Reducing Measurement and Sampling Biases in Nonprobability Surveys

Aditi Sen
asen123@umd.edu

Department of Mathematics
University of Maryland College Park

IASS Webinar

February 27, 2026

Joint work with

Prof. Partha Lahiri
Joint Program in Survey Methodology,
Department of Mathematics,
University of Maryland College Park
IASS President

Outline of talk

Outline of talk

1. **Motivation**

A discussion on nonprobability surveys

Nonprobability surveys: Introduction

- Participants selected based on non-random criteria, such as availability, convenience, or researcher judgment rather than random selection.
- Biased samples that do not represent the target population.
- Sample inclusion or participation mechanisms are unknown.
- Selection probabilities could be zero for some population units.
- Model-free estimation (design-based inference) is not possible.
- The estimates can be assessed only by subjective judgment about model assumptions.
- As in probability sampling, both sampling and nonsampling errors exist.

Nonprobability surveys: Introduction

- Nonprobability existed since early days of conducting surveys, but have recently gained increased popularity.
- Useful when probability survey is difficult to implement due to cost, time or lack of sampling frame.
- Applications are in market research, social science, public health, etc.
- Examples: Quota sampling, Snowball sampling, Respondent driven sampling, Venue-based sampling, Internet sampling.
- We focus on **opt-in panel**, recruited through invitations attached as banners to specific websites.
- Internet users voluntarily sign up to respond to surveys for incentives.
- Inexpensive to conduct and fast responses, but fail to cover the entire population.

Nonprobability surveys: Recent works

- Special issues of *Survey Methodology* and *Calcutta Statistical Association Bulletin* on statistical research in nonprobability sampling.

- Recent literature focus on integrating nonprobability with probability (reference) surveys (usually not having response).

- Some well-known methods include:
  - Matching and mass imputation - Rivers (2007), Kim, Park, Chen, and Wu (2021)
  - Inverse propensity score weighting (IPW) - Lee and Valliant (2009), Valliant and Dever (2011), Wang, Valliant, and Li (2021)
  - Doubly robust methods - Chen, Li, and Wu (2020)

- Additional references - Rao (2021); Wu (2022); Beaumont (2020); Beaumont, Bosa, Brennan, Charlebois, and Chu (2024), etc.

- Huge body of work on improving representativeness, **but not much focus on measurement error in nonprobability surveys**.

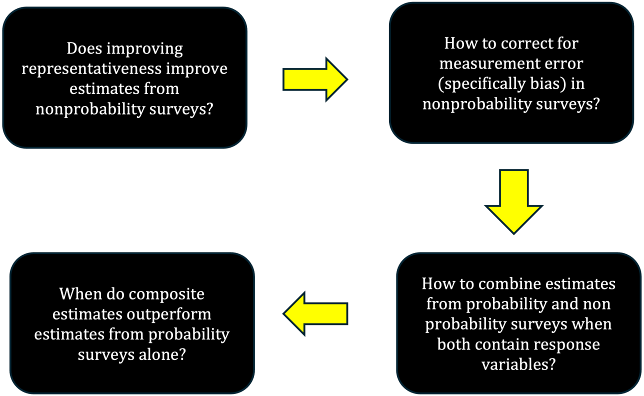
Outline of talk

2 **Objective**

Main research questions

Objective and Research Questions

- Our goal is to combine two avenues of research in nonprobability sampling for opt-in surveys: *representativeness* and *measurement error*.
- The main research questions are:



Outline of talk

3 **Data**

2021 Benchmarking Study by the Pew Research Center

Data: 2021 Benchmarking Study

- *Pew Research Center* designed the study - Mercer and Lau (2023).
- Objective was to determine accuracy of online surveys on general population estimates for all U.S. adults and key demographic subgroups.
- It is the first study to use benchmarking to identify subgroups where 'bogus responding' is concentrated.
- Benchmarks are obtained from NHANES, NHIS, ACS, CPS etc.
- For comparison with benchmark, Pew uses Mean Absolute Error (MAE):

$$\text{MAE} = m^{-1} \sum_{i=1}^{m} \left[{k_i}^{-1} \sum_{c=1}^{k_i} |\bar{y}_{ic} - \bar{Y}_{ic}| \right],$$

m : number of benchmark variables for a survey,
k_i : number of categories of a benchmark variable Y_i,
$\bar{y}_{ic}$: survey estimate for category c of Y_i
$\bar{Y}_{ic}$: estimate from benchmark source.

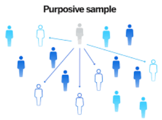
NHANES = National Health and Nutrition Examination Survey, NHIS = National Health Interview Survey, ACS = American Community Survey, CPS = Current Population Survey.

Survey details

1. Three probability-based (*ps*) online surveys:

- Obtained from Pew's American Trends Panel (ATP).
- Panelists are recruited offline using address-based sampling (ABS) from the U.S. Postal Service Computerized Delivery Sequence File.
- Typically, data collection is done online.

2. Three nonprobability (*nps*) commercial online surveys:



- Obtained from 'opt-in' (volunteer) panels through vendors.
- Vendors used quota sampling with a common set of quotas on age by gender, race and ethnicity, and educational attainment.

Survey Details

Table: Sample Sizes and Field Dates of 6 surveys administered between June 14 and July 21, 2021 with a common questionnaire in English/Spanish.

Sample type	Sample Name	ID	Sample size	Field dates
Probability	ABS panel 1	P_1	5,027	June 14 - 28, 2021
	ABS panel 2	P_2	5,147	June 14 - 27, 2021
	ABS panel 3	P_3	4,965	June 29 - July 21, 2021
Nonprobability	Opt-in panel 1	O_1	4,912	June 15 - 25, 2021
	Opt-in panel 2	O_2	4,931	June 11 – 27, 2021
	Opt-in panel 3	O_3	4,955	June 11 – 26, 2021

Table: Two of the $m = 25$ benchmarking variables used for MAE calculation: type, source, question and answer levels.

Variable	Source	Question	Answer levels
English proficiency (Categorical)	2019 ACS	How well do you speak English?	(i) Very well (ii) Well (iii) Not well (iv) Not at all
Received social security (Binary)	2021 CPS (March Supplement)	During 2020 did you receive any Social Security payments from the U.S. Government?	(i) Yes (ii) No

Pew's Key Findings

- Pew used calibration to weight all six samples.
- For probability sample design weights are known.
- For opt-in samples all respondents are assigned a base weight of 1.
- These weights for each sample are then calibrated to align with population benchmarks.
- The details are provided in the table from Pew's report.

Weighting dimensions

Variable	Benchmark source
Age x Gender Education x Gender Education x Age Race/Ethnicity x Education Born inside vs. outside the U.S. among Hispanics and Asian Americans Years lived in the U.S.	2019 American Community Survey
Census region x Metro/Non-metro	2020 CPS March Supplement
Volunteerism	2019 CPS Volunteering & Civic Life Supplement
Voter registration	2018 CPS Voting and Registration Supplement
Party affiliation Frequency of internet use Religious affiliation	2020 National Public Opinion Reference Survey (NPORS)

Note: Estimates from the ACS are based on non-institutionalized adults. Voter registration is calculated using procedures from Hur, Achen (2013) and rescaled to include the total U.S. adult population.

PEW RESEARCH CENTER

Pew's Key Findings

- MAE values are higher in opt-in samples as compared to ABS.
- At sub-group level, MAE values indicate a pattern in opt-in samples, which is not present in ABS samples.
- For *18-29 year-olds* and *Hispanic adults* MAE values are higher in opt-in samples.
- The above is attributed to 'bogus-responding' i.e. replying 'Yes' regardless of the question.
- Bogus responding is more prominent in questions regarding *government benefits* viz. social security, food stamps, unemployment compensation or workers' compensation.
- Large shares of $18-29$ year-old and Hispanic adults report having received at least three of four such benefits, which is very rare in the true population.

Outline of talk

4 **Improving representativeness**

Inverse Propensity Weighted estimation

Question 1

Does improving representativeness improve estimates from nonprobability surveys?

- Existing methods in literature can be used for this.
- We use the IPW method of Chen et al. (2020), assuming that IPW primarily removes selection bias from nonprobability samples.
- It is computationally advantageous - one set of estimated weights which can be used for any response variable.
- But assumes ignorable selection or missing at random (MAR) selection mechanism and requires good covariates.
- Goal is to estimate the unknown propensity scores for nps.
- Common auxiliary variables (from both ps and nps samples) and weights from ps samples are required and response variable is not needed.

Preliminaries

- For a finite population of size N, define
  S_A: *nps* of size n_A,
  S_B: *ps* of size n_B,
  d_i^B: design weight of unit i from S_B,
  R_i: selection indicator of unit i such that

$$R_i = \begin{cases} 1 & \text{if } i \in S_A, \\ 0 & \text{if } i \notin S_A, \end{cases}$$

  π_i^A: propensity score of unit i such that

$$\pi_i^A = P_q(R_i = 1 | x_i, y_i) = P_q(R_i = 1 | x_i), \quad i = 1, \ldots, N,$$

  where q: model for the selection mechanism for S_A.

- Assumptions are as follows:
  (A1) For i^{th} unit, R_i and y_i are independent given x_i.
  (A2) All units have a non-zero propensity score, i.e., $\pi_i^A > 0, \forall i$.
  (A3) R_i and R_j are independent given x_i and x_j for $i \neq j$.

Preliminaries

- For logistic propensity score model,

$$\pi(x_i, \theta_0) = \exp(x_i'\theta_0)/(1 + \exp(x_i'\theta_0)); \; i = 1, \ldots, N.$$

  with θ_0 as the true value of the unknown model parameters.

- The maximum likelihood estimator of π_i^A is computed as $\hat{\pi}_i^A = \pi(x_i, \hat{\theta})$, where $\hat{\theta}$ maximizes the log-likelihood function

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N} \left\{ R_i \log \pi_i^A + (1 - R_i) \log(1 - \pi_i^A) \right\} \\ &= \sum_{i \in S_A} \log \left\{ \frac{\pi(x_i, \theta)}{1 - \pi(x_i, \theta)} \right\} + \sum_{i=1}^{N} \log\{1 - \pi(x_i, \theta)\}. \end{aligned}$$

- The above cannot be used in practice since we do not observe x_i for all units in the finite population.
- Instead of using $l(\theta)$, we compute the estimator by maximizing the pseudo log-likelihood function $l^\star(\theta)$.

CLW Estimator: Method

- The pseudo-log-likelihood function is given by

$$l^\star(\theta) = \sum_{i \in S_A} \log\left\{\frac{\pi(x_i,\theta)}{1-\pi(x_i,\theta)}\right\} + \sum_{i \in S_B} d_i^B \log\{1-\pi(x_i,\theta)\}$$
$$= \sum_{i \in S_A} x_i'\theta - \sum_{i \in S_B} d_i^B \log\{1+\exp(x_i'\theta)\}$$

- Solution is obtained using Newton-Raphson iterative procedure.
- Using *nps* IPW estimator for the population mean $\mu_Y = N^{-1}\sum_{i=1}^N y_i$ is

$$\hat{\mu}_{CLW} \equiv (\hat{N}^A)^{-1} \sum_{i \in S_A} (y_i/\hat{\pi}_i^A),$$

  where $\hat{N}^A = \sum_{i \in S_A} (\hat{\pi}_i^A)^{-1}$.
- Using *ps*, we get the survey weighted estimator of y, denoted by

$$\hat{\mu}_{cal} \equiv (\hat{N}^B)^{-1} \sum_{i \in S_B} d_i^B y_i,$$

  where $\hat{N}^B = \sum_{i \in S_B} d_i^B$.

CLW Estimator: Method

- We estimate weights for 3 opt-in samples O_1, O_2, O_3, with P_1, P_2, P_3 respectively as reference, using `R` package `nonprobsvy`.
- For opt-in samples we calculate $\hat{\mu}_{CLW}$ and compare with $\hat{\mu}_{cal}$, calculated using calibrated weights from Pew.
- Evaluation criterion is Mean Squared Deviation/Difference (MSD) – error in terms of deviation from benchmark values.
- If $\bar{Y}_{ic}$ is the value for question i and subgroup c from benchmark source (considered as true value) and $\hat{\bar{Y}}_{ic}$ is the respective survey estimate (where $i = 1, \ldots, m;\ c = 1, \ldots, k$), then MSD for the estimator $\hat{\bar{Y}}$ is defined as

$$\text{MSD}(\hat{\bar{Y}}) \equiv m^{-1} \sum_{i=1}^{m} \left[k^{-1} \sum_{c=1}^{k} \left(\hat{\bar{Y}}_{ic} - \bar{Y}_{ic} \right)^2 \right].$$

CLW Estimator: Results

Table: MSD values (scaled to 10^2) of two population mean estimators, $\hat{\mu}_{cal}$ and $\hat{\mu}_{CLW}$, calculated using 12 benchmark questions from 6 surveys. For opt-in two estimators of population mean one using calibrated weights (from Pew) and another using IPW weights.

	P_1	P_2	P_3	O_1	O_2	O_3
MSD($\hat{\mu}_{cal}$)	0.080	0.210	0.097	1.039	1.024	0.480
MSD($\hat{\mu}_{CLW}$)	-	-	-	1.168	1.237	0.564

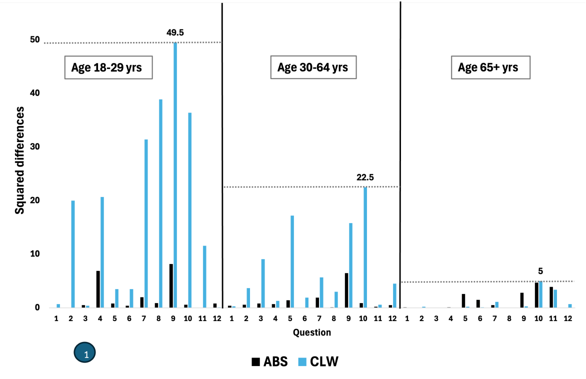
- Results show that MSD values for opt-in samples do not decrease even after using IPW weights as compared to calibrated weights.
- In comparison to ABS samples, MSD values for opt-in samples are still quite high.

- At sub-group level, results show that bias pertaining to specific sub-groups (as reported by Pew) is still present in opt-in samples.

Table: Subgroup level MSD values (scaled to 10^2) of $\hat{\mu}_{cal}$ and $\hat{\mu}_{CLW}$, at age groups $18-29, 30-64, 65+$ years and race groups White, Black and Hispanic.

	Factor	Group	P_1	P_2	P_3	O_1	O_2	O_3
MSD($\hat{\mu}_{cal}$)	age	18-29 (yrs)	0.158	0.507	0.177	2.869	3.162	1.790
		30-64 (yrs)	0.104	0.239	0.117	1.345	1.407	0.567
		65+ (yrs)	0.130	0.185	0.135	0.189	0.162	0.102
MSD($\hat{\mu}_{CLW}$)	age	18-29 (yrs)	-	-	-	2.312	3.901	1.930
		30-64 (yrs)	-	-	-	1.497	1.705	0.728
		65+ (yrs)	-	-	-	0.318	0.146	0.097
MSD($\hat{\mu}_{cal}$)	race	White	0.064	0.152	0.090	1.034	0.907	0.390
		Black	0.281	0.502	0.376	1.099	1.121	0.499
		Hispanic	0.102	0.375	0.236	2.625	2.341	2.147
MSD($\hat{\mu}_{CLW}$)	race	White	-	-	-	0.913	1.136	0.471
		Black	-	-	-	1.351	0.877	0.464
		Hispanic	-	-	-	3.441	2.781	2.444

Figure 1: *Plot of squared differences of two population mean estimates with benchmark* $\bar{Y}$, *i.e.,* $(\bar{Y} - \hat{\mu}^{P_3}_{cal})^2$ *in black and* $(\bar{Y} - \hat{\mu}^{O_3}_{CLW})^2$ *in blue, for* 12 *questions and* 3 *age groups* $18-29, 30-64, 65+$ *years. Highest values of blue bars in each age group are denoted in dotted lines.* Y*-axis is scaled to* 10^3.



No.	Question
1	Health insurance
2	High blood pressure
3	Parent
4	Food allergy
5	Job last year
6	Retirement account
7	Unemployment compensation
8	Workers' compensation
9	Food stamps
10	Social Security
11	Union membership
12	U.S. citizenship

- Thus improving representativeness doesn't improve *nps* estimates.

5 **Bias Correction**

Proposed method and results

Question 2

How to correct for measurement difference in nps?

- Prior knowledge from Pew's findings confirms that bias due to age and race is present.
- We compute estimates of bias at question × group level using O_1, O_2 and P_1, P_2.
- We keep O_3 and P_3 for validating the bias-corrected estimates.
- For this analysis, we only consider 12 binary questions (yes/no answer).
- Assume that the distribution of bias is similar between *nps*.
- For the lack of availability of benchmarks at more granular level (age × race) we do not consider both factors together.

Bias Correction: Method

- For the i^{th} question and c^{th} category of a factor, estimate of bias is

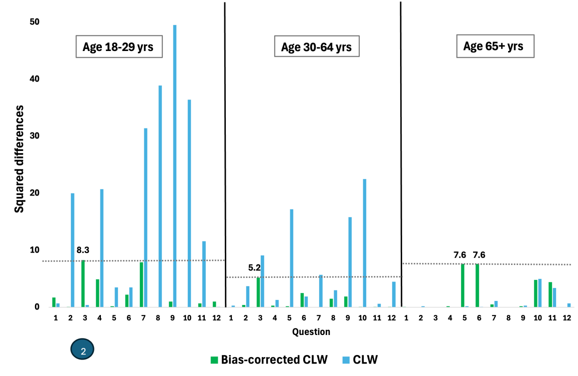
$$\begin{aligned}\hat{\epsilon}^{ic} &= 4^{-1}\Big\{(\hat{\mu}_{CLW}^{ic;O_1} - \hat{\mu}_{cal}^{ic;P_1}) + (\hat{\mu}_{CLW}^{ic;O_1} - \hat{\mu}_{cal}^{ic;P_2}) + (\hat{\mu}_{CLW}^{ic;O_2} - \hat{\mu}_{cal}^{ic;P_1}) \\ &\qquad + (\hat{\mu}_{CLW}^{ic;O_2} - \hat{\mu}_{cal}^{ic;P_2})\Big\} \\ &= 2^{-1}\Big\{(\hat{\mu}_{CLW}^{ic;O_1} - \hat{\mu}_{cal}^{ic;P_1}) + (\hat{\mu}_{CLW}^{ic;O_2} - \hat{\mu}_{cal}^{ic;P_2})\Big\};\; i = 1, \ldots, n; c = 1, \ldots, k.\end{aligned}$$

- Bias-corrected CLW estimate of proportion of yes for question j in sub-group c from validation sample O_3 is

$$\hat{\mu}_{bc;CLW}^{ic;O_3} = \hat{\mu}_{CLW}^{ic;O_3} - \hat{\epsilon}^{ic}.$$

- We compare $\hat{\mu}_{bc;CLW}^{ic;O_3}$ and $\hat{\mu}_{cal}^{ic;P_3}$ with benchmarks from Pew's analysis.

Figure 2: *Plot of squared differences of* $\hat{\mu}^{O_3}_{CLW}$ *and* $\hat{\mu}^{O_3}_{bc;CLW}$ *with benchmark, i.e.,* $(\bar{Y} - \hat{\mu}^{O_3}_{CLW})^2$ *in blue and* $(\bar{Y} - \hat{\mu}^{O_3}_{bc;CLW})^2$ *in green, for* 12 *questions and age groups* $18-29, 30-64, 65+$ *years.*



No.	Question
1	Health insurance
2	High blood pressure
3	Parent
4	Food allergy
5	Job last year
6	Retirement account
7	Unemployment compensation
8	Workers' compensation
9	Food stamps
10	Social Security
11	Union membership
12	U.S. citizenship

- Squared differences are lower for bias-corrected CLW estimators.
- Thus improving measurement error as well produces better estimates.

Outline of talk

6. **Composite estimators**

Existing and proposed method and Results

Question 3

How to combine estimates from ps and nps when both contain response variables?

- Consider following sampling model:

$$\begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \sim \left(\begin{pmatrix} \mu \\ \mu + \epsilon \end{pmatrix}, \begin{pmatrix} v_1 & 0 \\ 0 & v_2 \end{pmatrix} \right).$$

$\bar{y}_1$: survey-weighted estimator of population mean from *ps* ($\hat{\mu}_{cal}$ in our case)
$\bar{y}_2$: estimator from *nps* ($\hat{\mu}_{CLW}$ in our case)
μ : true mean of the *ps* estimator (unknown)
ϵ : bias in *nps* estimator (known),
v_1 : variance of *ps* estimators,
v_2 : variance of *nps* estimators.

- Since the *nps* and *ps* are sampled independently, we assume correlation between them to be 0.

- We focus on combining $\bar{y}_1$ and $\bar{y}_2$ to create composite estimators.

Composite estimators: Method

- Elliott and Haviland (2007) consider a composite estimator of the form:

$$\hat{\mu}_{EV} = \frac{(\epsilon^2 + v_2)\bar{y}_1 + v_1\bar{y}_2}{\epsilon^2 + v_1 + v_2}.$$

- $\hat{\mu}_{EV}$ is a biased estimator for μ, with remaining bias

$$b_{EV} = E(\hat{\mu}_{EV}) - \mu = \frac{\epsilon v_1}{\epsilon^2 + v_1 + v_2}.$$

- We propose an unbiased composite estimator – a weighted combination of $\bar{y}_1$ and 'bias-corrected' estimates $\bar{y}_2 - \epsilon$, as follows:

$$\hat{\mu}_{comb} = \left(\frac{v_2}{v_1 + v_2}\right) \times \bar{y}_1 + \left(\frac{v_1}{v_1 + v_2}\right) \times (\bar{y}_2 - \epsilon).$$

- $\hat{\mu}_{comb}$ and $\hat{\mu}_{EV}$ both minimize MSE, but in disjoint classes of estimators.
- It can be shown that

$$\text{MSE}(\hat{\mu}_{comb}) = \frac{v_1 v_2}{v_1 + v_2} < \frac{v_1(\epsilon^2 + v_2)}{\epsilon^2 + v_1 + v_2} = \text{MSE}(\hat{\mu}_{EV})$$

Composite estimators: Method

- The bias term ϵ needs to be known possibly from alternative sources.
- Assume that $\hat{\epsilon}$, an estimator of ϵ, is constructed from the auxiliary survey that is independent of $\bar{y}_2$.
- Let $E(\hat{\epsilon}) = \epsilon$, $\text{Var}(\hat{\epsilon}) = v_b$.
- Under these assumptions, the composite estimator with optimal weight i.e. that minimizes MSE is:

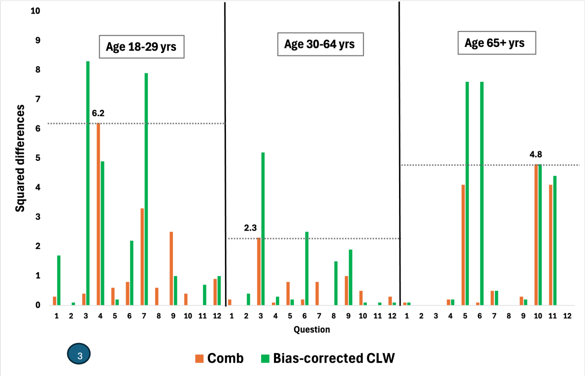
$$\tilde{\mu}_{comb} \equiv \left(\frac{v_2 + v_b}{v_1 + v_2 + v_b}\right) \bar{y}_1 + \left(\frac{v_1}{v_1 + v_2 + v_b}\right) (\bar{y}_2 - \hat{\epsilon})$$

  and the MSE of $\tilde{\mu}_{comb}$ is $\text{MSE}(\tilde{\mu}_{comb}) = v_1(v_2 + v_b)/(v_1 + v_2 + v_b)$.

- $\text{MSE}(\tilde{\mu}_{comb}) \leqslant \text{MSE}(\hat{\mu}_{EV})$ provided $v_b \leqslant \epsilon^2$.

Composite estimators: Results

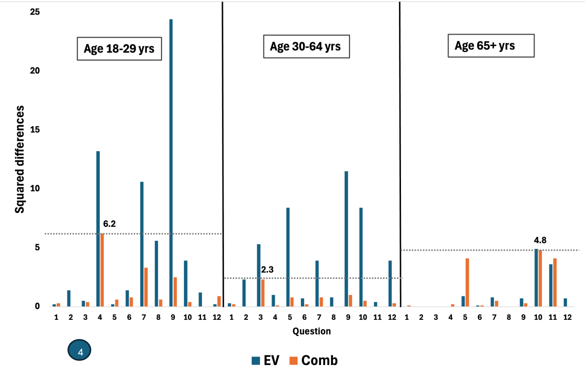
Figure 3: *Plot of squared differences with benchmark of* $\hat{\mu}^{O_3}_{bc:CLW}$ *(using* O_3*) and* $\tilde{\mu}_{comb}$ *(using* O_3 *and* P_3*), i.e.,* $(\bar{Y} - \hat{\mu}^{O_3}_{bc:CLW})^2$ *in green and* $(\bar{Y} - \tilde{\mu}_{comb})^2$ *in orange, for* 12 *questions and* 3 *age groups* $18-29, 30-64, 65+$ *years.*



No.	Question
1	Health insurance
2	High blood pressure
3	Parent
4	Food allergy
5	Job last year
6	Retirement account
7	Unemployment compensation
8	Workers' compensation
9	Food stamps
10	Social Security
11	Union membership
12	U.S. citizenship

Composite estimators: Results

Figure 4: *Plot of squared differences with benchmark of composite estimators* $\hat{\mu}_{EV}$ *and* $\tilde{\mu}_{comb}$*, i.e.,* $(\bar{Y} - \hat{\mu}_{EV})^2$ *in dark blue and* $(\bar{Y} - \tilde{\mu}_{comb})^2$ *in orange, for* 12 *questions and* 3 *age groups* $18 - 29, 30 - 64, 65+$ *years.*



No.	Question
1	Health insurance
2	High blood pressure
3	Parent
4	Food allergy
5	Job last year
6	Retirement account
7	Unemployment compensation
8	Workers' compensation
9	Food stamps
10	Social Security
11	Union membership
12	U.S. citizenship

Predictive model for bias correction

- When other sources of information about bias are not available, we fit a machine learning (ML) model on P_3.
- Using fitted parameter estimates from ML model, we predict the unit level responses in O_3 (or equivalently, the probabilities of responding 'Yes' : $\hat{p}_i^y$).
- We then use the same IPW method on the predicted probabilities and finally produce the following estimator of *nps*

$$\hat{\mu}_{ML;CLW} \equiv (\hat{N}^A)^{-1} \sum_{i \in S_A} (\hat{p}_i^y / \hat{\pi}_i^A),$$

  - $\hat{p}_i^y$ is the estimated likelihood that individual i responds 'Yes' to variable y given the auxiliary characteristics of i,
  - $\hat{\pi}_i^A$ is obtained from IPW method,
  - $\hat{N}^A = \sum_{i \in S_A} (\hat{\pi}_i^A)^{-1}$.

Composite estimators

- Composite estimator under this scenario, denoted by $\hat{\mu}_{ML;comb}$, is

$$\hat{\mu}_{ML;comb} \equiv \left(\frac{v_2}{v_1 + v_2}\right)\hat{\mu}_{cal} + \left(\frac{v_1}{v_1 + v_2}\right)\hat{\mu}_{ML;CLW}.$$

- We compare the above estimator with a different version of $\hat{\mu}_{EV}$, denoted by $\hat{\mu}_{ML;EV}$, defined as

$$\hat{\mu}_{ML;EV} \equiv \frac{\{(\bar{y}_1 - \bar{y}_2)^2 + v_2\}\bar{y}_1 + v_1\bar{y}_2}{(\bar{y}_1 - \bar{y}_2)^2 + v_1 + v_2}.$$

- In this estimator, ϵ is estimated by the difference of weighted means from *ps* and *nps*, as suggested by Elliott and Haviland (2007).

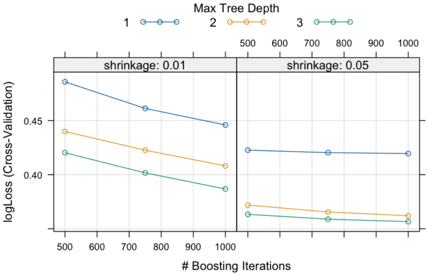
Modeling Results

- For model building we use the R package caret to fit - Random Forest (RF) and Gradient Boosting Model (GBM), using cross validation (CV) as the re-sampling method and tuning several hyper-parameters (number of boosting iterations, maximum tree depth, shrinkage, etc).

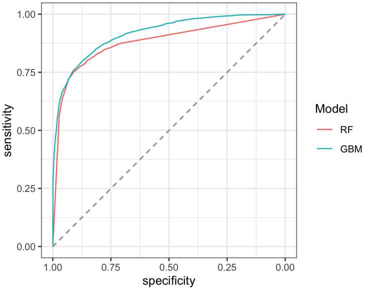
Data structure	Question × respondent level data
Size (Train + test)	59 thousand observations split into train (80%) and test (20%)
Response variable type	Binary (Refusals are not considered)
Auxiliary variables (no. of levels)	Question (12), age (3), race-ethnicity (5), education (3), gender (3), region (4)
Predictive models (AUC, Accuracy)	Random Forest (0.88, 0.85), Gradient Boosting (0.92, 0.85)

Modeling Results

(a) Tuning parameters of GBM



(b) ROC curve of RF and GBM



Composite estimator Results

Table: Composite Estimators $\hat{\mu}_{ML;comb}$ (proposed) and $\hat{\mu}_{ML;EV}$ (existing) and their MSD values (in 10^3 scale) for 12 binary benchmark questions.

Question	Benchmark	$\hat{\mu}_{ML;EV}$	MSD($\hat{\mu}_{ML;EV}$)	$\hat{\mu}_{ML;comb}$	MSD($\hat{\mu}_{ML;comb}$)
1. Insurance	90.800	88.979	4.450	90.755	4.447
2. Blood Pressure	31.100	37.725	5.271	35.709	5.109
3. Parent	26.000	22.336	10.189	25.977	10.099
4. Food allergy	9.400	14.038	0.984	12.823	0.966
5. Job last year	64.200	58.262	20.105	63.476	19.817
6. Retirement account	49.900	49.866	25.016	51.119	25.015
7. Unemployment compensation	9.300	16.846	0.970	12.189	0.825
8. Workers' compensation	0.400	6.637	0.284	1.144	0.010
9. Food stamps	11.100	21.549	5.328	16.416	5.247
10. Social Security	21.800	29.898	26.574	26.189	26.413
11. Union membership	5.600	10.410	0.657	8.055	0.595
12. U.S. citizenship	92.500	96.299	4.413	96.064	4.332

- We observe that for all questions, MSD($\hat{\mu}_{ML;comb}$) $<$ MSD($\hat{\mu}_{ML;EV}$).

Summary : Composite estimator

- Finally, we summarize the two cases of bias-correction –

Case 1: When bias is computed from alternative sources.
Case 2: When bias-corrected estimates are predicted using modeling.

Table: Comparison of MSD values (in 10^2 scale, given inside parenthesis) of proposed and competing estimators calculated using 12 benchmark questions from *nps* O_3 and *ps* P_3 (in case of composite estimators) for two cases of bias correction.

	Proposed Composite	Existing Composite	Other
Case 1:	$\tilde{\mu}_{comb}(0.101)$	$\hat{\mu}_{EV}(0.337)$	$\hat{\mu}_{CLW}(0.564)$ $\hat{\mu}_{bc;CLW}(0.183)$
Case 2:	$\hat{\mu}_{ML;comb}(0.092)$	$\hat{\mu}_{ML;EV}(0.355)$	$\hat{\mu}_{ML;CLW}(0.099)$

- Bias-correction is effective in both cases by significantly lowering MSD.
- In both cases proposed estimators ($\tilde{\mu}_{comb}$ and $\hat{\mu}_{ML;comb}$) have lower MSD values than existing composite estimators ($\hat{\mu}_{EV}$ and $\hat{\mu}_{ML;EV}$) as well as other estimators ($\hat{\mu}_{CLW}$, $\hat{\mu}_{bc;CLW}$ and $\hat{\mu}_{ML;CLW}$).

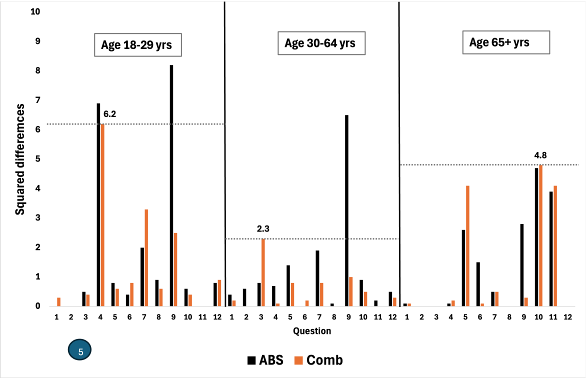
Question 4

When do composite estimates outperform estimates from probability surveys alone?

- In our case study *ps* already has a large sample size (close to 5k).
- As a result *ps* itself is good enough (in most cases) for estimation of finite population characteristics.
- For the age group $18 - 29$ years, we see that the estimator from *ps* alone is better than proposed composite estimator for some questions.
- We want to investigate if the performance of composite estimator can be further improved.
- In order to explain situations where performance of the proposed composite estimator is significantly better in terms of MSD , we generate *ps* of smaller sample sizes.

Question 4

Figure 6: *Plot of squared differences of* $\hat{\mu}_{cal}^{P_3}$ *(using* P_3*) and* $\tilde{\mu}_{comb}$ *(using* P_3 *and* O_3*) with benchmark, i.e.,* $(\bar{Y} - \hat{\mu}_{cal}^{P_3})^2$ *in black and* $(\bar{Y} - \tilde{\mu}_{comb})^2$ *in orange, for* 12 *questions and* 3 *age groups* $18 - 29, 30 - 64, 65+$ *years.*



No.	Question
1	Health insurance
2	High blood pressure
3	Parent
4	Food allergy
5	Job last year
6	Retirement account
7	Unemployment compensation
8	Workers' compensation
9	Food stamps
10	Social Security
11	Union membership
12	U.S. citizenship

Question 4

- We draw samples from P_3 using a stratified sampling procedure with proportional allocation.
- Strata are created by segmentation based on survey weights in P_3, viz., $(0, 0.5), [0.5, 1)$ and so on.
- Calculate MSD values of $\hat{\mu}_{EV}$, $\tilde{\mu}_{comb}$ and $\hat{\mu}_{cal}$ for smaller samples from P_3, keeping the original *nps* O_3 intact.

Table: MSD values (in 10^2 scale) of estimator $\hat{\mu}_{cal}$, $\tilde{\mu}_{comb}$ and $\hat{\mu}_{EV}$. Numbers presented in $(\cdot)$ correspond to the relative change in MSD (in %) of the estimators compared to $\hat{\mu}_{cal}$.

ps sample size	$\hat{\mu}_{cal}$	$\hat{\mu}_{EV}$	$\tilde{\mu}_{comb}$
4912	0.097	0.337 (-71.284)	0.101 (-3.767)
1000	0.231	0.295 (-21.718)	0.158 (45.896)
500	0.399	0.385 (3.642)	0.289 (38.293)
100	1.073	0.748 (43.476)	0.626 (71.399)

- When *ps* has an adequate sample size, composite estimator is not better, than the estimator from *ps*.
- As sample size of *ps* decreases the variance of $\hat{\mu}_{cal}$ increases - $\tilde{\mu}_{comb}$ outperforms $\hat{\mu}_{cal}$.

Discussion

- We focus on improving finite population inference using statistical data integration from probability and nonprobability surveys.
- We have observed that only improving representativeness is not enough for nonprobability surveys (especially opt-in ones as in this data analysis).
- Effective bias correction can be done if alternate data sources and benchmarks are available.
- We used modern machine learning techniques for bias correction and proposed new composite estimators.
- While comparing such composite estimator with estimator from probability survey, we observed that upon reducing sample size of P_3, some subgroups do not have any observations.
- In such cases, small area modeling techniques are required (Nandram and Rao (2024)) - currently under investigation.

References

Beaumont, J., K. Bosa, A. Brennan, J. Charlebois, and K. Chu (2024). Handling nonprobability samples through inverse probability weighting with an application to statistics canada's crowdsourcing data. *Survey Methodology, Statistique Canada 50*(1), 77–106.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology 46*(1), 1–29.

Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association 115*(532), 2011–2021.

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

Kim, J. K., S. Park, Y. Chen, and C. Wu (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society 184*(3), 941–963.

Lee, S. and R. Valliant (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research 37*(3), 319–343.

Mercer, A. and A. Lau (2023). Comparing two types of online survey samples. *Pew Research Center*. https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/.

Nandram, B. and J. Rao (2024). Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. *Statistics and Applications ISSN 2454-7395 (online) 22*, 343–374.

Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B 83*(1), 242–272.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*, Volume 4. American Statistical Association Alexandria, VA.

Valliant, R. and J. A. Dever (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research 40*(1), 105–137.

Wang, L., R. Valliant, and Y. Li (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med. 40*(4), 5237–5250.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology, Statistics Canada 48*(2), 5237–5250.

Sen and Lahiri (2026) - To appear in *Survey Methodology* June issue

Thank You!!