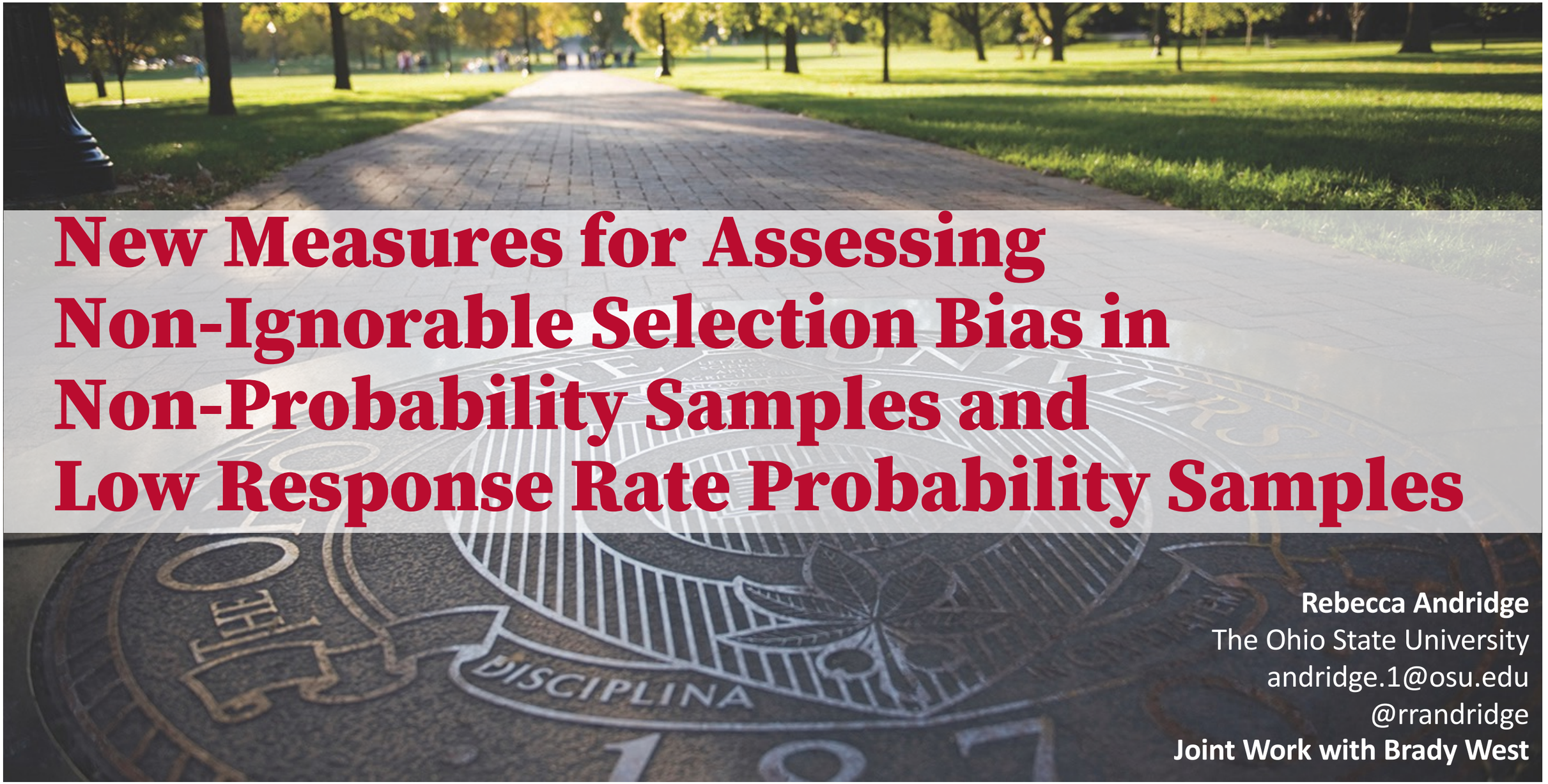


New Measures for Assessing Non-Ignorable Selection Bias in Non-Probability Samples and Low Response Rate Probability Samples

Rebecca Andridge
The Ohio State University
andridge.1@osu.edu
@rrandridge
Joint Work with Brady West

Acknowledgements

- This work was supported by an R21 grant from NIH (PI: West; NIH Grant No. 1R21HD090366-01A1)
- Thank you to my co-authors: Brady West*, Phil Boonstra, Rod Little, and Fernanda Alvarado-Leiton
- Thanks to Dr. Mick Couper for letting us work with the NSFG data!

The National Survey of Family Growth (NSFG) is conducted by the Centers for Disease Control and Prevention's (CDC's) National Center for Health Statistics (NCHS), under contract # 200-2010-33976 with University of Michigan's Institute for Social Research with funding from several agencies of the U.S. Department of Health and Human Services, including CDC/NCHS, the National Institute of Child Health and Human Development (NICHD), the Office of Population Affairs (OPA), and others listed on the NSFG webpage (see http://www.cdc.gov/nchs/nsfg/). The views expressed here do not represent those of NCHS nor the other funding agencies.

Problem Statement

- "Big Data" = everywhere, inexpensive, nonprobability samples
  - Need model-based approaches for inference (Elliott and Valliant 2017)
- Selection bias a (the?) major concern with nonprobability samples
  - Also a concern with low response rate probability samples!
- Existing methods for assessing sample representativeness inadequate
  - **R-indicator** (Schouten et al. 2009) depends only on response propensity and is agnostic about the survey variables of interest
  - **H_1 indicator** (Sarndal and Lundstrom 2010) is based on the variables of interest, but assumes an ignorable selection mechanism
- **Our Goal: Develop tools exist for gauging the amount of non-ignorable selection bias in survey estimates arising from non-probability sampling**

Approach for Means (Continuous Variables)

- Assume we have:
  - Continuous variable of interest Y, covariates Z from non-probability sample
  - **Aggregate population information for Z**, via administrative records or some other source (e.g., a large probability sample producing small standard errors)
- Use Z to develop the **best predictor** of Y from Z
  - E.g., linear predictor of Y from a regression of Y|Z (in selected sample)
- Call this "best" predictor of Y an *auxiliary proxy* for Y, and call it X
  - Based on aggregate Z info we can obtain $\bar{X}$ = mean of X for the population
- Let U = other covariates in Z that are are orthogonal to X
- Let S = selection indicator (in non-probability sample)

Approach for Means, cont'd

- Basic idea:
  - We can measure the amount of selection bias in X (Z)
  - If Y is correlated with X (Z), this tells you something about the potential selection bias in Y

- Use pattern-mixture models to explicitly model non-ignorable selection (i.e., selection dependent on Y)

Approach for Means, cont'd

- Bivariate normal pattern-mixture model:

$$(X, Y|S = j) \sim N_2\left(\left(\mu_X^{(j)}, \mu_Y^{(j)}\right), \begin{bmatrix} \sigma_{XX}^{(j)} & \rho_{XY}^{(j)}\sqrt{\sigma_{XX}^{(j)}\sigma_{YY}^{(j)}} \\ \rho_{XY}^{(j)}\sqrt{\sigma_{XX}^{(j)}\sigma_{YY}^{(j)}} & \sigma_{YY}^{(j)} \end{bmatrix}\right)$$

$$\Pr(S = 1|X, Y, U) = g\big((1-\phi)X^* + \phi Y, U\big)$$

Note: X^*= X, rescaled to have same variance as Y

- Selection probability allowed to depend on both X* (rescaled proxy) and Y through ϕ; g() arbitrary
  - ϕ = 0 → selection is ignorable, depending on X* (and U) only
  - 0 < ϕ < 1 → selection is non-ignorable, depends at least partially on Y
  - ϕ = 1 → selection is "extremely" non-ignorable, depending on Y (and U) only
- No information in the data about ϕ → vary ϕ in a **sensitivity analysis**

Approach for Means, cont'd

- Andridge and Little (2011) show that the ML estimate of the mean of Y, given ϕ, is the following (note that rescaling of X is incorporated):

$$\hat{\mu}_Y(\phi) = \bar{y}^{(1)} + \frac{\phi + (1-\phi)r_{XY}^{(1)}}{\phi r_{XY}^{(1)} + (1-\phi)} \sqrt{\frac{s_{YY}^{(1)}}{s_{XX}^{(1)}}} \left(\bar{X} - \bar{x}^{(1)}\right)$$

- $r_{XY}^{(1)}$ = estimated correlation of X and Y in the *non-probability sample*
- Two proposed measures based on this result:
  - **Measure of unadjusted bias (MUB)** and **Standardized MUB (SMUB)**:

$$MUB(\phi) = \bar{y}^{(1)} - \hat{\mu}_Y(\phi) = \frac{\phi + (1-\phi)r_{XY}^{(1)}}{\phi r_{XY}^{(1)} + (1-\phi)} \sqrt{\frac{s_{YY}^{(1)}}{s_{XX}^{(1)}}} \left(\bar{X} - \bar{x}^{(1)}\right)$$

$$SMUB(\phi) = \frac{MUB(\phi)}{\sqrt{s_{YY}^{(1)}}} = \frac{\phi + (1-\phi)r_{XY}^{(1)}}{\phi r_{XY}^{(1)} + (1-\phi)} \frac{\left(\bar{X} - \bar{x}^{(1)}\right)}{\sqrt{s_{XX}^{(1)}}}$$

Using the Proposed Index

- Proposed index is **simple** as it only depends on:
  - Means, SDs, and correlation from the observed non-probability sample
  - Population mean for proxy X
  - Sensitivity parameter ϕ
- Little et al. (2020) propose an intermediate choice of ϕ = 0.5 for computing SMUB, along with an "interval" for the selection bias based on the extreme cases of ϕ:

$$SMUB(0.5) = \frac{\left(\bar{X} - \bar{x}^{(1)}\right)}{\sqrt{s_{XX}^{(1)}}}$$

Interval:

$$SMUB(0) = r_{XY}^{(1)} \frac{\left(\bar{X} - \bar{x}^{(1)}\right)}{\sqrt{s_{XX}^{(1)}}} \quad \text{and} \quad SMUB(1) = \frac{1}{r_{XY}^{(1)}} \frac{\left(\bar{X} - \bar{x}^{(1)}\right)}{\sqrt{s_{XX}^{(1)}}}$$

Additional Remarks on (S)MUB

- **SMUB(1)** = unstable when X (Z) is not a good predictor of Y
  - Bias in Y cannot be reliably estimated
- **SMUB(0.5) =** related to the Bias Effect Size proposed by Biemer and Peytchev at the 2011 Nonresponse Workshop
  - We show that this has a model-based justification
  - We use the full bias expression in the numerator ($\bar{X} - \bar{x}^{(1)}$), rather than difference between non-selected and selected means ($\bar{x}^{(0)} - \bar{x}^{(1)}$)
- MLEs don't incorporate uncertainty in creation of X
- Alternatively, a **fully Bayesian approach** to computing (S)MUB available that incorporates uncertainty
  - Caution: Bayesian approach additionally requires variance/covariance matrix for Z for the non-selected cases
  - The proposed MLE-based interval is a recommendation for practice, but we have an R function available for the Bayesian approach given the necessary information

Simulation Results: Normal Case

- A comprehensive simulation study replicating Nishimura et al. (2016) has demonstrated excellent performance of the proposed SMUB index relative to other competing tools across a variety of scenarios
- Relative to other potential diagnostics (e.g., the R-indicator and the coefficient of variation of the selection propensities), the SMUB index has a much stronger correlation with, and is more predictive of, the true (unknown) selection bias
- See **Boonstra et al. (2021)** for details

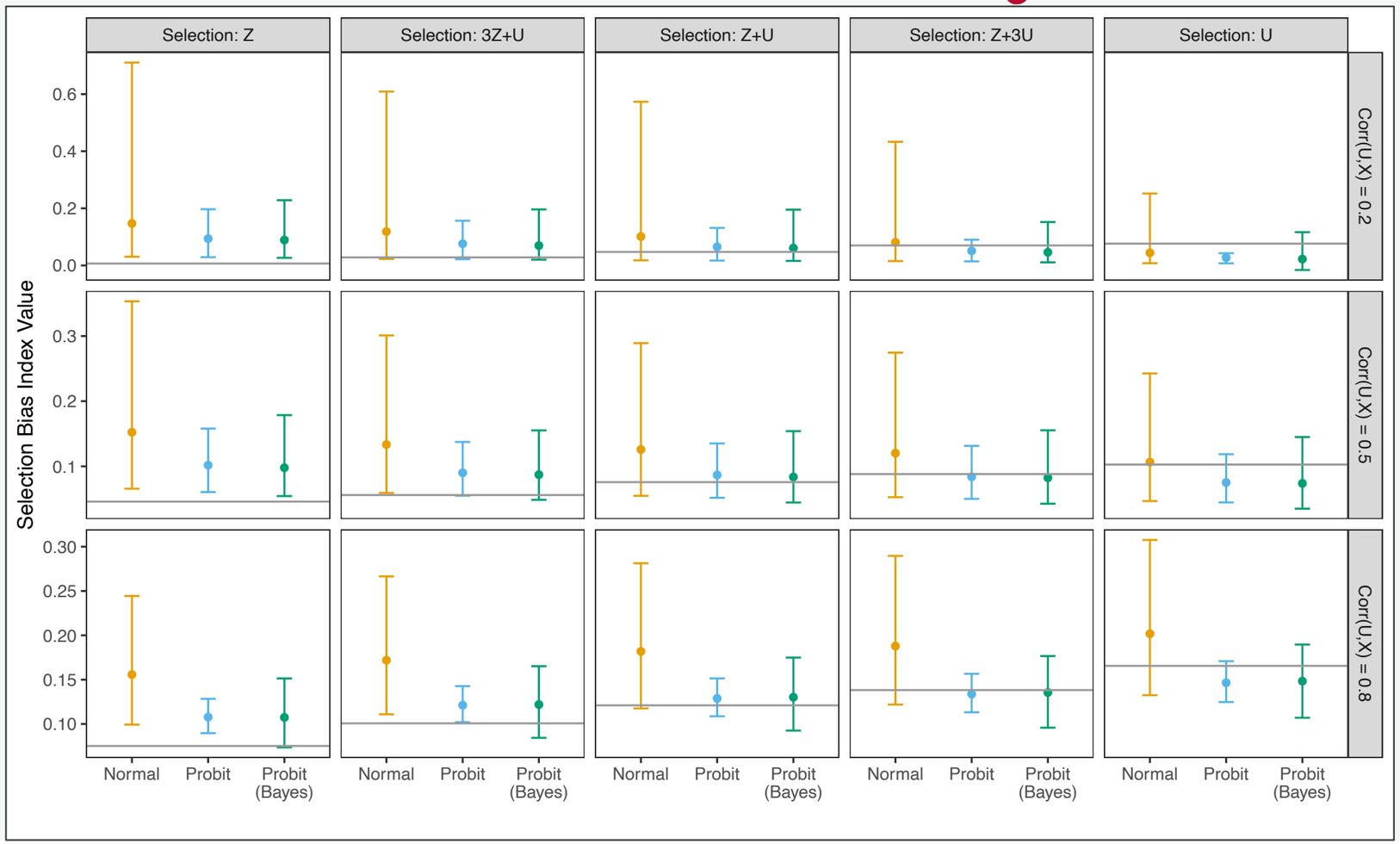
Approach for Binary Y

- Per Andridge and Little (2009, 2020), assume binary variable Y arises from a latent variable U that follows a normal distribution
  - Create proxy X from a probit regression of Y on Z
  - Key measure = **biserial correlation** of X and Y in selected sample = $r_{XU}^{(1)}$
- Follow similar approach as (S)MUB using a pattern-mixture model for U and X
- Form similar indices of selection bias based on the observed respondent proportion $\bar{y}^{(1)}$ and ("modified") MLEs for pattern-mixture model
- Proposed measure: **Measure of Unadjusted Bias for a Proportion (MUBP)**
  - See **Andridge et al. 2019** for details
  - No need for standardization
  - As with SMUB, can use MUBP(0), MUBP(0.5), and MUBP(1) indices for forming intervals

$$MUBP(\phi) = \bar{y}^{(1)} - \hat{\mu}_Y(\phi)$$

- Alternative estimation via a **fully Bayesian approach** for forming credible intervals for MUBP (given sufficient statistics on Z for the non-selected cases)

Simulated Data: Binary Case



- The gray horizontal line represents actual bias
- SMUB intervals for the bias (orange, the normal model) are **substantially wider**
- MUBP intervals (blue, green, based on the probit model) are much narrower and tend to cover the bias equally well (especially with Bayesian approach)

Example: NSFG Smartphone Users

- Hypothetical population = Data from 16 quarters (2012-2016) of the National Survey of Family Growth (NSFG)
- Hypothetical non-probability sample = smartphone users in this population as our hypothetical non-probability sample
  - Simulates a selection process, see Couper et al. (2018) for details
- Our Y variables were variables of interest to NSFG data users
  - Considered both continuous and binary Y variables
- Our Z variables included those where pop. aggregates may be available:
  - age, race/ethnicity, marital status, education, household income, region of the U.S. (based on definitions from the U.S. Census Bureau), current employment status, and presence of children under the age of 16 in the household
- We regressed Y on Z for smartphone users to form our linear predictor X
  - Separately for males and females*

**The binary representation of gender in the NSFG does not reflect our current understanding of the complex construct of gender*

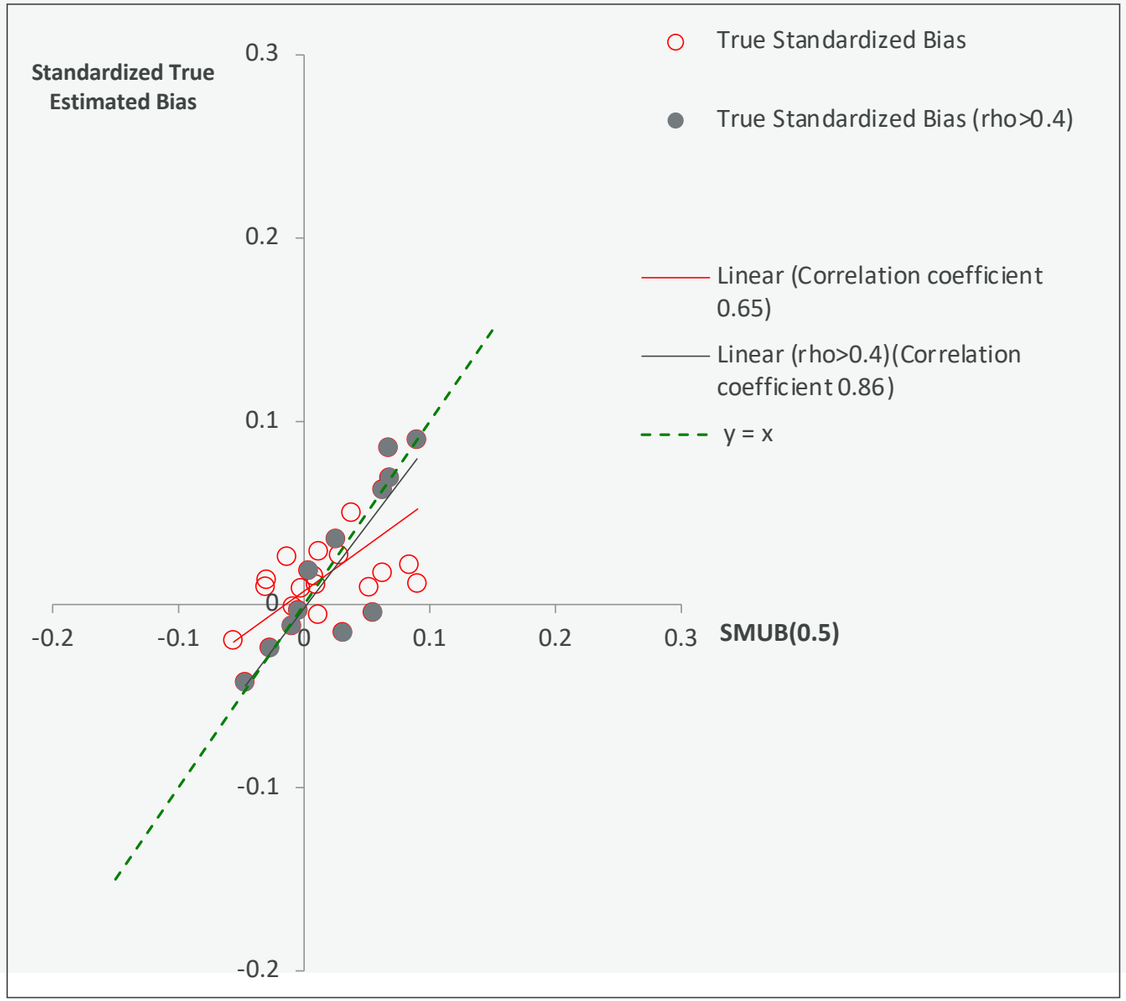
Evaluation

- Computed our proposed indices (SMUB and MUBP) and intervals for each of several Y variables, for males and females separately
- Computed the **standardized true estimated bias (STEB)** for each of our estimates (the “truth”)
- Also considered the **fraction of missing information** (FMI) as a competing index, given some of the results in Nishimura et al. (2016)
- Assessed how often our proposed interval covered the STEB, and the correlations of our indices with the STEB values

Results: Proposed Interval (SMUB)

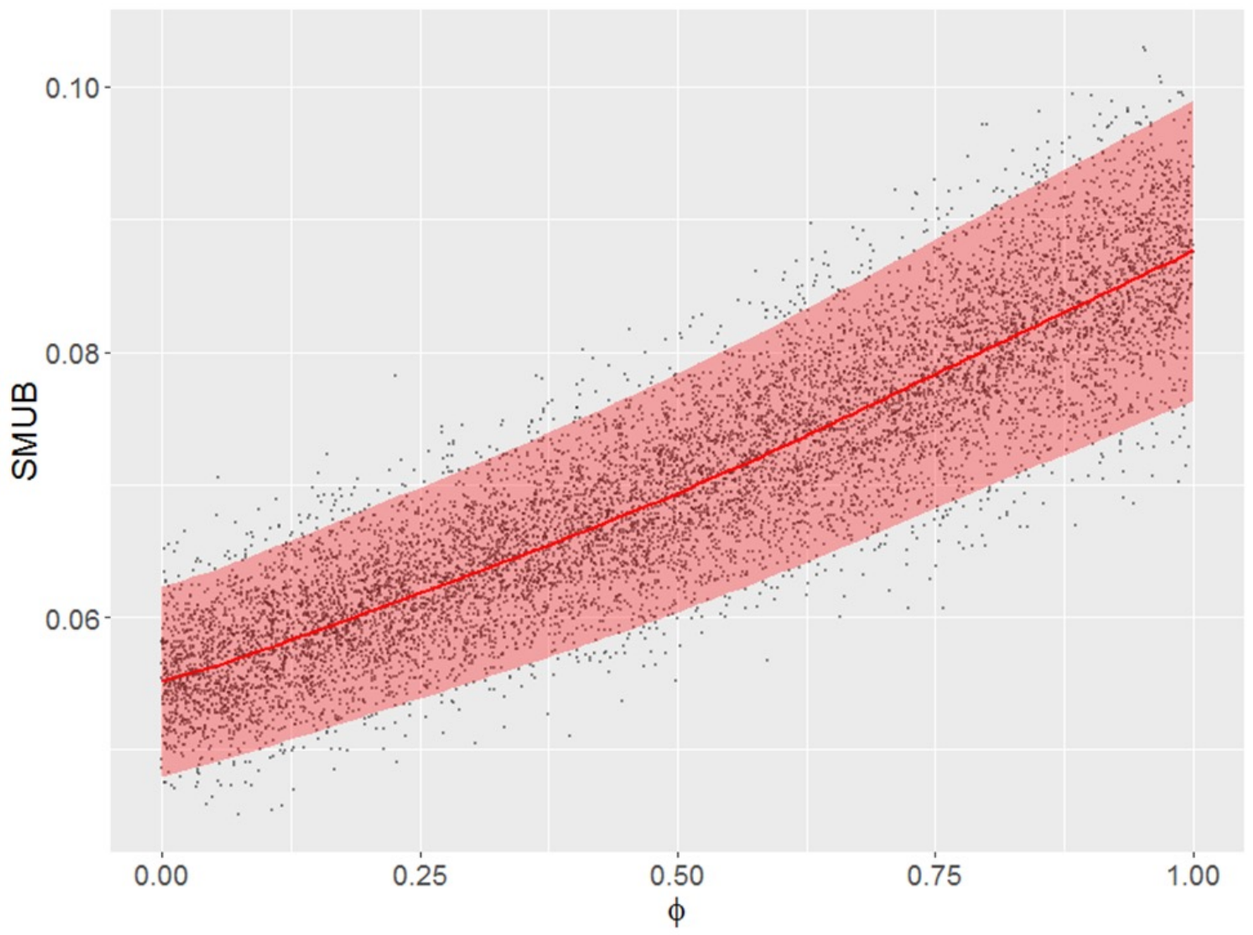
- The proposed interval covers the actual STEB in **9/12 cases** where the correlation of X with Y in the non-probability sample was above 0.4 (**good proxies**)
- The proposed interval only covered the STEB in 5/16 cases where the correlation was less than 0.4
- The proposed interval becomes much wider when the correlation becomes smaller!

Results: Correlation of SMUB(0.5) with STEB



- Perfect correlation = dashed green line
- Strong correlation (r=0.65) of the SMUB(0.5) values with the actual STEB (true bias) values
- **We see an even stronger correlation (r=0.86) when focusing on variables with proxy strengths (correlations) greater than 0.4 in the non-probability sample**
- The FMI had a small negative correlation: not a useful measure (Nishimura et al. 2016)

Results: Bayesian Approach (SMUB)



- This plot shows draws of SMUB(ϕ) given draws of ϕ [from Uniform(0,1)], with predicted values of SMUB(ϕ) as a function of ϕ and 95% credible intervals
- NSFG variable shown = number of months worked in the past year for females, with **STEB = 0.069**
- **This plot is automatically generated when running our `nisb_bayes()` function in R for a given set of variables**
- Note that a choice of ϕ = 0.5 results in draws of SMUB that closely reflect the true bias (**for THIS variable**)
- The proposed interval, allowing for uncertainty, also covers the true bias in this case; **the Bayesian approach had better coverage for smaller correlations**

Applying the Binary Approach (MUBP)

- MUBP indices/intervals for **16 proportions** based on binary variables in NSFG:
  - Wide range of biserial correlations: .16 (income) to .82 (never been married)
  - MUBP intervals **significantly less wide** than the SMUB intervals (regardless of correlation)
    - Reflects tailoring of the MUBP index to the discrete nature of the binary Y
  - 10/16 estimated bias values were covered by the proposed intervals, representing an improvement over the SMUB approach (only 8/16)
- Simulation results for the MUBP indices similarly show improvements over the SMUB index for binary Y

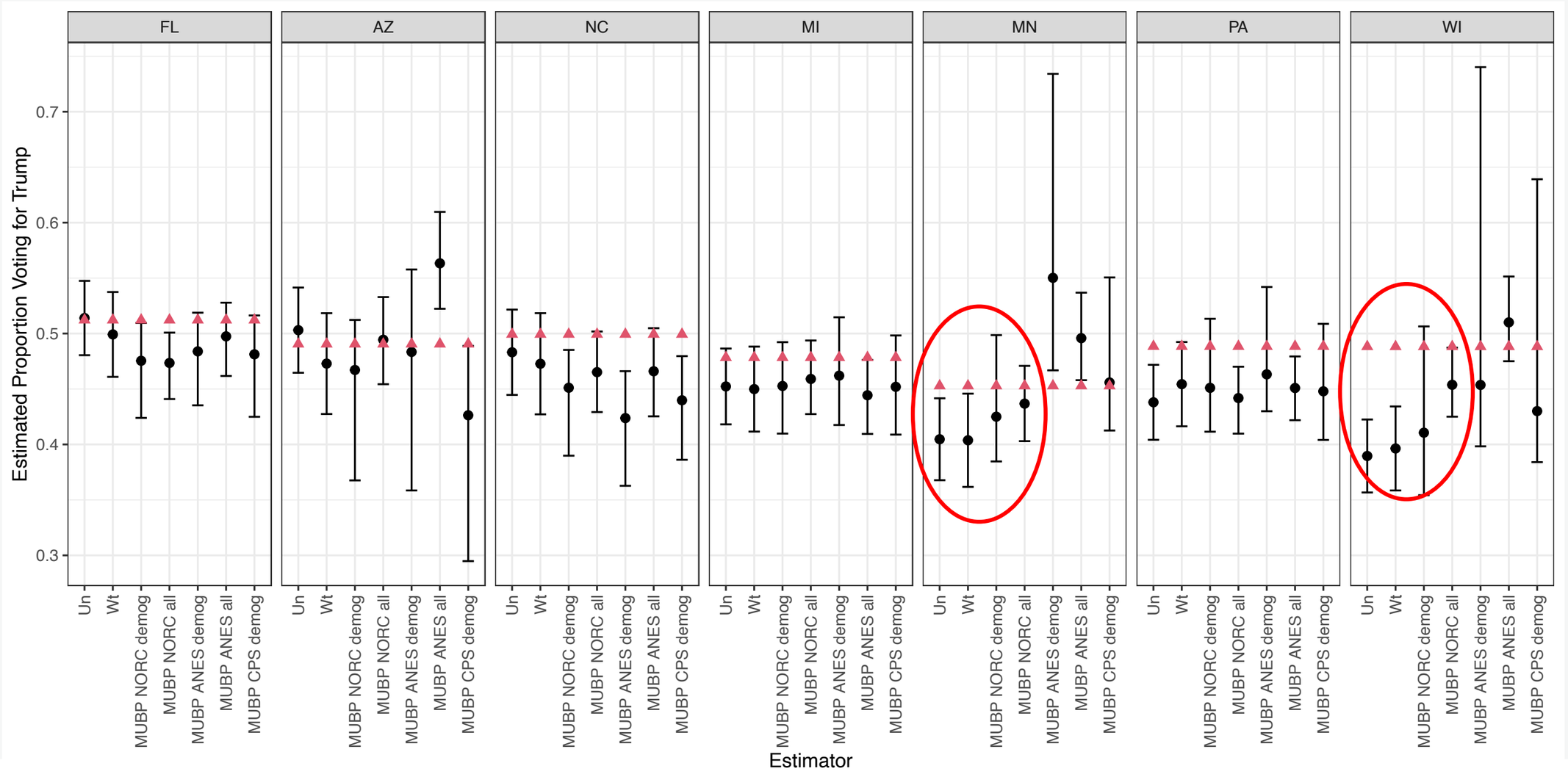
Current Work: Assessing Non-Ignorable Selection Bias in Pre-Election Polling Estimates

- We are currently working on applications of the MUBP measure using pre-election polling data from 18 different polls in the U.S. and U.K.
- Some of these data are now publicly available!
- **Approach:**
  - Find a good source of population data on likely voters, with predictors of voting for a given candidate that are also available in the poll data; **not trivial!**
  - Use the Bayesian approach to compute measures of potential non-ignorable selection bias for each poll, adjusted estimates (and adjusted credible intervals for the proportions), and comparisons of the adjusted estimates with direct weighted estimates (and their design-based SEs)

MUBP: Polling Application

- Paper is under second review at POQ
- Main challenge is finding the “right” population source (we are doing a post-hoc evaluation)
- You can walk through examples (including U.S. polls) in R: https://github.com/bradytwest/IndicesOfNISB
- Both the SMUB and MUBP measures could be computed in real-time for selected estimates, providing a sense of the amount of selection bias in the estimates at a **particular point** in a data collection
  - Useful for responsive survey design strategies and prioritizing particular cases expected to reduce the estimated bias!

Selected Results from Polling Application



Bonus Application*: MUBP for Estimating COVID Vaccine Uptake

Nature | Vol 600 | 23/30 December 2021 | 695

Article

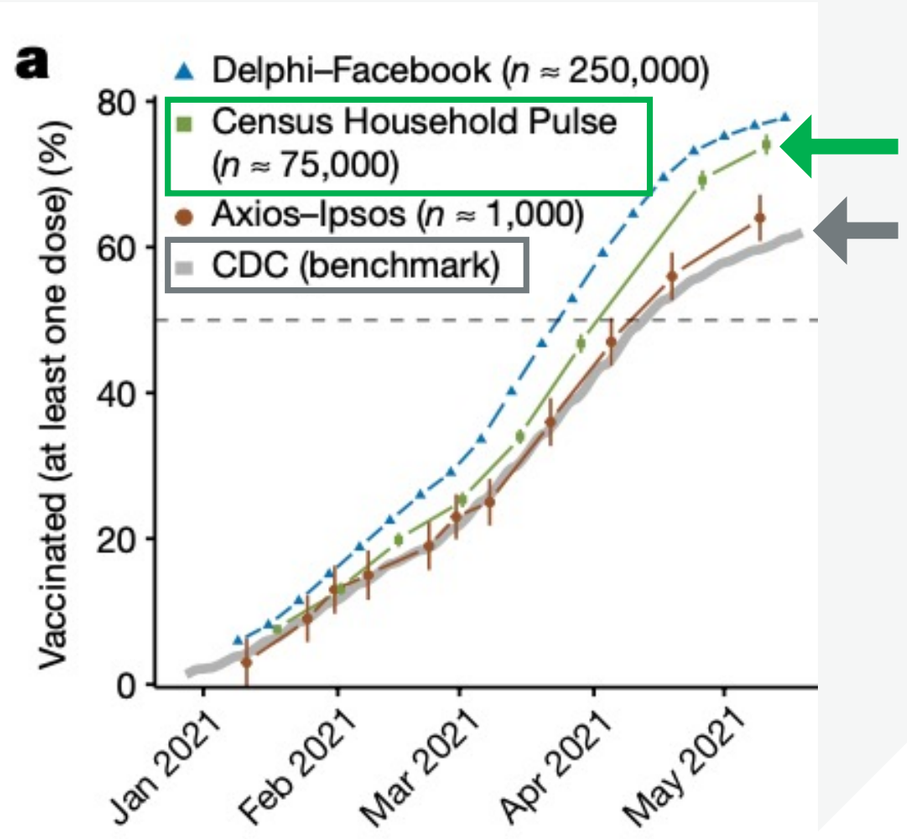
Unrepresentative big surveys significantly overestimated US vaccine uptake

https://doi.org/10.1038/s41586-021-04198-4

Received: 18 June 2021

Valerie C. Bradley[1,6], Shiro Kuriwaki[2,6], Michael Isakov[3], Dino Sejdinovic[1], Xiao-Li Meng[4] & Seth Flaxman[5]✉

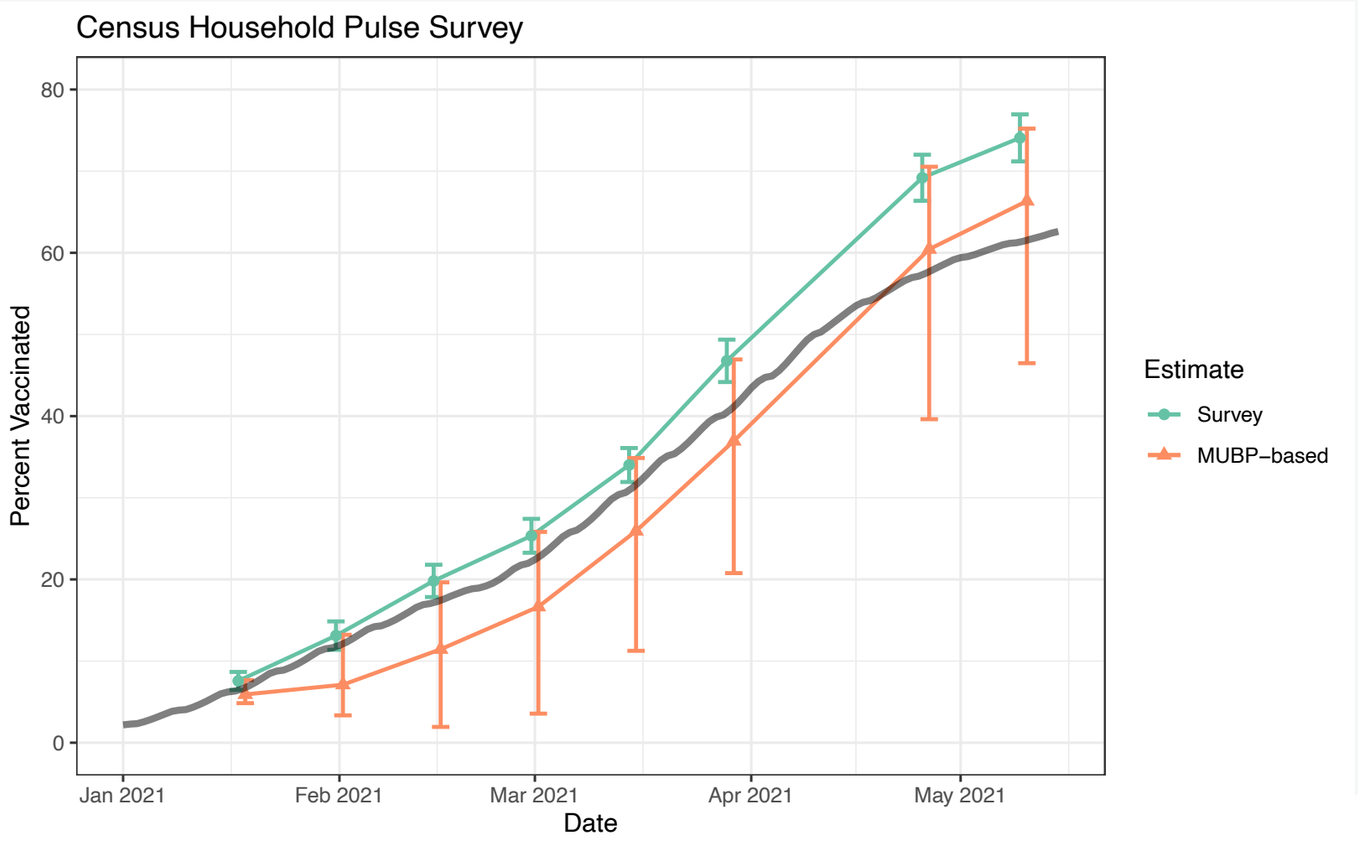
- Could MUBP-based analysis have "predicted" at least the direction of this (selection) bias?



Bonus Application*: MUBP for Estimating COVID Vaccine Uptake

- **Yes!** (with admittedly wide CIs)

- Easier to get aggregate Z for population than in polling application
  - Auxiliary variables: gender, education, race, ethnicity, age, income
  - Means/covariances from ACS
- Biserial correlations range .28 to .53



Conclusions: Descriptive Estimates

- We propose simple model-based indices of potential non-ignorable selection bias for descriptive estimates based on non-probability samples (or low response rate probability samples)
- Indices are easy to compute and only require aggregate information on relevant covariates for the target population
- Proposed indices perform quite well when based on moderately informative auxiliary information (correlation > 0.4? or > 0.3?)
- R functions enabling all SMUB and MUBP computations reported in this presentation; these functions are available here:
  https://github.com/bradytwest/IndicesOfNISB

The Case of Regression Coefficients

- But wait – we've done more!
- Extended these ideas to model-based indices of selection bias for the coefficients in linear and probit regression models estimated from non-probability samples (West et al. 2021)
  - Indices are effective when informative auxiliary variables are available
    - Interested in coefficients from Y|Z
    - Auxiliary information = covariates that are predictive of Y after conditioning on Z
- Easy-to-use R code, with examples, can be found at our GitHub page: https://github.com/bradytwest/IndicesOfNISB
- Future work needs to consider models for other non-normal outcomes; thus far, we have only considered extensions for the case of probit regression models

Thoughts on Real-Time Monitoring

- All of the measures of selection bias outlined here are easy to compute given the necessary input information
- The measures would be straightforward to monitor during a data collection, to determine whether responsive and adaptive design activities are helping to reduce the bias in selected estimates computed based on the measured sample
- They offer the important advantage of allowing for non-ignorable selection bias in selected estimates
- As a result, post-data collection adjustments (for which these measures also can be used) may not need to be as extreme

Thank You! / Questions?

Please direct any and all inquiries to

Rebecca Andridge (andridge.1@osu.edu)

or*

Brady West (bwest@umich.edu)

**Depending on whether you're a Buckeyes or a Wolverines fan, I suppose*

References

- Andridge, R.R. and Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 2, 153-180.
- Andridge, R.R., & Little, R.J. (2020). Proxy pattern-mixture analysis for a binary variable subject to nonresponse. *Journal of Official Statistics*, 36(3), 703-728.
- Andridge, R.R., West, B.T., Little, R.J.A., Boonstra, P., and Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society (Series C)*, 68(5), 1465-1483.
- Biemer, P. and Peytchev, A. (2011). A standardized indicator of survey nonresponse bias based on effect size. Paper presented at the International Workshop on Household Survey Nonresponse, Bilbao, Spain, September 5, 2011.
- Boonstra, P.S., Andridge, R.R., West, B.T., Little, R.J.A., and Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of Official Statistics*, 37(3), 751-769.
- Couper, M.P., Gremel, G., Axinn, W. G., Guyer, H., Wagner, J., and West, B.T. (2018), New Options for National Population Surveys: The Implications of Internet and Smartphone Coverage, *Social Science Research*, available at https://www.sciencedirect.com/science/article/pii/S0049089X17307871
- Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249-264.

References, cont'd

- Little, R.J. and Rubin, D.B. (2019). *Statistical Analysis with Missing Data, 3rd edition*. New York: Wiley.
- Little, R.J.A., West, B.T., Boonstra, P., and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 932-964.
- Nishimura, R., Wagner, J., and Elliott, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *International Statistical Review*, 84(1), 43-62.
- Särndal, C.-E., and S. Lundström (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36, 131–144.
- Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the Representativeness of Survey Response. *Survey Methodology*, 35(1), 101-113.
- West B.T. and Little R.J.A. (2013). Nonresponse adjustment of survey estimates based on auxiliary variables subject to error. *Journal of the Royal Statistical Society, Series C*, 62(2), 213-231.
- West, B.T., Little, R.J.A., Andridge, R.R., Boonstra, P., Ware, E.B., Pandit, A., and Alvarado-Leiton, F. (2021). Assessing Selection Bias in Regression Coefficients Estimated from Nonprobability Samples with Applications to Genetics and Demographic Surveys. *Annals of Applied Statistics*, 15(3), 1556-1581.