data. The second is the subgroup of the Uganda population corresponding to lower-educated, rural women.

This subgroup constitutes 26% of the target population, but just 2.3% of the SMS sample in these simulations.

b. **Source data.** The estimates can be based on the SMS sample alone (SMS), the face-to-face sample alone (FTF), or a combined sample including both SMS and face-to-face data (CMB).

c. **Estimation method.** We consider the following estimation approaches. Note that not all estimation methods apply to each data source (discussed below).

   i. *Unadjusted*. We calculate a simple mean of the outcomes in the source data.

   ii. *Inverse probability weighting (IPW)*. We follow the quasi-randomization procedure described by Valliant (2020). First, we generate pseudo-inclusion probabilities for each observation of the SMS data based on a logistic regression model. This model is a logistic regression model fit by combining the source data (SMS, FTF, or CMB) with the reference data, to estimate the likelihood that each observation is in the source data. We then generate estimates for each target population using inverse-probability weighting.

   iii. *Standard MRP*. The standard implementation of MRP, described in Section 3.2. We estimate two versions: one that uses a multilevel regression model with main effects only (MRP1, following equation 3 or Model A1 in Appendix A), and another that includes two-way interactions between poststratification variables (MRP2, following Model A2).

   iv. *Blended*. We follow the procedure of Elliott and Haviland (2007) to combine the IPW-based estimates from the SMS sample with that from the FII sample. The final estimate is a weighted average of the IPW estimates corresponding to SMS-only and FTF-only sources, where the weights are based on the estimated mean-squared error of each sub-estimate.

   v. *Calibrated MRP*. Calibrated MRP as described in Section 3.4. We estimate two versions: one that uses a multilevel regression model following equation 5 (cMRP1, following equation 5 or Model A3), and another that includes two-way interactions between poststratification variables (cMRP2, following Model A4).

For the estimates based on source data that comes from a single sample (either the SMS only or the face-to-face only sample), we apply four estimation methods (i-iii, using two versions of MRP). For those based on the combined (SMS and face-to-face) data, we apply five estimation methods (i, ii, iv, and v, using two versions of cMRP). This leads to a total of 13 combinations of source data and estimation method. Each of these 13 approaches are used to estimate the rate in both target populations (26 total estimates).

For each of the 36 scenarios, we compare the 26 estimates to the true rate of the outcome in the corresponding target population (full country or subgroup). The true rate is defined by first applying the GBM to predict the outcome in each observation of the Uganda FII dataset, and then taking a weighted average of these predicted probabilities over the corresponding target population, using the raked weights provided with the FII dataset. We summarize the performance of each method in each scenario by calculating their bias, variance, and mean squared error (MSE).

1.2 Results

Table 1 presents the simulation results corresponding to the subset of the scenarios where $N^{FTF} = 150$, $N^{ref} = \infty$, and the estimation target is the full population. Results are presented as the MSE, with the absolute value of the bias in parentheses. Approaches that use only SMS data perform very well when there is no residual bias and very little observable sampling bias. However, these approaches perform extremely poorly in the presence of residual bias. On the other hand, the approaches that use only the face-to-face dataset perform similarly across all scenarios – this is expected because the observable sampling bias and residual bias do not affect the face-to-face sample. Though all approaches that use only FTF data are unbiased, they also tend to have more variance than the SMS-only approaches due to the small sample of face-to-face data. Among the eight single-source approaches (SMS or FTF), the MRP methods consistently outperform the unadjusted and IPW approaches. Including interactions in the multilevel regression model improves the performance of calibrated MRP in the presence of survey mode bias (cMRP2 vs. cMRP1) but slightly worsens its performance when no survey mode bias is present due to an increase in variance. The increase in variance observed in the models that include interactions is due to using a regression model with more terms in it that are ultimately unnecessary (this complexity is not part of the data-generating mechanism). The interactions do not have a substantial effect on the performance of standard MRP.

Table 1. Simulation results by level of bias (α and γ), for the scenarios with $N^{FTF} = 150$, $N^{ref} = \infty$, and the target population is the full population. Estimates are presented as MSE on the percentage point scale (outcomes are percentages out of 100, not out of 1), with the absolute bias in parentheses.

		No residual bias ($\gamma = 0$)			With residual bias ($\gamma = 0.5$)		
Source	Method	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	1.5 (0.5)	301.2 (17.3)	1133.2 (33.6)	92.3 (9.5)	645.0 (25.4)	1710.8 (41.3)
SMS	IPW	1.4 (0.2)	12.6 (3.3)	91.6 (9.2)	85.9 (9.2)	155.2 (12.4)	372.9 (19.1)
SMS	MRP1	1.2 (0.0)	1.9 (0.5)	11.9 (2.4)	83.8 (9.1)	92.8 (9.5)	145.5 (11.7)
SMS	MRP2	1.3 (0.0)	1.9 (0.5)	11.7 (2.3)	84.1 (9.1)	92.1 (9.5)	144.4 (11.7)
FTF	Unadj.	12.9 (0.2)	14.8 (1.1)	14.1 (0.3)	18.8 (0.7)	13.8 (0.7)	16.4 (0.4)
FTF	IPW	14.3 (0.4)	11.5 (0.3)	13.8 (0.1)	12.2 (0.4)	11.0 (0.4)	14.6 (0.3)
FTF	MRP1	12.6 (0.3)	11.1 (0.6)	12.2 (0.0)	13.5 (0.5)	9.6 (0.5)	12.8 (0.4)
FTF	MRP2	12.7 (0.2)	10.6 (0.5)	12.1 (0.0)	13.5 (0.5)	9.3 (0.5)	13.4 (0.4)
CMB	Unadj.	1.4 (0.4)	252.2 (15.8)	938.2 (30.6)	77.5 (8.7)	536.2 (23.1)	1416.6 (37.6)
CMB	IPW	1.2 (0.2)	10.8 (3.0)	75.8 (8.4)	71.7 (8.4)	129.3 (11.3)	309.5 (17.4)
CMB	Blended	6.7 (0.3)	8.9 (1.4)	16.3 (1.5)	16.0 (1.9)	16.1 (1.8)	17.7 (1.3)
CMB	cMRP1	6.7 (0.0)	5.8 (0.6)	7.6 (0.3)	14.6 (1.6)	11.3 (1.6)	13.7 (1.5)
CMB	cMRP2	8.2 (0.1)	6.8 (0.6)	8.3 (0.2)	13.4 (1.4)	10.3 (1.4)	12.7 (1.3)

cMRP seems to offer both low bias and low variance, when possible, and consistently performs well across all six scenarios presented in the table. Among each of the six scenarios, the absolute bias of cMRP is less than two percentage points. For cases with no residual bias it does not perform quite as well as the SMS-only approaches when there is no observable sampling bias, but it does outperform them when the observable sampling bias is higher. When there is residual bias, the approach is successful in leveraging the face-to-face data to provide unbiased estimates, despite a highly biased SMS sample. Performance in these scenarios is similar to the performance of the approaches that use only face-to-face data. Among the four approaches that combine SMS and face-to-face data, cMRP has the lowest MSE, except in the case where $\gamma = \alpha = 0$.

The relative advantage of cMRP compared to MRP using face-to-face data alone is more apparent when the target population is the subgroup of lower-educated rural women (Table 2). In these cases, cMRP outperforms MRP using only face-to-face data by a wide margin, in all six scenarios. The relatively small sample size for the face-to-face data causes the estimates based only on this dataset to be highly uncertain for this subgroup, but cMRP is able to leverage the information available in the larger SMS sample to obtain much more stable estimates. Interestingly, the unadjusted and IPW estimates that use SMS data are much more accurate for the smaller subgroup than for the full population. This is because these procedures use only the subset of the SMS data that corresponds to the subgroup, thereby providing a more homogenous sample of respondents that is not as susceptible to sampling bias.

Table 2. Simulation results by level of bias (α and γ), for the scenarios with $N^{FTF} = 150$, $N^{ref} = \infty$, and the target population is the subgroup of lower-educated, rural women. Estimates are presented as MSE on the percentage point scale (outcomes are percentages out of 100, not out of 1), with the absolute bias in parentheses.

		No survey mode bias ($\gamma = 0$)			With survey mode bias ($\gamma = 0.5$)		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	3.2 (0.2)	6.4 (0.6)	90.1 (5.3)	66.9 (7.9)	80.3 (8.6)	223.9 (12.6)
SMS	IPW	3.1 (0.1)	5.8 (0.1)	81.3 (4.4)	59.0 (7.4)	65.3 (7.6)	194.9 (11.5)
SMS	MRP1	2.5 (0.8)	4.1 (1.2)	32.4 (4.4)	73.8 (8.4)	85.2 (9.0)	177.8 (12.5)
SMS	MRP2	2.3 (0.6)	3.9 (1.0)	32.6 (4.2)	68.6 (8.1)	78.6 (8.6)	168.1 (12.1)
FTF	Unadj.	30.1 (0.2)	29.6 (0.9)	37.0 (0.1)	27.9 (0.2)	27.0 (0.8)	32.8 (0.2)
FTF	IPW	29.4 (0.6)	29.6 (0.5)	37.8 (0.1)	28.8 (0.1)	26.3 (0.3)	37.2 (0.4)
FTF	MRP1	38.2 (4.2)	39.1 (4.3)	39.6 (3.6)	34.4 (3.9)	33.8 (4.1)	31.3 (3.5)
FTF	MRP2	30.2 (3.0)	33.5 (3.6)	35.9 (2.9)	29.4 (3.0)	28.6 (3.4)	28.9 (2.6)
CMB	Unadj.	2.6 (0.2)	5.5 (0.7)	27.8 (2.2)	55.8 (7.2)	60.0 (7.3)	54.1 (5.2)
CMB	IPW	2.5 (0.1)	5.2 (0.0)	60.6 (3.7)	49.0 (6.8)	53.7 (6.9)	143.4 (9.8)
CMB	Blended	14.6 (0.8)	16.0 (0.0)	31.5 (0.1)	27.2 (1.4)	26.4 (1.8)	37.0 (0.6)
CMB	cMRP1	5.9 (0.9)	5.6 (1.4)	13.4 (2.0)	13.5 (2.2)	12.4 (2.3)	19.3 (2.6)
CMB	cMRP2	7.4 (1.1)	8.4 (1.5)	14.4 (1.8)	12.5 (2.1)	11.7 (2.1)	17.7 (2.3)

The trends in performance across α and γ hold for the remaining scenarios (those that include different values of N^{FTF} and N^{ref}, see Tables C1-C10 in Appendix C of the online supplement). Therefore Table 3 explores the effects of varying these parameters by focusing on the most interesting scenario, the one in which there are high degrees of both types of bias ($\gamma = 0.5$, $\alpha = 1$). We also limit the table to the MRP-based models, which generally outperform the other estimation methods for each corresponding data source.

Table 3. Simulation results for MRP-based models by N^{FTF}, N^{ref}, and target population, for the scenarios where $\alpha = 1$ and $\gamma = 0.5$. Estimates are presented as MSE on the percentage point scale (outcomes are percentages out of 100, not out of 1), with the absolute bias in parentheses. The Subgroup target population refers to lower-educated, rural women. SMS refers to MRP estimates based on SMS only, FTF refers to MRP estimates based on FTF only, and CMB refers to cMRP estimates that use both SMS and FTF data.

Target Population		$N^{FTF} = 150$			$N^{FTF} = 300$		
	N^{ref}	3000	10000	∞	3000	10000	∞
Full	SMS	140.3 (11.4)	145.5 (11.7)	145.5 (11.7)	143.2 (11.6)	140.4 (11.5)	140.0 (11.5)
Full	FTF	10.8 (0.1)	12.8 (0.4)	12.8 (0.4)	7.9 (0.8)	6.0 (0.2)	5.9 (0.2)
Full	CMB	11.1 (0.8)	13.7 (1.5)	13.7 (1.5)	8.3 (1.3)	6.7 (0.5)	6.6 (0.5)
Subgroup	SMS	174.9 (12.4)	177.6 (12.5)	177.8 (12.5)	178.2 (12.5)	169.0 (12.2)	166.5 (12.1)
Subgroup	FTF	31.3 (3.1)	30.9 (3.5)	31.3 (3.5)	16.5 (2.3)	13.6 (1.7)	13.6 (1.7)
Subgroup	CMB	16.3 (2.3)	19.3 (2.6)	19.3 (2.6)	12.0 (2.3)	8.1 (1.4)	8.1 (1.4)

We find the MSE improves slightly as N^{ref} increases for the scenarios where $N^{FTF} = 300$, but it tends to worsen slightly with increasing N^{ref} for scenarios where $N^{FTF} = 150$. This latter pattern is surprising, and we suspect it could be a spurious finding due to only having run our simulation for 100 iterations.

Generally, we find that a reference dataset of size 3000 is sufficient to provide unbiased estimation with cMRP, though there is some evidence to suggest that variance can be improved with a larger reference dataset. Performance of MRP using face-to-face data only and cMRP improves when N^{FTF} increases from 150 to 300, with variance decreasing by approximately 25%-50% for both estimation procedures.

Finally, we assess the validity of the estimated standard errors by calculating the coverage rates of the 95% credible intervals across the 100 simulation iterations. Table 4 presents these results for the scenario with the largest amount of bias ($\gamma = 0.5$, $\alpha = 1$), for the four estimators that use the combined (SMS and face-to-face) sample. The unadjusted and IPW estimators, which are highly biased, show poor coverage as expected. Both the Blended and cMRP estimators are very close (and sometimes above) the nominal 95% level in most scenarios.

Table 4. Coverage rates of 95% credible intervals for MRP-based models without interactions by N^{FTF}, N^{ref}, and target population, for the scenarios where $\alpha = 1$ and $\gamma = 0.5$. The Subgroup target population refers to lower-educated, rural women. Coverage rates are based on 100 simulation iterations.

Target Population	N^{ref}	$N^{FTF} = 150$			$N^{FTF} = 300$		
		3000	10000	∞	3000	10000	∞
Full	Unadj.	0%	0%	0%	0%	0%	0%
Full	IPW	0%	0%	0%	0%	0%	0%
Full	Blended	95%	94%	94%	94%	98%	98%
Full	cMRP	94%	92%	93%	90%	97%	95%
Subgroup	Unadj.	75%	81%	81%	83%	85%	85%
Subgroup	IPW	64%	71%	71%	74%	68%	68%
Subgroup	Blended	96%	91%	91%	96%	95%	95%
Subgroup	cMRP	90%	91%	91%	88%	93%	93%

V. Applied Experiment

5.1. Experiment design

In this section, we further explore the performance of cMRP relative to other approaches by conducting an experiment using the Uganda financial inclusion data. We distinguish this experiment from the simulation described above because in this experiment we use the observed outcomes in the financial inclusion data, rather than simulating new outcomes. From this data we analyze seven different measures of financial inclusion (Figure 1). Since the Uganda FII dataset is relatively large (3,001 participants), a reasonable approach would be to ignore the SMS data and base all estimates off this face-to-face dataset. However, the purpose of this paper is to understand the performance of different approaches for cases when a much smaller face-to-face dataset is available. Therefore, we conduct an experiment to compare the performance of each approach when a smaller face-to-face dataset is sampled at random from the full FII, under varying sampling schemes.

The FII survey will play two important roles in this analysis. First, we will use a random subset of the FII to augment our SMS data. This procedure intends to simulate the situation in which a policymaker seeking to avoid the high cost of collecting the full FII instead collected a smaller FTF sample to calibrate a larger but less expensive SMS sample. Second, the population estimates from the FII will function as "ground truth" against which we will compare estimates from our new calibrated MRP approach. As we describe below, we will define the ground truth based on the portion of the FII that was not selected as part of the FTF sample used for calibration.

We hypothesize that we can replicate the full FII results by combining SMS data with small samples from the FII. These samples will be either 5% or 10% of the total size of the FII survey for each country In Uganda, the full FII consists of 3,001 respondents, so the 5% and 10% samples will consist of 150 and 300 respondents, respectively. For our study, these samples were selected as simple random draws, without replacement. We repeat this procedure 10 times for each of the two sample sizes, for a total of 20 FII samples.

5.2. Estimation and performance

We estimate the level of each of the seven indicators of financial inclusion, separately for two target populations: the national population of Uganda and the subpopulation of poor rural women in the country. Poor rural women were specifically selected as a stratum of the population with high likelihood of being under-represented in SMS-based data collection. This stratum is also of specific interest to the financial inclusion community as the sub-population most likely to be financially excluded in Uganda and elsewhere.

As we did in the simulation study, we consider three different data scenarios: one in which only the SMS data is available (1,362 respondents), one in which only the small sample of FII data (150 or 300 respondents) is available, and one in which both samples are available. We implement the same estimation approaches that were introduced in Section 4.3, with two modifications:

1. Instead of implementing two versions of the multilevel regression model for the MRP-based approaches, we only implement the version that includes interactions (MRP2 and cMRP2 from the simulation). We do this because exploratory analyses suggested large degrees of survey mode bias in the data.
2. In addition to the IPW approach that is based on the logistic regression model (IPW-LR), we add a similar IPW approach that uses a gradient-boosted model to estimate the inclusion probabilities (IPW-GBM). The gradient-boosted model considers all three-way interactions between adjustment variables and is implemented using the `twang` package in R (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2014). The use of the latter model is intended to provide insight as to whether any differences in performance between MRP and IPW weighting are due to the importance of interaction terms in the corresponding model-based adjustment.

In total, if we consider the two propensity score models to be different approaches, and we also consider the use of 5% or 10% of the FII data to be different approaches, we estimate each indicator for each target population in 24 different ways: 4 that use only SMS data, 8 that use only FII data, and 12 that use both SMS and FII data.

Based on a literature review (Null, Chaplin, Hartog, Jacobson, & Rangarajan, 2015; Blumenstock, Cadamuro, & On, 2015; Johnson, 2016; Lau et al., 2019), we identified nine poststratification variables that could be associated with both outcomes and sample selection: age, gender, education, urbanization, literacy, region, electricity adoption, poverty status, and phone ownership. We code each of these variables into discrete categories. These variables are used for the IPW-based methods (in the binary regression model) as well as for MRP and cMRP.

After calculating each estimate using each approach, we compare the estimate to its target. Unlike the simulation study, where we were able to control and know the true rate of the outcomes, we do not know the true rate of any of the indicators in the population. Instead, we define the target to be the estimated rate of each indicator based on all FII data that was not included in the corresponding sample. In other words, for estimates that use a 10% sample of the FII, we define the target based on the remaining 90% FII sample, whereas for estimates that do not use any of the FII, we define the target using 100% of the FII sample. Target estimates are calculated using the sampling weights included with the FII data, which are based on population data by age, urbanization, and gender (InterMedia, 2016).

For each combination of country, target population (national or poor rural women), indicator, and data scenario, we summarize the performance of each estimation method by estimating its bias, variance, and mean squared error (MSE). We define the bias as the difference between the point estimate and the target and variance as the model-based estimate of the variance of the point estimate (i.e., the squared standard error). For data scenarios that include FII data, these numbers are averaged across the 10 repeated samples; for those that do not include FII data, they are based on a single sample. We estimate the MSE as the sum of the squared bias and the variance.

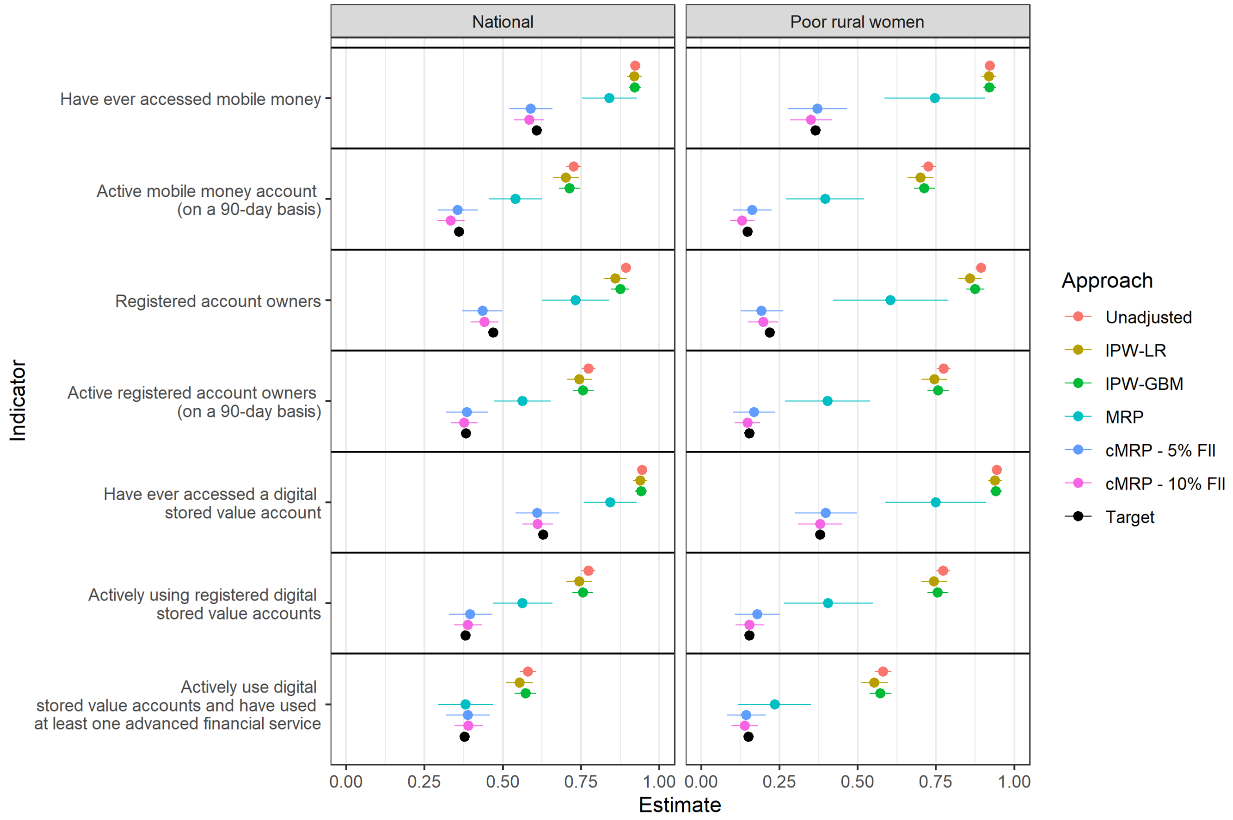
5.3. Results

As previously mentioned, we estimate the rate of each indicator in the two target populations using 24 different approaches. Two of these correspond to our cMRP approach (one using a 5% sample of FII and the other using a 10% sample), whereas the other 22 are potential competing approaches. Rather than comparing the results from all 24 approaches at once, we divide the presentation of results into three substantive groups.

5.3.1. How does cMRP compare to approaches that only use the SMS data?

First, we compare the two cMRP estimates to the four competing methods that only use SMS data. This comparison measures the marginal benefits of collecting a small, unbiased sample to augment an SMS sample. We display in Figure 1 the estimates for each of the seven outcomes and two target populations, using each of the six methods. Note that the cMRP estimates are based on only the first of the 10 repeated samples of FII data.

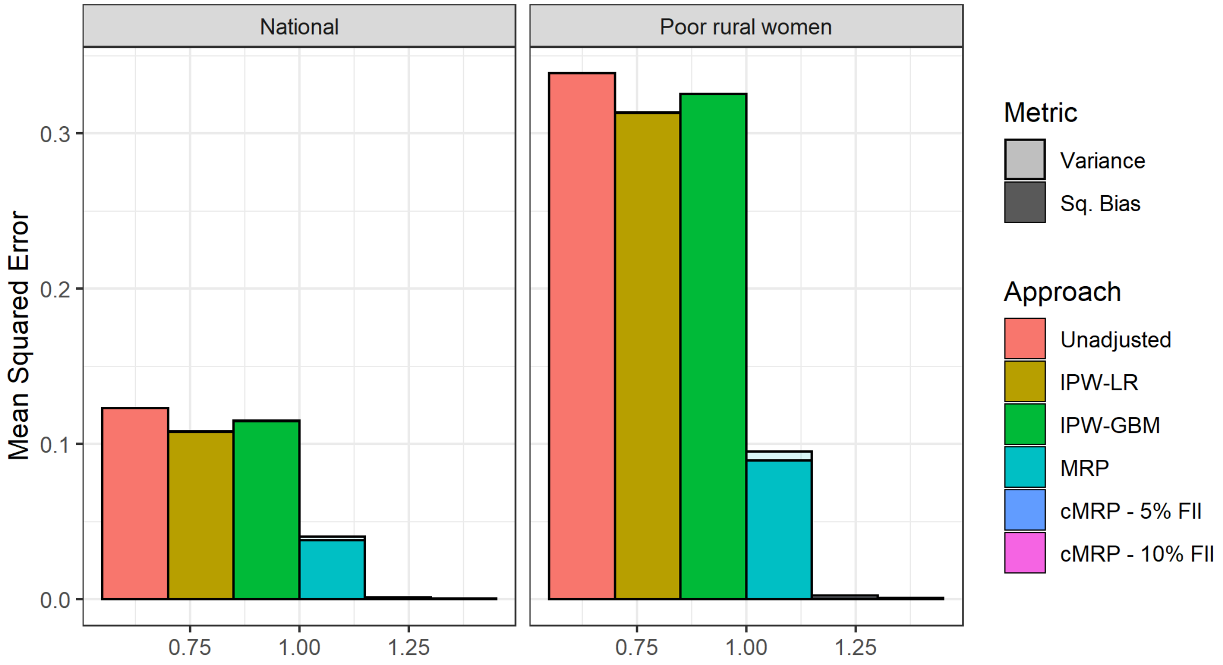
Figure 1. Comparison of estimates for each outcome and target population using cMRP to those based on only SMS data only. The two cMRP estimates are based on the first of the 10 repeated samples of FII data. Lines represent either 95% confidence intervals (for frequentist procedures) or 95% credible intervals (for MRP-based procedures). IPW-LR uses the logistic regression propensity model, and IPW-GBM uses the gradient boosted propensity score model.



We see that the unadjusted SMS results (red) are highly biased for each of the 14 target estimands (black), in that they all overestimate the rates of financial inclusion in both target populations. This is not surprising. Since SMS respondents must have access to a cellular phone, the sample overrepresents more highly educated, younger, wealthier men – exactly the population that is more likely to be financially included. IPW procedures (gold and green) do little to decrease bias. The unadjusted and IPW estimates also have very small standard errors, indicating that not only are these estimates wrong, but they are very confident in their wrong estimates. Standard MRP estimates using only SMS data (teal) are somewhat better in that they are consistently less biased than the IPW estimates and estimate larger standard errors, but the resulting 95% credible intervals rarely cover the target. The cMRP estimates (blue and pink), on the other hand, are all much closer to the target than any of the approaches based on SMS data alone, and the credible intervals for every estimate cover their target. Performing cMRP using a 10% sample of FII reduces the standard error of the resulting estimates by an average of 33% (averaged across the 14 estimands), compared to using a 5% sample of FII.

These observations are reiterated by comparing the MSE of the six different approaches (Figure 2). As discussed in Section 5.2, the MSE values for the cMRP approaches are averaged across the 10 repeated FII samples. For this visualization, we then average the MSE values for each approach and target population across the seven indicators and decompose MSE into variance and squared bias to highlight these differences. We see that the unadjusted and IPW methods have very high MSE, and this MSE is dominated by squared bias. Standard MRP (using only SMS data) is also fairly biased. cMRP (using either the 5% or 10% sample) appears essentially unbiased in comparison, as evidenced by the fact that the corresponding bars are barely visible in the plot. This observation holds for both the national population and the subgroup of poor rural women, for which each of the other approaches fairs much worse.

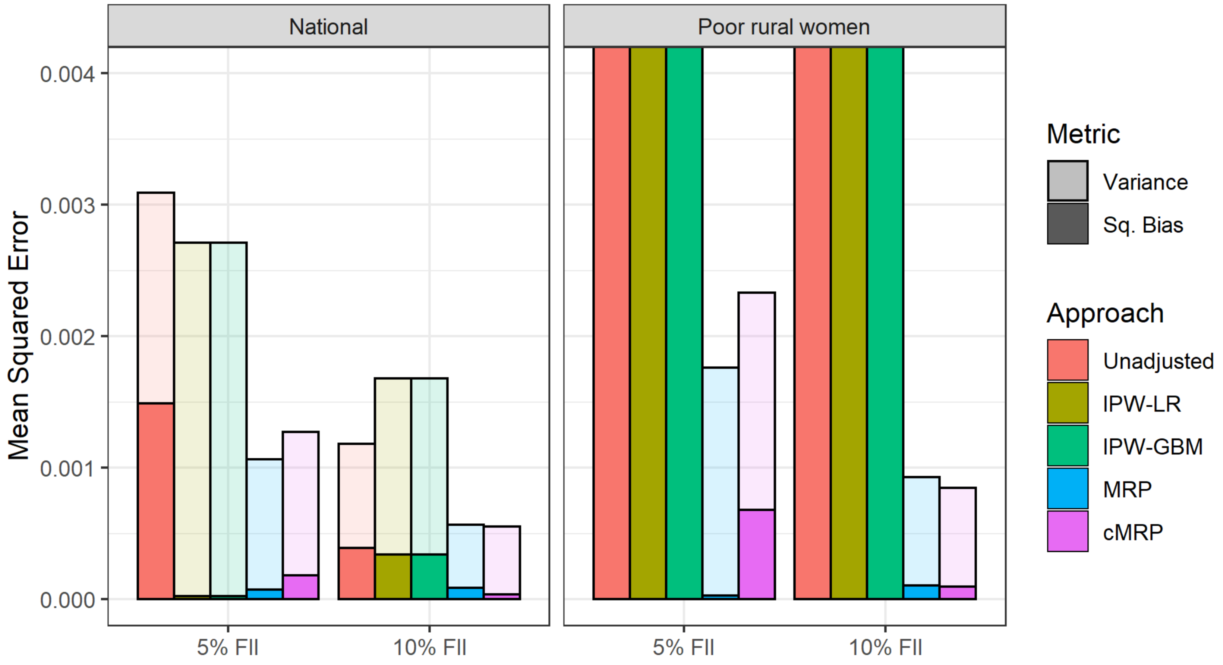
Figure 2. Comparison of the performance of each estimation approach that uses only SMS data to that of cMRP, averaging across the seven indicators. MSE is decomposed into the variance (light color on top) and squared bias (dark color below). IPW-LR uses the logistic regression propensity model, and IPW-GBM uses the gradient boosted propensity score model.



5.3.2. How does cMRP compare to approaches that only use small samples of the FII data? We compare in Figure 3 the performance of each estimation method that only uses FII data to that of cMRP. As in Figure 2, we display the performance as MSE and decompose the MSE into its two components (squared bias and variance). Note that we also truncate the MSE axis at 0.004 in order to highlight differences among the higher-performing approaches; the MSE for the unadjusted and IPW approaches among poor rural women all exceeded 0.04 (10 times the maximum value on the axis).

For the national target population, all five estimation methods perform reasonably well. Though the unadjusted estimates are biased (average bias is 2.8% for the 5% FII sample and 1.6% for the 10% FII sample), this bias can be nearly eliminated by using either IPW procedure. This is not the case among the subgroup of poor rural women because both the unadjusted and the IPW estimates are highly biased (bias is above 20% for each of the six approaches). These findings support the claims of Lau (2008), who found that weighting-based approaches tend to be biased for subgroup-specific estimates if the weights do not account for interactions between the subgroup-defining variables and other variables used for adjustment.

Figure 3. Comparison of the performance of each estimation approach that uses only FII data to that of cMRP, averaging across the seven indicators. Results are presented as MSE, which is decomposed into squared bias and variance. We truncate the vertical axis at 0.004 to highlight differences among the higher-performing approaches. IPW-LR uses the logistic regression propensity model, and IPW-GBM uses the gradient boosted propensity score model.



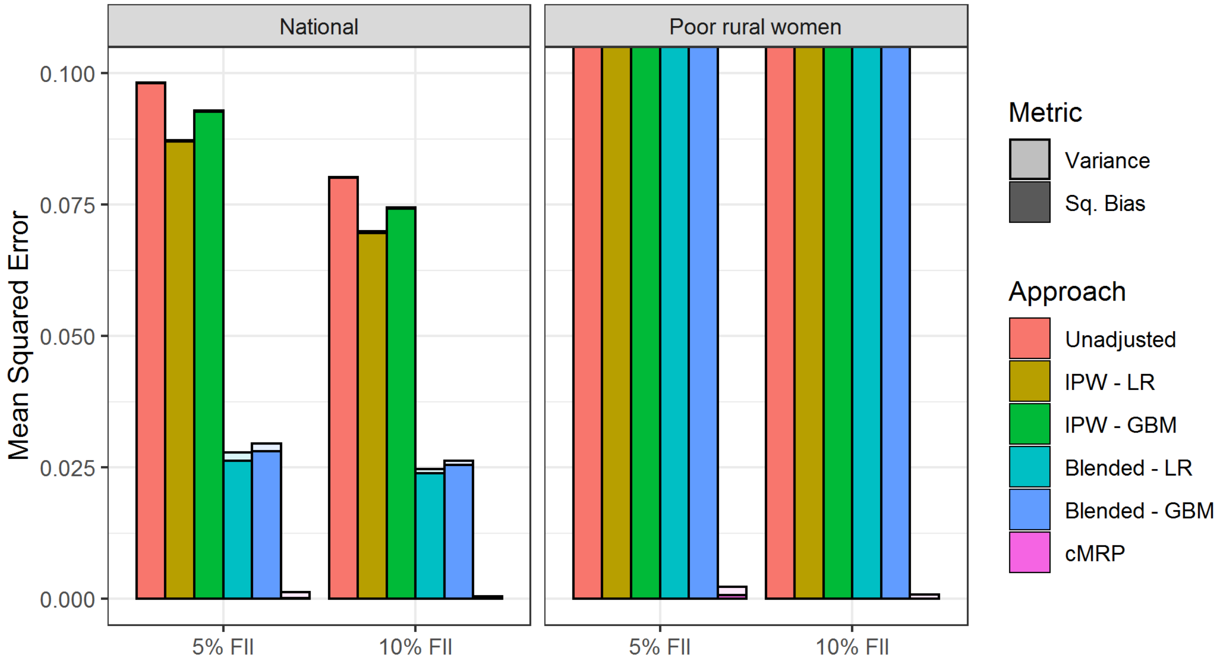
Both standard MRP and cMRP offer improvements over the IPW-based approaches, both for the national target population and for poor rural women. In the case of the subgroup-specific estimates, this improvement is extreme: whereas the IPW approaches were highly biased, the estimates based on standard MRP are essentially unbiased. For the national population, the use of MRP instead of IPW adds a small amount of bias, but it more than makes up for this increase with a large reduction in the variance of the resulting estimates, resulting in lower overall MSE. Interestingly, standard MRP performs slightly better than cMRP when a 5% sample of FII data is used, whereas cMRP offers a slight improvement over standard MRP when a 10% sample of FII data is used. However, these differences are very small and likely negligible in practice. These results do suggest that a viable alternative to collecting both SMS and a small sample of FII data may be to simply collect the small sample of FII data and perform standard MRP. We discuss this issue further in the Discussion section.

5.3.3. How does cMRP compare to other approaches that combine both SMS and FII data?

Finally, we compare cMRP to other approaches that combine both SMS and FII data (Figure 4). Among these methods, cMRP is far and away the top performing approach based on MSE—so much so that the pink bars corresponding to the MSE of cMRP are barely visible in the plot. This is especially true among poor rural women, where the MSE exceeds 0.14 for all blending approaches, 0.24 for all stacking approaches, and 0.26 for the unadjusted approaches (these numbers exceed the range of the plot). Moreover, most of the inferior performance of the non-cMRP approaches is predominantly due to bias: squared bias accounts for at least 94% of MSE in each of these 20 cases and above 99% in most of them (which is why the variance contribution is barely visible in the plot). For cMRP, on the other hand, squared bias accounts for between 7% and 30% of total MSE. We also note that the blending algorithm of

Elliott and Haviland (2007) outperforms the more basic stacking procedure for both target populations, FII sample sizes, and propensity score methods, but it still performs much worse than cMRP.

Figure 4. Comparison of the performance of each estimation approach that combines SMS and FII data, averaging across the seven indicators. Results are presented as MSE, which is decomposed into squared bias and variance. We truncate the vertical axis at 0.1 to highlight differences among the higher-performing approaches. IPW-LR uses the logistic regression propensity model, and IPW-GBM uses the gradient boosted propensity score model.



VI. Discussion

In this paper, we explored methods for the analysis of SMS survey data, with a focus on the application of MRP to obtain unbiased estimates of outcomes. We found that the standard application of MRP was insufficient to obtain unbiased estimates for outcomes related to digital financial inclusion because it could not correct for two potential sources of residual bias in the SMS data: residual sampling bias and survey mode bias. We then hypothesized that we can correct for these biases by collecting a relatively small sample of FTF data known to be free of survey mode bias. We also introduced calibrated MRP, a procedure that leverages the larger sample of inexpensive SMS data along with the unbiasedness of the FTF data in order to produce estimates either at the population level or for subgroups of interest.

Through both a simulation study and an applied experiment, we showed that cMRP produces efficient, unbiased estimates of target parameters across a wide range of scenarios. Whenever a substantial amount of bias was present in the simulations and in every scenario of the applied experiment, cMRP had lower bias and MSE than all methods that were based on SMS data alone, as well as all other methods that combined SMS and face-to-face data. These results held consistently across all scenarios, target populations, and outcomes that were explored.

We also found that calibrated MRP did not always outperform standard MRP that was applied to the small FII sample alone. This observation would indicate that for these cases, collecting the SMS sample is not worth the effort and a small FTF probability survey is sufficient for obtaining accurate estimates. However, we caution against attempting this approach because in practice it will be difficult to collect a small FTF sample that has sufficient coverage to obtain reliable results. In these analyses, adequate coverage was likely because our FTF samples were generated as simple random samples from the full population. In practice, collecting such a sample would be cost prohibitive because it would require interviewers to travel to each randomly selected region of the country, only collecting a small number of responses in each location. Fielding a survey in this manner would cause the cost per response to increase dramatically. We believe that combining SMS data with this small FTF sample (and performing estimation using calibrated MRP) can increase the coverage of the resulting sample at a much lower cost. Future work is needed to understand the performance of calibrated MRP using more realistic sampling designs for the FTF data, and explore the optimal design for the small FTF sample under these real-world constraints.

This study has several additional limitations. A limitation of both the simulation study and the applied experiment is that each was run for a relatively low number of iterations (100 for the simulation, 10 for the applied experiment). This was done due to the computational burden of fitting multiple Bayesian regression models under many different scenarios, but it could lead to spurious conclusions. Also in the applied experiment, the “target” we use to define the true indicator values in the population are not known but are estimated based on sampling weights that only account for age, gender, and urbanicity (InterMedia, 2016). Both the IPW and MRP procedures use a more exhaustive list of variables to adjust estimates to the target populations, and it is possible that some of these estimates could be closer to the true rate of the indicator in the country than the values we are using as the target.

Despite these limitations, this study illustrates the potential for obtaining accurate, cost-effective estimates of national indicators from SMS data using calibrated MRP, by combining the SMS sample with a second survey mode that is less susceptible to residual bias. More work should be done to further explore the conditions under which such an approach is effective, both in terms of estimation accuracy and in reducing data collection costs.

References

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073–1076.

Bürkner, P.C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics, 11(4), 427–444.

Demirguc-Kunt, A., Klapper, L. F., Singer, D., & Oedheusden, P. V. (2015). The Global Findex Database 2014: Measuring financial inclusion around the world. The World Bank.

Elliott, M. R., & Davis, W. W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3), 595-609.

Elliott, M. N., & Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. Survey Methodology, 33(2), 211–215.

Financial Inclusion Insights (2020). Data Fiinder. Retrieved from http://finclusion.org/data_fiinder/.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. Annals of statistics, 33(1), 1-53.

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge university press.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. Survey Methodology, 23, 127-135.

Ghitza, Y. and Gelman, A. (2013). “Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups.” American Journal of Political Science, 57(3), 762–776.

InterMedia (2016). Uganda Financial Inclusion Insights Survey, Technical Report, Wave Four 2016. Retrieved from http://finclusion.org/uploads/file/technical_report_uganda_fii_iv_final.pdf.

Jeoffreys-Leach, S., Grundling, I., Robertson, G. & Berkowitz, B. (2020). Innovative and Sustainable Data: Mixed Modes Data Collection and Modelling project. FinMark Trust insight2impact facility. November 2020. Retrieved from i2ifacility.org

Johnson, D. (2016). Collecting data from mHealth users via SMS surveys: A case study in Kenya. Survey Practice, 9(8).

Lau, A. (2018). Comparing MRP to raking for online opt-in polls. Decoded, Pew Research Center. November 13, 2018. Retrieved from https://medium.com/pew-research-center-decoded/comparing-mrp-to-raking-for-online-opt-in-polls-b05444d9931d.

Lau, C. Q., Lombaard, A., Baker, M., Eyerman, J., & Thalji, L. (2019). How representative are SMS surveys in Africa? Experimental evidence from four countries. International Journal of Public Opinion Research, 31(2), 309–330.

Levy, P. S., & Lemeshow, S. A. (2008). Sampling of populations: Methods and applications (4th ed.). Hoboken, NJ: Wiley.

Little, R. J. A. (1993). Post-stratification: A modeler’s perspective. Journal of the American Statistical Association, 88, 1001–1012.

Null, C., Chaplin, D., Hartog, J. Jacobson, J. Mamun, A., & Rangarajan, A. (2015). Findings from the Financial Inclusion Insights Surveys in East Africa: Inequities in use of mobile money. Report submitted to the Bill & Melinda Gates Foundation. Princeton, NJ: Mathematica Policy Research. Retrieved from https://cipre.mathematica-mpr.com/our-publications-and-findings/publications/findings-from-the-financial-inclusion-insights-surveys-in-east-africa-inequities-in-use-of-mobile

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. Political Analysis, 12(4), 375–385.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., & Feuer, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. Journal of the American Statistical Association, 102(478), 474-486.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2014). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the R TWANG Package. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/tools/TL136z1.html

Selb, P., & Munzert, S. (2011). Estimating constituency preferences from sparse survey data using auxiliary geographic information. Political Analysis, 19(4), 455–470.

Si, Y., Trangucci, R., Gabry, J.S., and Gelman, A. (2019). Bayesian hierarchical weighting adjustment and survey inference. arXiv preprint arXiv:1707.08220.

Trangucci, R., Ali, I., Gelman, A., & Rivers, D. (2018). Voting patterns in 2016: Exploration using multilevel regression and poststratification (MRP) on pre-election polls. arXiv preprint arXiv:1802.00842.

Valliant, R (2020). Comparing alternatives for estimation from nonprobability samples. Journal of Survey Statistics and Methodology, 8(2), 231-263.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). Finite population sampling and inference. New York, NY: John Wiley & Sons.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). “Forecasting elections with non-representative polls.” International Journal of Forecasting, 31(3), 980–991.

World Bank Group. (2017). Global financial inclusion and consumer protection survey, 2017 report. Washington, DC: World Bank.

Calibrated Multilevel Regression with Poststratification for the Analysis of SMS Survey Data
Supplementary Material

Appendix A. Model specification and sample code

In this section, we present the full specification for each multilevel regression model discussed in the paper. Sample code that uses the brms package in R is also provided. We follow the same notation that is used in the paper, which we repeat here for convenience. Let $k \in \{1, \dots, K\}$ index the poststratification variables, and $l \in \{1, \dots, L_k\}$ index the level of the k^{th} poststratification variable. Also let $y_i \in \{0,1\}$ be the binary outcome for respondent i, and $x_{ikl} \in \{0,1\}$ be the indicator that poststratification variable k takes the value l for respondent i. Also let $p_i = \Pr(Y_i = 1)$.

Model A1. Standard MRP model without interactions

This specification is repeated from the text (Equation 3) for the reader’s convenience:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{k=1}^{K} \sum_{l=1}^{L_k} \alpha_l^k x_{ikl}$$

$$\begin{aligned} \alpha &\sim N(0,1) \\ \alpha_l^k &\sim N(0, \sigma_k) \\ \sigma_k &\sim \frac{1}{2} N(0, \tau) \\ \tau &\sim \frac{1}{2} N(0,1) \end{aligned}$$

Model A2. Standard MRP model with interactions

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{k=1}^{K} \sum_{l=1}^{L_k} \alpha_l^k x_{ikl} + \sum_{j,k:j<k} \sum_{l=1}^{L_k} \sum_{m=1}^{L_j} \phi_{lm}^{jk} x_{ikl} x_{ijm}$$

In addition to those listed for Model 1, we have the following priors (note the superscript “I” indicates “Interaction”):

$$\begin{aligned} \phi_{lm}^{jk} &\sim N\left(0, \sigma_{jk}^I\right) \\ \sigma_{jk}^I &\sim \frac{1}{2} N(0, \tau^I) \end{aligned}$$

Model A3. Calibrated MRP without interactions between poststratification variables

Let $M_i \in \{0,1\}$ be the indicator that respondent i was surveyed through the SMS survey.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta M_i + \sum_{k=1}^{K}\sum_{l=1}^{L_k}(\alpha_l^k x_{ikl} + \gamma_l^k x_{ikl} M_i)$$

In addition to those listed for Model 1, we have the following priors (note the superscript "M" indicates "Mode"):

$$\beta \sim N(0,1)$$
$$\gamma_l^k \sim N\left(0, \sigma_k^M\right)$$
$$\sigma_k^M \sim \frac{1}{2} N(0, \tau^I)$$

Model A4. Calibrated MRP with interactions between poststratification variables

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta M_i + \sum_{k=1}^{K}\sum_{l=1}^{L_k}(\alpha_l^k x_{ikl} + \gamma_l^k x_{ikl} M_i) + \sum_{j,k:j<k}\sum_{l=1}^{L_k}\sum_{m=1}^{L_j} \phi_{lm}^{jk} x_{ikl} x_{ijm}$$

This model includes all priors that are appear above in Models 1, 2, or 3. Note that for simplicity, the same hyperprior $\frac{1}{2} N(0, \tau^I)$ applies to the variance components associated with all interaction terms, regardless of whether the term includes the survey mode indicator.

Appendix B. Sample code

This section presents the R code that we used to fit the multilevel models and perform the poststratification. The code uses the `brms` package (Bürkner, 2018) to fit the model. `brms` is a wrapper for the `rstan` package (Stan Development Team, 2020), which uses Stan on the backend to fit the Bayesian model.

The code is parameterized so that it can fit all four of the models described in Appendix A. The main function for fitting the multilevel regression model is `fit_model()`, which takes four arguments: the data as a data.frame (`dat`), a vector of poststratification variables (`xvars`), whether or not interactions should be used (`use_interactions`), and whether to fit cMRP or standard MRP (`cMRP`). Two accessory functions follow the declaration of `fit_model()`. Required packages include `tidyverse` and `brms`, as well as their dependencies.

```
fit_model <- function(dat, xvars, use_interactions, cMRP) {

  interactions <- if (use_interactions){
    combn(xvars,2) %>%
      t() %>%
      data.frame(stringsAsFactors=F) %>%
      filter(X1!=X2) %>%
      mutate(interaction = paste0(X1,":",X2)) %>%
      pull(interaction)
  } else NULL

  if (cMRP) {
    interactions <- c(interactions,
                      paste(xvars, "Mode", sep = ":"))
    xvars <- c(xvars, "Mode")
  }

  fmla <- build_fmla("Y", xvars, interactions)
  prior <- build_priors_hier(xvars, interactions)

  brm(fmla, prior = prior$prior, stanvar = prior$stanvar,
      family = "bernoulli", data = dat, iter = 1000, chains = 4)
}

build_fmla <- function(outcome, ps_vars, interactions=NULL){
  paste(outcome, "~", paste0("(1|", c(ps_vars, interactions) ,")",
                             collapse=" + ")) %>%
    as.formula()
}


build_priors_hier <- function(ps_vars, interactions){
  stanvar <- stanvar(scode = c("real<lower=0> tau_sigma;"),
                     block = "parameters")

  prior <-
    prior_string("normal(0,tau_sigma)",class="sd") +
    prior(normal(0,1), class = "Intercept") +
    prior_string("target += normal_lpdf(tau_sigma | 0, 1) - 1 *
normal_lccdf(0 |0, 1)", check=F)

  if (!is.null(interactions)){
    prior_interactions <-
      map_df(interactions,
     ~prior_string("normal(0,tau_sigma2)",class="sd",group=.)) +
      prior_string("target += normal_lpdf(tau_sigma2 | 0, 1) - 1 *
normal_lccdf(0 |0, 1)", check=F)
    prior <- prior + prior_interactions
    stanvar = stanvar +
      stanvar(scode = c("real<lower=0> tau_sigma2;"),
              block = "parameters")
  }

  list(prior=prior, stanvar=stanvar)
}
```

The following function, `poststratify_mrp()`, performs the poststratification. It requires four arguments: the fitted brms model (`fit`), a data.frame that has one row for every combination of poststratification variables and a `Weight` column containing the poststratification weight (`psw`), whether cMRP is being performed (`cMRP`), and whether the target population is the subgroup or the full population (subgroup). It returns the posterior for the poststratified estimate in the target population, as a vector.

```
poststratify_mrp <- function(fit, psw, cMRP, subgroup) {

  # Predict poststratification cells (returns the full posterior)
  if (cMRP) {
    # Predict assuming Mode is F2F
    psw$Mode <- "F2F"
  }
  pred <- posterior_linpred(fit, newdata = psw, transform = TRUE,
                            allow_new_levels = TRUE)
  psw$pred <- t(pred)

  if (subgroup) {
    # Filter to lower educated rural women and rescale weights
    psw <- filter(psw,
                  Education == "Primary education incomplete",
                  Rural == "rural",
                  Gender == "Woman") %>%
      mutate(Weight = Weight/sum(Weight))

  }

  # Poststratify
  t(psw$pred) %*% psw$Weight %>%
    as.numeric()
}
```

Appendix C. Additional simulation results

Tables 1-2 of the paper show simulation results for the scenarios where $N^{FTF} = 150$ and $N^{ref} = \infty$. Here, we present the remaining 10 tables, which correspond to five other combinations of N^{F2F} and N^{ref}, and for each target population.

Table C2. Simulation results for the scenarios with $N^{FTF} = 150$, $N^{ref} = 3000$, and the target population is the full population.

		No survey mode bias $(\gamma = 0)$			With survey mode bias $(\gamma = 0.5)$		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	1.6 (0.4)	291.3 (17.0)	1140.2 (33.7)	92.3 (9.5)	647.0 (25.4)	1710.8 (41.3)
SMS	IPW	1.5 (0.1)	39.9 (6.2)	206.4 (14.2)	87.6 (9.3)	232.3 (15.2)	536.9 (23.0)
SMS	MRP1	1.6 (0.2)	1.9 (0.4)	12.7 (2.6)	83.7 (9.1)	93.7 (9.6)	140.3 (11.4)
SMS	MRP2	1.5 (0.2)	1.9 (0.3)	12.6 (2.5)	84.1 (9.1)	93.0 (9.5)	138.9 (11.3)
FTF	Unadj.	16.0 (0.9)	14.3 (0.4)	14.3 (0.1)	16.2 (0.4)	12.9 (0.5)	13.8 (0.2)
FTF	IPW	15.1 (0.6)	12.7 (0.0)	12.6 (0.4)	11.1 (0.3)	15.3 (0.2)	12.4 (0.3)
FTF	MRP1	13.9 (0.7)	10.8 (0.1)	11.1 (0.4)	11.8 (0.2)	14.0 (0.4)	10.8 (0.1)
FTF	MRP2	13.5 (0.6)	10.7 (0.0)	10.9 (0.4)	11.8 (0.3)	13.8 (0.4)	11.2 (0.1)
CMB	Unadj.	1.6 (0.4)	242.1 (15.5)	941.6 (30.7)	77.2 (8.7)	537.0 (23.2)	1415.2 (37.6)
CMB	IPW	1.4 (0.1)	33.1 (5.6)	169.9 (12.9)	73.0 (8.5)	192.7 (13.8)	442.3 (20.9)
CMB	Blended	6.8 (0.4)	14.6 (1.5)	14.3 (0.7)	16.2 (1.9)	19.6 (1.3)	13.6 (0.4)
CMB	cMRP1	7.9 (0.3)	5.4 (0.1)	8.0 (0.2)	12.6 (1.2)	14.1 (1.4)	11.1 (0.8)
CMB	cMRP2	9.0 (0.3)	6.6 (0.1)	8.5 (0.1)	11.4 (1.1)	13.2 (1.2)	10.6 (0.7)

Table C2. Simulation results for the scenarios with $N^{FTF} = 150$, $N^{ref} = 3000$, and the target population is the subgroup of lower-educated, rural women.

		No survey mode bias $(\gamma = 0)$			With survey mode bias $(\gamma = 0.5)$		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	3.2 (0.1)	6.2 (0.5)	103.5 (6.8)	64.8 (7.8)	71.9 (8.0)	290.6 (14.8)
SMS	IPW	3.0 (0.1)	5.7 (0.1)	78.9 (5.2)	58.8 (7.5)	59.4 (7.2)	241.9 (13.1)
SMS	MRP1	2.8 (0.8)	4.8 (1.2)	37.1 (5.0)	74.5 (8.5)	78.1 (8.6)	174.9 (12.4)
SMS	MRP2	2.5 (0.6)	4.2 (0.9)	36.6 (4.9)	68.5 (8.1)	71.8 (8.2)	168.3 (12.1)
FTF	Unadj.	37.0 (0.6)	31.2 (0.6)	25.7 (0.0)	39.6 (0.7)	37.0 (0.1)	28.2 (0.2)
FTF	IPW	34.1 (0.0)	30.7 (1.0)	24.8 (0.3)	40.5 (0.3)	39.5 (0.4)	28.4 (0.2)
FTF	MRP1	41.0 (4.2)	36.8 (3.6)	27.1 (3.3)	37.3 (3.7)	33.5 (3.1)	31.3 (3.1)
FTF	MRP2	35.4 (3.3)	29.9 (2.5)	23.3 (2.5)	34.2 (3.1)	30.7 (2.3)	27.9 (2.5)
CMB	Unadj.	2.9 (0.2)	5.5 (0.3)	26.9 (2.8)	54.3 (7.2)	52.6 (6.8)	64.2 (6.5)
CMB	IPW	2.8 (0.1)	5.1 (0.2)	54.7 (4.2)	49.5 (6.8)	47.6 (6.4)	166.8 (10.8)
CMB	Blended	17.7 (0.3)	17.7 (1.0)	21.1 (0.2)	36.6 (1.6)	35.4 (1.0)	28.2 (1.1)
CMB	cMRP1	7.5 (1.3)	6.1 (1.0)	12.0 (2.3)	11.8 (2.0)	12.4 (1.7)	16.3 (2.3)
CMB	cMRP2	8.9 (1.3)	7.4 (0.9)	11.2 (2.0)	12.0 (1.8)	11.9 (1.5)	14.8 (2.0)

Table C3. Simulation results for the scenarios with $N^{FTF} = 150$, $N^{ref} = 10000$, and the target population is the full population.

		No survey mode bias $(\gamma = 0)$			With survey mode bias $(\gamma = 0.5)$		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	1.5 (0.5)	301.2 (17.3)	1133.2 (33.6)	92.3 (9.5)	645.0 (25.4)	1710.8 (41.3)
SMS	IPW	1.4 (0.2)	12.6 (3.3)	91.6 (9.2)	85.9 (9.2)	155.2 (12.4)	372.9 (19.1)
SMS	MRP1	1.3 (0.0)	1.9 (0.5)	11.9 (2.4)	83.6 (9.1)	92.9 (9.5)	145.5 (11.7)
SMS	MRP2	1.3 (0.1)	1.8 (0.4)	11.8 (2.3)	84.0 (9.1)	92.1 (9.5)	144.5 (11.7)
FTF	Unadj.	12.9 (0.2)	14.8 (1.1)	14.1 (0.3)	18.8 (0.7)	13.8 (0.7)	16.4 (0.4)
FTF	IPW	14.3 (0.4)	11.5 (0.3)	13.8 (0.1)	12.2 (0.4)	11.0 (0.4)	14.6 (0.3)
FTF	MRP1	12.6 (0.3)	10.9 (0.5)	12.2 (0.0)	13.3 (0.5)	9.5 (0.5)	12.8 (0.4)
FTF	MRP2	12.9 (0.2)	10.6 (0.6)	12.1 (0.0)	13.4 (0.5)	9.2 (0.5)	13.5 (0.3)
CMB	Unadj.	1.4 (0.4)	252.2 (15.8)	938.2 (30.6)	77.5 (8.7)	536.2 (23.1)	1416.6 (37.6)
CMB	IPW	1.2 (0.2)	10.8 (3.0)	75.8 (8.4)	71.7 (8.4)	129.3 (11.3)	309.5 (17.4)
CMB	Blended	6.7 (0.3)	8.9 (1.4)	16.3 (1.5)	16.0 (1.9)	16.1 (1.8)	17.7 (1.3)
CMB	cMRP1	6.7 (0.0)	5.8 (0.6)	7.7 (0.3)	14.5 (1.6)	11.2 (1.6)	13.7 (1.5)
CMB	cMRP2	8.2 (0.1)	6.7 (0.6)	8.3 (0.2)	13.2 (1.4)	10.2 (1.4)	12.8 (1.3)

Table C4. Simulation results for the scenarios with $N^{FTF} = 150$, $N^{ref} = 10000$, and the target population is the subgroup of lower-educated, rural women.

		No survey mode bias $(\gamma = 0)$			With survey mode bias $(\gamma = 0.5)$		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	3.2 (0.2)	6.4 (0.6)	90.1 (5.3)	66.9 (7.9)	80.3 (8.6)	223.9 (12.6)
SMS	IPW	3.1 (0.1)	5.8 (0.1)	81.3 (4.4)	59.0 (7.4)	65.3 (7.6)	194.9 (11.5)
SMS	MRP1	2.4 (0.9)	4.0 (1.2)	32.5 (4.4)	74.2 (8.4)	85.2 (9.0)	177.6 (12.5)
SMS	MRP2	2.3 (0.6)	3.7 (0.9)	32.9 (4.2)	69.1 (8.1)	78.7 (8.6)	168.5 (12.1)
FTF	Unadj.	30.1 (0.2)	29.6 (0.9)	37.0 (0.1)	27.9 (0.2)	27.0 (0.8)	32.8 (0.2)
FTF	IPW	29.4 (0.6)	29.6 (0.5)	37.8 (0.1)	28.8 (0.1)	26.3 (0.3)	37.2 (0.4)
FTF	MRP1	38.2 (4.2)	38.8 (4.2)	39.5 (3.7)	34.6 (3.9)	34.0 (4.1)	30.9 (3.5)
FTF	MRP2	30.7 (3.1)	33.3 (3.6)	36.1 (2.9)	29.8 (3.0)	28.5 (3.4)	28.5 (2.6)
CMB	Unadj.	2.6 (0.2)	5.5 (0.7)	27.8 (2.2)	55.8 (7.2)	60.0 (7.3)	54.1 (5.2)
CMB	IPW	2.5 (0.1)	5.2 (0.0)	60.6 (3.7)	49.0 (6.8)	53.7 (6.9)	143.4 (9.8)
CMB	Blended	14.6 (0.8)	16.0 (0.0)	31.5 (0.1)	27.2 (1.4)	26.4 (1.8)	37.0 (0.6)
CMB	cMRP1	6.0 (0.9)	5.6 (1.4)	13.5 (2.0)	13.6 (2.2)	12.3 (2.3)	19.3 (2.6)
CMB	cMRP2	7.5 (1.1)	8.3 (1.5)	14.5 (1.8)	12.9 (2.1)	11.6 (2.0)	17.8 (2.3)

Table C5. Simulation results for the scenarios with $N^{FTF} = 300$, $N^{ref} = 3000$, and the target population is the full population.

		No survey mode bias ($\gamma = 0$)			With survey mode bias ($\gamma = 0.5$)		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	1.4 (0.3)	303.7 (17.4)	1145.3 (33.8)	95.9 (9.7)	662.4 (25.7)	1713.4 (41.4)
SMS	IPW	1.3 (0.2)	44.7 (6.5)	212.6 (14.4)	92.5 (9.5)	242.2 (15.5)	547.5 (23.3)
SMS	MRP1	1.2 (0.0)	2.6 (0.7)	13.7 (2.5)	88.7 (9.3)	98.3 (9.8)	143.2 (11.6)
SMS	MRP2	1.2 (0.0)	2.6 (0.7)	13.8 (2.5)	88.8 (9.4)	97.5 (9.8)	141.4 (11.5)
FTF	Unadj.	6.5 (0.4)	7.2 (0.0)	7.5 (0.2)	11.5 (0.6)	6.6 (0.7)	8.4 (0.8)
FTF	IPW	6.6 (0.3)	6.7 (0.1)	6.7 (0.1)	10.0 (0.5)	5.7 (0.3)	8.3 (0.8)
FTF	MRP1	6.2 (0.2)	6.5 (0.2)	6.2 (0.1)	9.4 (0.4)	5.4 (0.2)	7.9 (0.8)
FTF	MRP2	6.1 (0.2)	6.6 (0.2)	6.3 (0.1)	9.4 (0.4)	5.4 (0.2)	8.0 (0.8)
CMB	Unadj.	1.1 (0.3)	211.2 (14.5)	797.2 (28.2)	68.7 (8.2)	464.9 (21.5)	1199.0 (34.6)
CMB	IPW	1.1 (0.2)	31.0 (5.4)	148.1 (12.0)	65.9 (8.0)	169.6 (13.0)	385.6 (19.5)
CMB	Blended	3.4 (0.2)	8.1 (0.9)	7.9 (0.7)	13.1 (1.4)	6.8 (0.8)	9.3 (1.2)
CMB	cMRP1	3.9 (0.1)	4.3 (0.0)	4.6 (0.2)	10.3 (1.1)	6.1 (0.7)	8.3 (1.3)
CMB	cMRP2	4.5 (0.1)	4.7 (0.1)	4.7 (0.2)	9.7 (1.0)	5.8 (0.7)	8.2 (1.3)

Table C6. Simulation results for the scenarios with $N^{FTF} = 300$, $N^{ref} = 3000$, and the target population is the subgroup of lower-educated, rural women.

		No survey mode bias ($\gamma = 0$)			With survey mode bias ($\gamma = 0.5$)		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	2.8 (0.1)	3.6 (0.7)	57.6 (3.9)	66.0 (7.9)	83.0 (8.7)	263.6 (13.9)
SMS	IPW	2.8 (0.2)	2.8 (0.1)	44.0 (2.1)	60.1 (7.5)	67.8 (7.8)	209.1 (11.8)
SMS	MRP1	2.4 (0.7)	4.4 (1.4)	33.3 (4.1)	74.6 (8.5)	84.3 (8.9)	178.2 (12.5)
SMS	MRP2	2.1 (0.4)	3.3 (1.0)	31.7 (3.9)	69.2 (8.1)	78.4 (8.6)	171.8 (12.2)
FTF	Unadj.	13.7 (0.9)	18.5 (0.7)	12.6 (0.7)	19.0 (0.3)	13.9 (0.3)	19.0 (0.3)
FTF	IPW	12.6 (0.4)	16.0 (0.1)	11.7 (0.3)	18.1 (0.1)	12.6 (0.2)	18.1 (0.1)
FTF	MRP1	14.4 (2.5)	16.7 (2.2)	12.6 (2.1)	19.4 (2.4)	14.3 (2.3)	16.5 (2.3)
FTF	MRP2	12.9 (2.1)	14.5 (1.7)	11.1 (1.6)	16.8 (1.8)	12.2 (1.8)	15.9 (1.8)
CMB	Unadj.	2.2 (0.2)	3.5 (0.7)	12.1 (1.6)	47.2 (6.6)	45.5 (6.4)	31.8 (3.9)
CMB	IPW	2.2 (0.1)	2.5 (0.1)	22.4 (1.6)	42.8 (6.3)	43.5 (6.3)	105.5 (8.3)
CMB	Blended	6.6 (0.1)	8.6 (0.1)	9.6 (0.4)	18.9 (1.0)	13.7 (0.9)	18.1 (0.6)
CMB	cMRP1	3.7 (0.8)	5.0 (0.9)	7.1 (1.4)	10.2 (1.6)	5.7 (1.3)	12.0 (2.3)
CMB	cMRP2	4.4 (0.9)	5.6 (0.8)	6.7 (1.1)	9.8 (1.4)	5.8 (1.2)	11.0 (1.9)

Table C7. Simulation results for the scenarios with $N^{FTF} = 300$, $N^{ref} = 10000$, and the target population is the full population.

		No survey mode bias ($\gamma = 0$)			With survey mode bias ($\gamma = 0.5$)		
Source	Method	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	2.0 (0.6)	299.6 (17.3)	1134.5 (33.7)	94.3 (9.6)	649.0 (25.5)	1719.3 (41.5)
SMS	IPW	1.6 (0.4)	13.7 (3.4)	94.8 (9.4)	89.7 (9.4)	155.3 (12.4)	362.0 (18.8)
SMS	MRP1	1.5 (0.3)	2.2 (0.5)	10.8 (2.4)	87.9 (9.3)	93.4 (9.5)	140.4 (11.5)
SMS	MRP2	1.5 (0.3)	2.1 (0.5)	10.4 (2.3)	88.0 (9.3)	92.9 (9.5)	138.5 (11.4)
FTF	Unadj.	7.7 (0.4)	9.6 (0.1)	6.3 (0.1)	10.2 (0.2)	9.5 (0.5)	7.9 (0.0)
FTF	IPW	8.3 (0.4)	8.4 (0.1)	7.4 (0.0)	8.9 (0.4)	7.2 (0.1)	6.3 (0.2)
FTF	MRP1	8.1 (0.4)	8.0 (0.1)	6.8 (0.1)	8.7 (0.2)	7.0 (0.1)	6.0 (0.2)
FTF	MRP2	7.9 (0.4)	7.9 (0.0)	6.8 (0.1)	8.6 (0.2)	6.9 (0.1)	5.9 (0.2)
CMB	Unadj.	1.8 (0.6)	208.7 (14.4)	788.2 (28.1)	66.4 (8.1)	454.7 (21.3)	1194.6 (34.6)
CMB	IPW	1.5 (0.4)	9.5 (2.8)	66.3 (7.8)	61.6 (7.8)	108.5 (10.3)	250.8 (15.7)
CMB	Blended	4.6 (0.4)	7.0 (0.6)	8.8 (0.8)	10.3 (0.5)	8.8 (0.8)	6.6 (0.2)
CMB	cMRP1	5.1 (0.3)	5.4 (0.0)	4.8 (0.2)	9.1 (0.3)	7.5 (0.8)	6.7 (0.5)
CMB	cMRP2	5.8 (0.4)	6.1 (0.0)	5.2 (0.1)	8.9 (0.2)	7.0 (0.7)	6.4 (0.4)

Table C8. Simulation results for the scenarios with $N^{FTF} = 300$, $N^{ref} = 10000$, and the target population is the subgroup of lower-educated, rural women.

		No survey mode bias ($\gamma = 0$)			With survey mode bias ($\gamma = 0.5$)		
Source	Method	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	2.4 (0.4)	6.4 (0.5)	84.2 (4.5)	64.3 (7.9)	70.9 (7.9)	250.4 (14.1)
SMS	IPW	2.1 (0.0)	5.8 (0.1)	75.8 (3.4)	55.8 (7.3)	56.8 (7.0)	217.9 (13.0)
SMS	MRP1	2.6 (1.0)	5.7 (1.4)	32.7 (4.1)	71.9 (8.4)	79.6 (8.6)	169.0 (12.2)
SMS	MRP2	2.1 (0.7)	4.8 (1.0)	31.9 (3.8)	68.0 (8.1)	73.1 (8.2)	162.6 (11.9)
FTF	Unadj.	18.8 (0.5)	18.2 (0.5)	17.0 (0.0)	16.1 (0.3)	20.9 (0.7)	15.7 (0.4)
FTF	IPW	19.4 (0.2)	17.4 (0.1)	17.5 (0.4)	16.4 (0.9)	21.1 (0.4)	16.1 (1.0)
FTF	MRP1	17.4 (2.4)	15.4 (2.3)	15.3 (2.0)	14.7 (2.0)	17.7 (2.5)	13.6 (1.7)
FTF	MRP2	15.7 (1.8)	14.2 (1.7)	13.4 (1.5)	13.0 (1.3)	16.2 (2.0)	12.3 (1.2)
CMB	Unadj.	2.2 (0.4)	4.7 (0.5)	14.2 (1.3)	44.6 (6.5)	41.8 (6.0)	25.1 (3.5)
CMB	IPW	1.9 (0.0)	4.4 (0.1)	42.5 (2.5)	37.7 (5.9)	39.7 (5.8)	119.2 (9.5)
CMB	Blended	11.5 (0.0)	11.1 (0.0)	16.2 (0.3)	17.4 (0.3)	21.5 (1.4)	16.4 (0.3)
CMB	cMRP1	4.5 (1.0)	5.9 (1.0)	7.2 (1.4)	6.6 (0.9)	7.8 (1.3)	8.1 (1.4)
CMB	cMRP2	5.5 (0.9)	6.2 (0.9)	7.4 (1.2)	6.6 (0.8)	8.1 (1.2)	7.6 (1.1)

Table C9. Simulation results for the scenarios with $N^{FTF} = 300$, $N^{ref} = \infty$, and the target population is the full population.

		No survey mode bias $(\gamma = 0)$			With survey mode bias $(\gamma = 0.5)$		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	2.0 (0.6)	299.6 (17.3)	1134.5 (33.7)	94.3 (9.6)	649.0 (25.5)	1719.3 (41.5)
SMS	IPW	1.6 (0.4)	13.7 (3.4)	94.8 (9.4)	89.7 (9.4)	155.3 (12.4)	362.0 (18.8)
SMS	MRP1	1.5 (0.3)	2.2 (0.5)	10.9 (2.4)	87.5 (9.3)	93.4 (9.5)	140.0 (11.5)
SMS	MRP2	1.5 (0.4)	2.1 (0.5)	10.5 (2.3)	87.7 (9.3)	93.1 (9.5)	138.5 (11.4)
FTF	Unadj.	7.7 (0.4)	9.6 (0.1)	6.3 (0.1)	10.2 (0.2)	9.5 (0.5)	7.9 (0.0)
FTF	IPW	8.3 (0.4)	8.4 (0.1)	7.4 (0.0)	8.9 (0.4)	7.2 (0.1)	6.3 (0.2)
FTF	MRP1	8.2 (0.4)	7.9 (0.1)	6.7 (0.1)	8.6 (0.2)	7.1 (0.1)	5.9 (0.2)
FTF	MRP2	8.0 (0.4)	7.9 (0.0)	6.9 (0.1)	8.5 (0.2)	6.9 (0.1)	5.8 (0.2)
CMB	Unadj.	1.8 (0.6)	208.7 (14.4)	788.2 (28.1)	66.4 (8.1)	454.7 (21.3)	1194.6 (34.6)
CMB	IPW	1.5 (0.4)	9.5 (2.8)	66.3 (7.8)	61.6 (7.8)	108.5 (10.3)	250.8 (15.7)
CMB	Blended	4.6 (0.4)	7.0 (0.6)	8.8 (0.8)	10.3 (0.5)	8.8 (0.8)	6.6 (0.2)
CMB	cMRP1	5.2 (0.3)	5.4 (0.0)	4.9 (0.2)	8.9 (0.2)	7.6 (0.8)	6.6 (0.5)
CMB	cMRP2	5.9 (0.4)	6.0 (0.0)	5.2 (0.1)	8.7 (0.2)	7.1 (0.7)	6.2 (0.4)

Table C10. Simulation results for the scenarios with $N^{FTF} = 300$, $N^{ref} = \infty$, and the target population is the subgroup of lower-educated, rural women.

		No survey mode bias $(\gamma = 0)$			With survey mode bias $(\gamma = 0.5)$		
Source	**Method**	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
SMS	Unadj.	2.4 (0.4)	6.4 (0.5)	84.2 (4.5)	64.3 (7.9)	70.9 (7.9)	250.4 (14.1)
SMS	IPW	2.1 (0.0)	5.8 (0.1)	75.8 (3.4)	55.8 (7.3)	56.8 (7.0)	217.9 (13.0)
SMS	MRP1	2.5 (0.9)	5.6 (1.4)	32.8 (4.1)	72.3 (8.4)	79.5 (8.6)	166.5 (12.1)
SMS	MRP2	2.1 (0.6)	4.7 (1.0)	32.1 (3.9)	68.4 (8.2)	73.0 (8.2)	161.2 (11.9)
FTF	Unadj.	18.8 (0.5)	18.2 (0.5)	17.0 (0.0)	16.1 (0.3)	20.9 (0.7)	15.7 (0.4)
FTF	IPW	19.4 (0.2)	17.4 (0.1)	17.5 (0.4)	16.4 (0.9)	21.1 (0.4)	16.1 (1.0)
FTF	MRP1	17.0 (2.4)	15.0 (2.3)	15.4 (2.0)	14.5 (2.0)	18.0 (2.5)	13.6 (1.7)
FTF	MRP2	15.2 (1.8)	14.2 (1.7)	13.3 (1.5)	12.9 (1.3)	16.5 (2.0)	12.3 (1.1)
CMB	Unadj.	2.2 (0.4)	4.7 (0.5)	14.2 (1.3)	44.6 (6.5)	41.8 (6.0)	25.1 (3.5)
CMB	IPW	1.9 (0.0)	4.4 (0.1)	42.5 (2.5)	37.7 (5.9)	39.7 (5.8)	119.2 (9.5)
CMB	Blended	11.5 (0.0)	11.1 (0.0)	16.2 (0.3)	17.4 (0.3)	21.5 (1.4)	16.4 (0.3)
CMB	cMRP1	4.4 (1.0)	5.7 (1.0)	7.2 (1.4)	6.5 (0.9)	7.9 (1.3)	8.1 (1.4)
CMB	cMRP2	5.3 (0.9)	6.0 (0.9)	7.4 (1.2)	6.6 (0.8)	8.2 (1.2)	7.5 (1.1)

References

Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.19.3. http://mc-stan.org/.

About the Series

Policymakers and researchers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers access to our most current work.

For more information about this paper, contact Jonathan Gellar, Senior Statistician, at JGellar@mathematica-mpr.com.

Suggested citation: Gellar, Jonathan E., Constance Delannoy, Erin Lipman, Shirley Jeoffreys-Leach, Bobby Berkowitz, Grant J. Robertson, and Sarah M. Hughes. "Calibrated Multilevel Regression with Poststratification for the Analysis of SMS Survey Data." Working Paper 66. Washington, DC: Mathematica Policy Research, June 26, 2021.

Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

Mathematica
Progress Together

mathematica.org