Data Integration in Survey Research:

Possible Approaches to Addressing Future Challenges

Joe Sakshaug

FedCASIC

16th April 2024

Data Integration

- The last decades have seen a growing interest in integrating surveys with alternative data sources
  - E.g. administrative, commercial, social media, digital trace data, etc..

- Basic idea: Use the strengths of one data source to offset limitations of the other

- Purposes of integration
  - Methodological
    - Assist with stratification, responsive survey design, investigating and correcting for nonresponse and measurement error
  - Substantive
    - Enhance substantive capabilities
    - Address complex research questions difficult to answer using single data source
  - Reduce costs / increase efficiencies


ISSN 0033-362X (Print)
ISSN 1537-5331 (Online)

Volume 87 • Special Issue • 2023
https://academic.oup.com/poq

Public Opinion Quarterly

SPECIAL ISSUE: AUGMENTING SURVEYS WITH PARADATA, ADMINISTRATIVE DATA, AND CONTEXTUAL DATA
Joseph W. Sakshaug and Bella Struminskaya, Editors

INTRODUCTION

Augmenting Surveys with Paradata, Administrative Data, and Contextual Data
Joseph W. Sakshaug and Bella Struminskaya

ARTICLES

Factors Associated with Interviewers' Evaluations of Respondents' Performance in Telephone Interviews: Behavior, Response Quality Indicators, and Characteristics of Respondents and Interviewers
Dana Garbarski, Jennifer Dykema, Nora Cate Schaeffer, Cameron P. Jones, Tiffany S. Neman, and Dorothy Farrar Edwards

How to Detect and Influence Looking Up Answers to Political Knowledge Questions in Web Surveys
Tobias Gummer, Tanja Kunz, Tobias Rettig, and Jan Karem Höhne

Income Source Confusion Using the SILC
Christopher Robert Bollinger and Iva Valentinova Tasseva

Evaluating Pre-election Polling Estimates Using a New Measure of Non-ignorable Selection Bias
Brady T. West and Rebecca R. Andridge

Continued on back cover

AAPOR
AMERICAN ASSOCIATION FOR
PUBLIC OPINION RESEARCH

OXFORD
UNIVERSITY PRESS


ISSN 2325-0984 (Print)
ISSN 2325-0992 (Online)

Volume 11 • Number 3 • June 2023
academic.oup.com/jssam

Journal of Survey Statistics and Methodology

SPECIAL ISSUE: RECENT ADVANCES IN DATA INTEGRATION

INTRODUCTION

Recent Advances in Data Integration
Joseph W. Sakshaug and Rebecca C. Steorts

SURVEY METHODOLOGY

Experiments on Multiple Requests for Consent to Data Linkage in Surveys
Sandra Walzenbach, Jonathan Burton, Mick P. Couper, Thomas F. Crossley, and Annette Jäckle

Augmenting Survey Data with Digital Trace Data: Is There a Threat to Panel Retention?
Mark Trappmann, Georg-Christoph Haas, Sonja Malich, Florian Keusch, Sebastian Bähr, Frauke Kreuter, and Stefan Schwarz

SURVEY STATISTICS

A Primer on the Data Cleaning Pipeline
Rebecca C. Steorts

Bayesian Graphical Entity Resolution using Exchangeable Random Partition Priors
Neil G. Marchant, Benjamin I. P. Rubinstein, and Rebecca C. Steorts

Continued on back cover

OXFORD
UNIVERSITY PRESS

JSSAM Special Issue: Overview of Topics

- Presenting multiple data linkage consent requests in online surveys
  - Walzenbach et al. (2023)
- Effects of linkage requests to mobile sensor data on panel retention
  - Trappmann et al. (2023)
- The data-cleaning pipeline
  - Steorts (2023)
- Entity resolution / correcting for linkage biases
  - Marchant et al. (2023); Patki and Shapiro (2023)

JSSAM Special Issue (cont.)

- Data fusion methods – relaxing the conditional independence assumption
  - Moretti and Shlomo (2023); Emmenegger et al. (2023)
- Linking WIC administrative records with the ACS
  - McBride et al (2023)
- Combining CDC vaccination data with inter-decennial population data to produce national and state-level estimates of vaccination rates
  - Raghunathan et al. (2023)
- Combining the UK Labour Force Survey with the Living Costs and Food Survey to improve the precision of estimates
  - Merkouris et al. (2023)

POQ Special Issue: Overview of Topics

- Using interviewers' evaluations of respondents' performance to study the respondents' behaviors and response quality
  - Garbarski et al. (2023)
- Using client-side paradata to examine the issue of respodnents looking up answers to political knowledge questions in web surveys
  - Gummer et al. (2023)
- Leveraging linked administrative data to examine misreporting in benefit programs and earnings
  - Bollinger and Tasseva (2023)
- Evaluating non-ignorable selection bias in pre-election polling estimates using aggregate data
  - West and Andridge (2023)

POQ Special Issue (cont.)

- Investigating attitudes toward privacy in relation to mouse-tracking paradata collection
  - Henninger et al. (2023)
- Research ethics and challenges of augmenting surveys with alternative data sources
  - Struminskaya and Sakshaug (2023)

Aims (and Challenges) of Data Integration

- Linkage consent
  - Ensuring informed consent
  - Maximizing consent rates

- Improving survey representativeness
  - Nonresponse bias evaluation
  - Enhancing NR bias adjustments

- Increasing estimation efficiency / cost savings
  - Supplementing probability sample surveys with non-probability information

Linkage Consent

Ensuring informed consent

Maximizing consent rates

Informed Consent

- Prior to linkage, respondent consent is usually required
  - In Germany, this is law (Federal Data Protection Act, 2013, Part I, Section 4; Code of Social Law X, 2013, Section 75)

- The purpose of the consent process is to ensure respondents are informed about:
  - Which data sources will be linked
  - Intended uses of the linked data
  - Possible benefits (and risks, if any)
  - Responsibility of ensuring data confidentiality
  - Voluntary nature of request

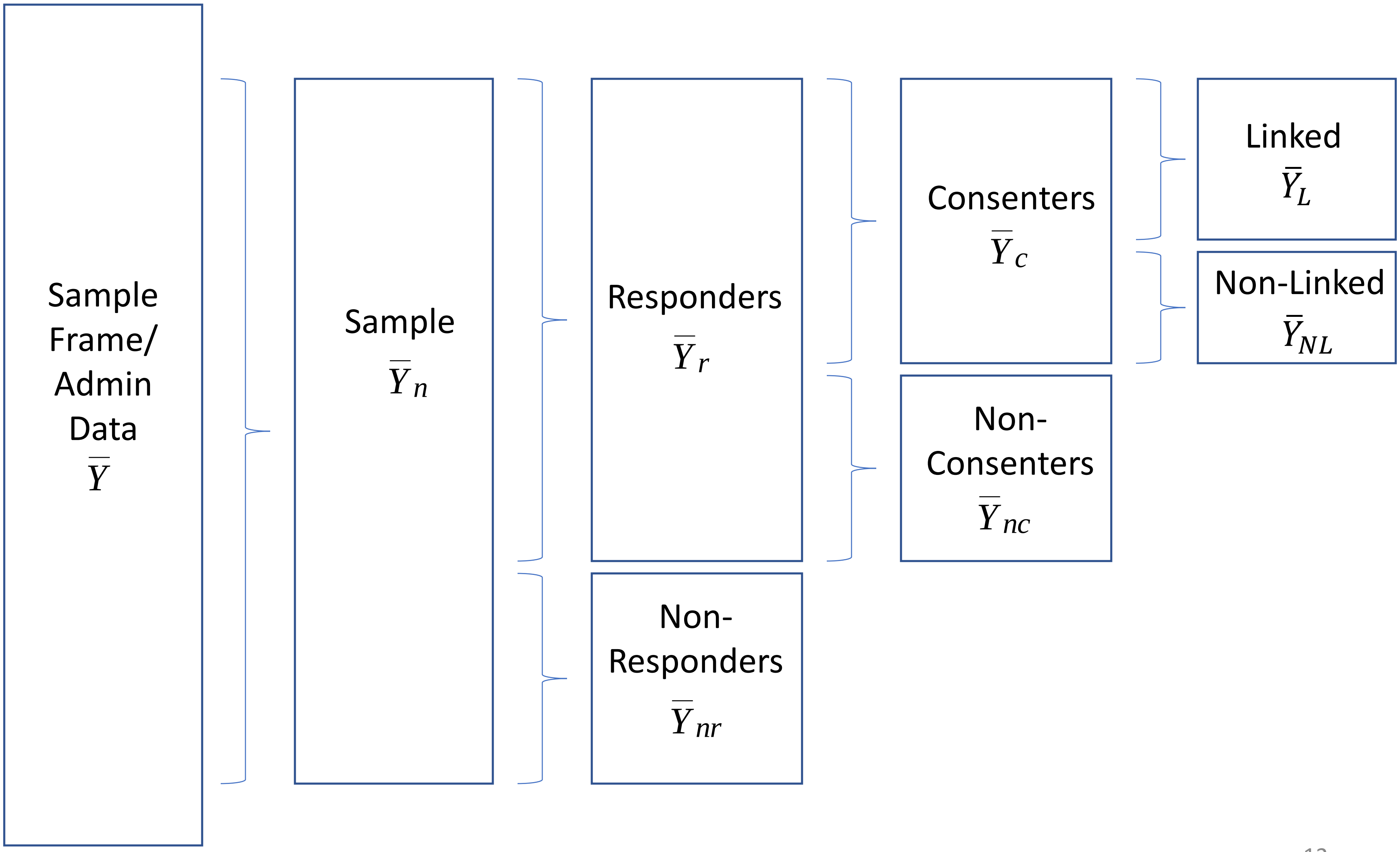
Linkage Consent Rates

- Consent rates vary from study-to-study
  - Range: 39 to 97 percent (da Silva et al. 2012)
  - Range: 24 to 89 percent (Sakshaug and Kreuter, 2012)

- Some evidence that consent rates were decreasing (in the U.S.)
  - National Health Interview Survey (1993-2005): 85 to 50%
  - Survey of Income and Program Participation (1996-2004): 88 to 65%
  - Current Population Survey (1994-2003): 90 to 76%

- Concern: non-consent error
  - Reduction in analytic sample size, increased variance estimates
  - Respondents who consent to linkage may be systematically different from those who don't
    - Many studies show this to be the case

Conceptual Pathway to Linkage



Bias in Survey Estimates

- Consumer Expenditure Quarterly Interview Survey (CEQ)

	Respondent Mean	Consenter Mean	Difference
Family income	$50,939.00	$52,869	**$1,930.00****
Vehicle cost	$599.59	$619.14	$19.55
Property taxes	$454.15	$429.12	**-$25.02****
Property value	$247,216.00	$243,507.00	-$3,709.00
Rental value	$1,378.03	$1,351.92	**-$26.11****

Yang, Fricker, and Eltinge, 2015

Bias in Administrative Estimates

- IAB PASS Study (welfare recipient sample)

Variable	Nonresponse Bias	Measurement Bias	Linkage Consent bias
Age	0.1	0.03	-0.3*
Foreign citizen (%)	-5.6*	-2.5*	-0.9*
Welfare receipt (%)	3.2*	-7.1*	-0.3
Disability (%)	0.4	6.0*	0.01
Employed (%)	1.0	-0.6	0.3
Income (30 days)	-71.4*	394.5*	1.7

- Non-consent bias is present, but relatively small compared to other error sources

Sakshaug and Kreuter, 2012

Optimizing Linkage Consent

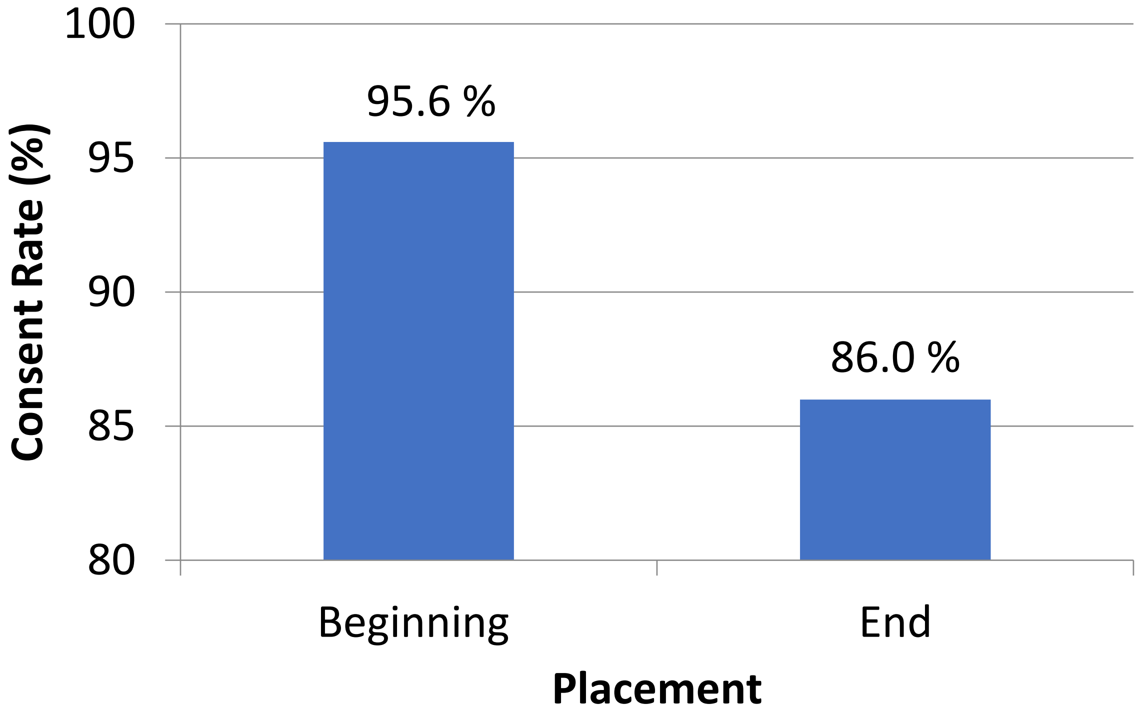
- Recent efforts have largely focused on methods of increasing the consent rate
  - **Placement**
  - **Wording/framing**
  - Re-asking for consent among prior refusers
  - Active vs. passive consent

Placement of Consent Question

- Historically linkage consent question has been asked at the end of interview
- *Conventional wisdom* is that interviewer-respondent rapport reaches peak at the end
  - However, relationship between rapport and linkage consent is mixed
    - Jenkins et al. (2006): positive effect
    - Sala et al. (2012): negative effect
- Experimental evidence suggests end-placement is suboptimal compared to:
  - Asking in the context of topic-related items (Sala, Knies, and Burton, 2014);
  - Asking at the beginning of the interview (Sakshaug, Tutz, and Kreuter, 2013)

Placement in a Household Survey

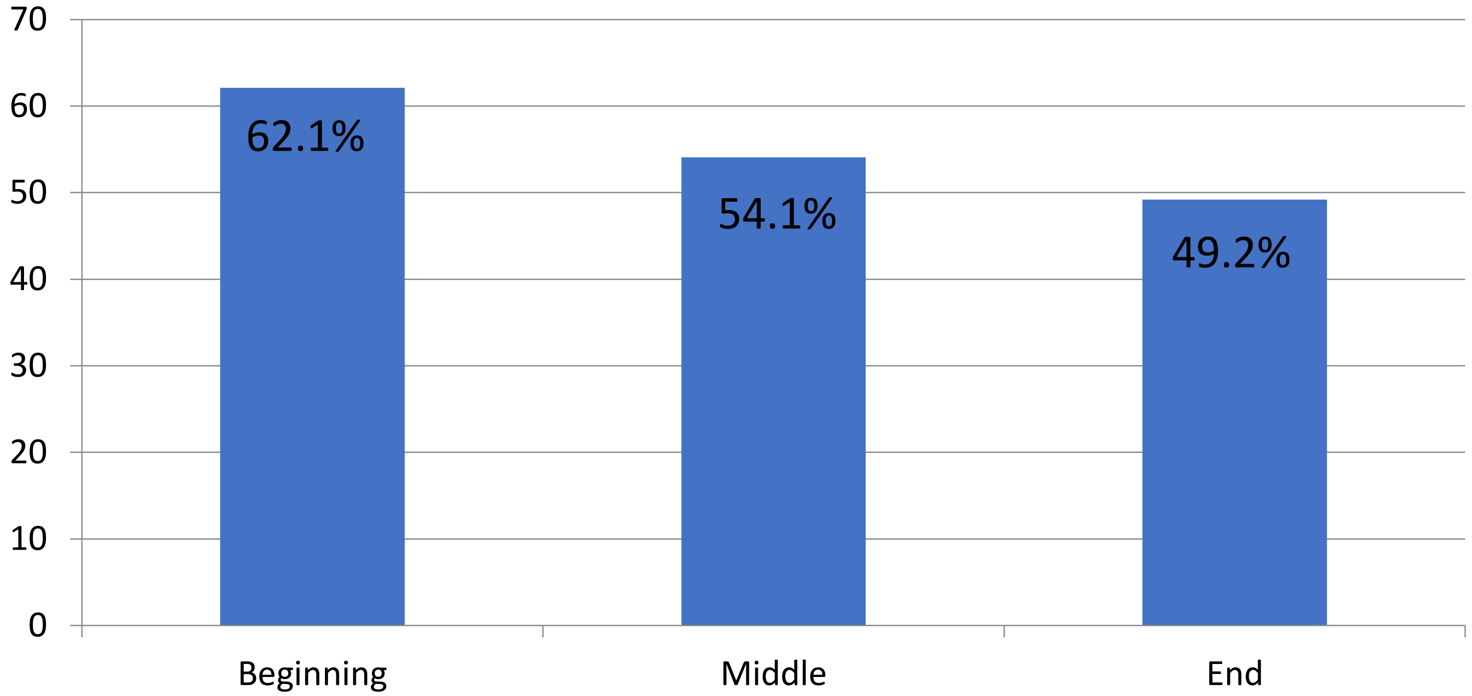
- N = 2,400 telephone interviews in Germany



Sakshaug, Tutz, and Kreuter, 2013

Placement in an Establishment Survey

- N = 4,222 responding establishments in Germany



Sakshaug and Vicari, 2018

Wording of the Consent Question

- Surveys have some flexibility in scripting the consent question
  - Exact wording varies across studies

- Often the **benefits** of linkage are emphasized to respondents
  - E.g., saves time, reduces costs and burden, improves data accuracy

- However, empirical support for this strategy is mixed
  - No effect on consent rates (Pascale, 2011; Sakshaug, Tutz, and Kreuter 2013)
    - Telephone survey
  - Positive effect of time-saving argument (Sakshaug and Kreuter, 2014)
    - Web survey

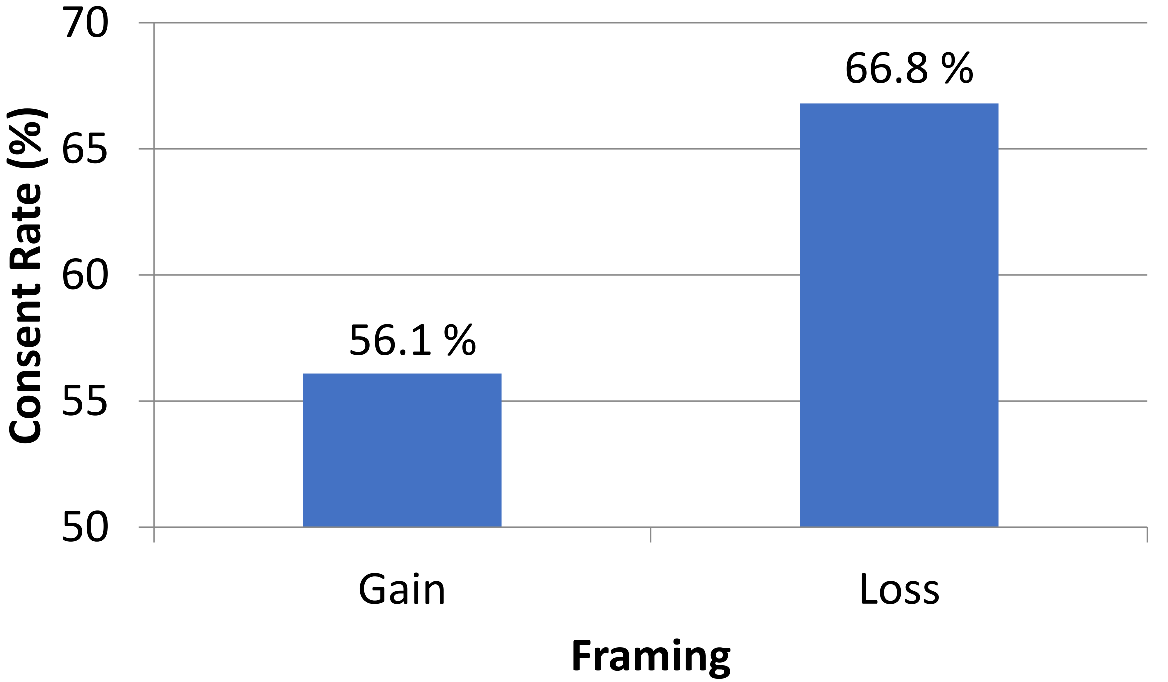
Loss Framing

- Instead of emphasizing the positive benefits of linkage, emphasize the *negative* consequences of not linking one's data
  - Based on the tenets of *Prospect Theory* (Kahneman and Tversky, 1979; 1984)

- <u>Gain frame: </u>"The information you have provided so far would be *a lot more valuable* to us if we could link it to…"

- <u>Loss frame:</u> The information you have provided so far would be *much less valuable* to us if we can't link it to…"

Gain-Loss Framing Experiment



- Respondents in the *loss framing group* were more likely to consent than those in the gain framing group

Kreuter, Sakshaug, and Tourangeau, 2015

Interaction: Placement vs. Framing

Phone	Beginning	End	Total n
Gain	90.8	78.7	598
Loss	90.5	81.2	610
Total n	613	595	1208

Web	Beginning	End	Total
Gain	82.6	62.4	520
Loss	86.3	75.4	489
Total	511	498	1009

Kreuter et al., 2015

Consent Understanding

- “Informed consent” implies that respondents are well-informed about the linkage process

- How much of the linkage consent process is understood by respondents?

- Are less informed respondents less likely to consent than those who are more informed?

Consent Understanding: IAB Study

- Percent answered correctly by linkage consent

	Consenters % correct	N	Non-consenters % correct	N
Answers send to IAB	88.3	977	57.8	142
Merged with IAB	93.3	982	36.7	147
Name/Adress saved	68.3	981	38.8	147
Result lead to you	63.4	995	--	--
IAB only access	85.6	998	--	--
Public access to identifiabled data	87.5	1009	--	--

Kreuter et al., 2015

Improving Survey Representativeness

Nonresponse bias evaluation

Enhancing NR bias adjustments

Data Integration for Reducing Nonresponse Error

- Nonresponse poses risks to survey inference

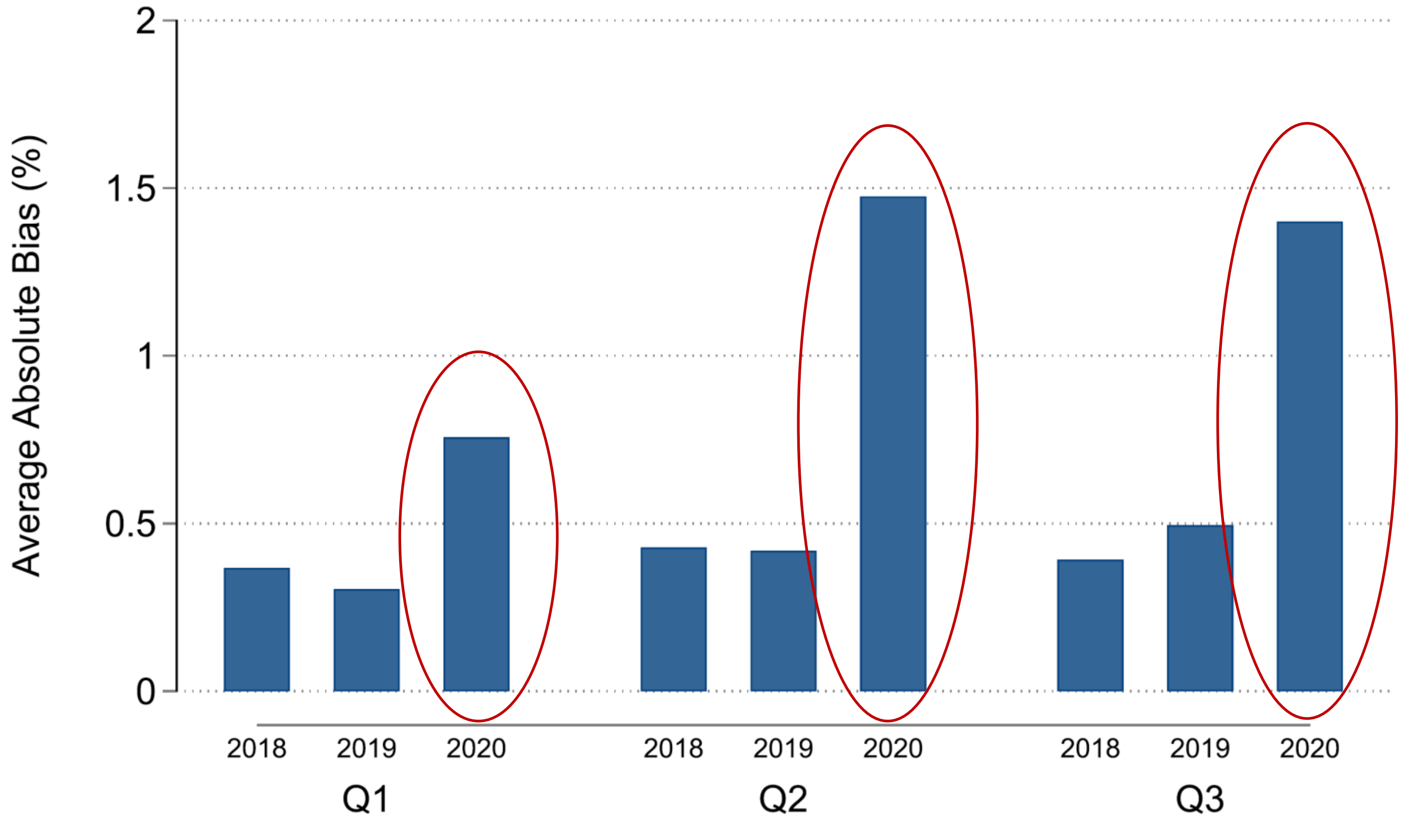
- Nonresponse likely related to substantive phenomena → bias
  - industry, estab size, employment status, job change, life events

- Available auxiliary data (e.g. paradata) may be limited for bias adjustment

- *Administrative data* offers viable source of auxiliary data
  - correlated with substantive variables of interest

- Recent work: Incorporate administrative data into nonresponse adjustments
  - Adjusting for COVID-19-related nonresponse
  - Combining administrative data with machine learning methods

IAB-JVS: Quarterly Nonresponse Bias *Increased* During COVID-19 Pandemic

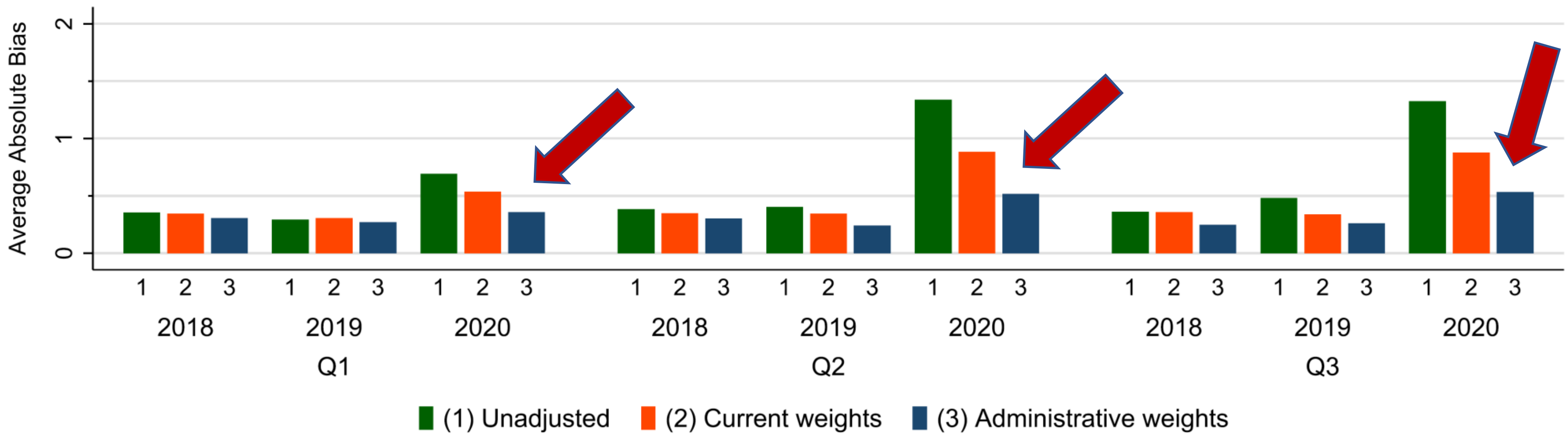


- Are standard weighting adjustments still effective?
    - Can augmenting with administrative data improve bias reduction?

Küfner, Sakshaug, and Zins (2022)

Comparing Current JVS Weighting Scheme vs. Enhanced Administrative Data Weighting Scheme

- Current JVS weighting scheme (propensity score estimation)
  - Only 3 covariates: industry, establishment size, paradata
- Enhanced administrative data weighting scheme
  - Additional 16 admin variables (establishment + employee characteristics)

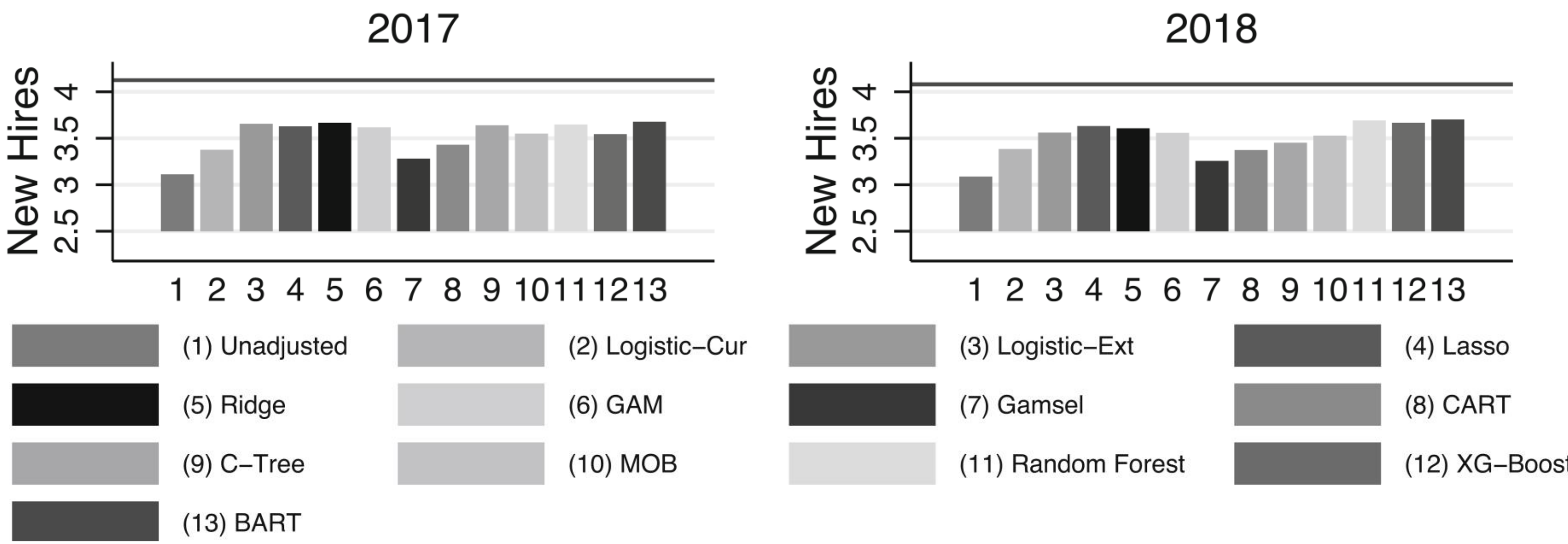


- Enhanced administrative data weights improve NR bias reduction

Küfner, Sakshaug, and Zins (2022)

Does Admin Data + Machine Learning Improve NR Adjustment?

- IAB-JVS: Mean number of new hires at t+1
- Current weighting scheme vs. Enhanced (admin) weighting vs. Enhanced (admin) + ML modeling of propensity scores



- Enhanced administrative data improves bias adjustment
  - ML methods do not provide much added value

Küfner, Sakshaug, and Zins (2022)

Increasing Estimation Efficiency / Cost Savings

Supplementing probability sample surveys with non-probability sample information

The context

Problem

A researcher is interested in making inferences from a probability sample (PS) survey but cannot afford a large sample size

Alternatives

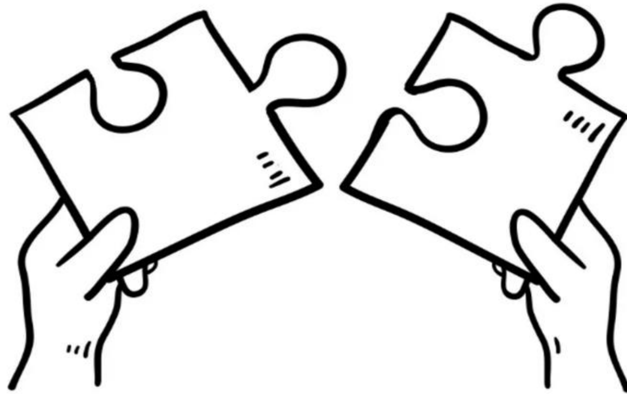
1. **Reduce the sample size**: small PS size → large variance but "unbiased" estimates
2. **Opt for a non-probability sample (NPS) survey**: biased but low variance

The proposal

The data integration puzzle

Small size to reduce costs

PS (high quality)
Unbiased
Large variance



NPS (lower quality)
Selection Bias
Lower variance

Basic Idea

The data integration perspective

- **Field** small PS survey + larger NPS survey in parallel with the same variables
- Integrate both surveys under Bayesian framework to improve inference on **regression coefficients and reduce survey costs**

Inference

- Based on **small PS data** (unbiased, high variance)
- **Incorporation** of (possibly) **biased NPS data** into the estimation process (low variance)
- Posterior estimates are likely to have more bias than PS estimates but possibly less variance (**bias/var trade off)**

Two aims

1. **Enhance inference (MSE)**
    - **Baseline situation:** analysis of small PS only (gold standard)
    - **Data Integration:** can we reduce MSE with respect to the baseline situation?

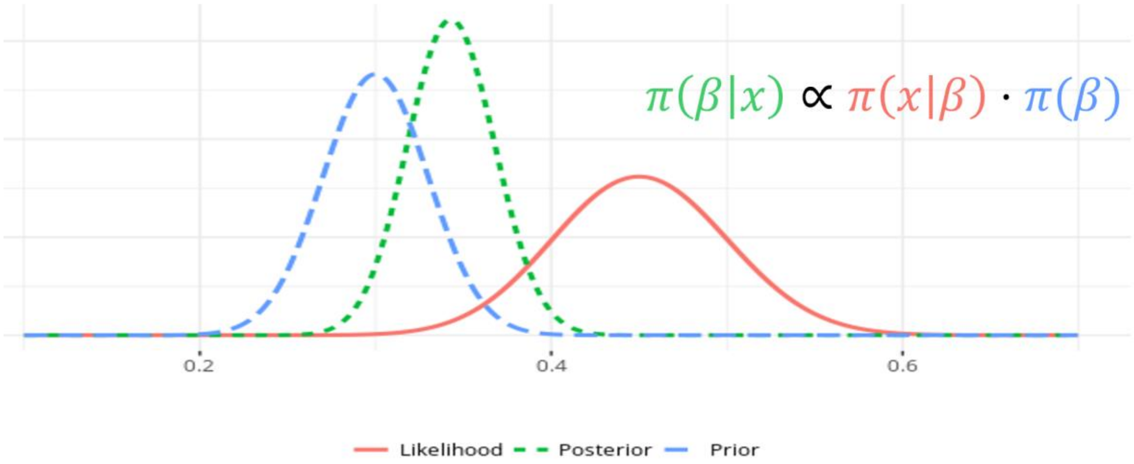
2. **Reduce survey costs**
    - Can we obtain at a **lower cost** the same **MSE** that we would obtain analyzing a much **larger** and **costlier PS-only survey**?

Why Bayesian? (Kruschke, 2014; Gelman et al., 2013)

- **Natural choice** to integrate data with varying levels of quality
- Its structure can be exploited in order to **incentivize high-quality** data

The prior is based on NPS data. How much should it influence the posterior inference?

We borrow information based on the **similarity** between **PS** and **NPS estimates**



Priors

Baseline (No data integration; PS data only)

- A weakly informative prior proposed by Gelman et al. (2008)
- **Baseline prior** against which compare data integration results

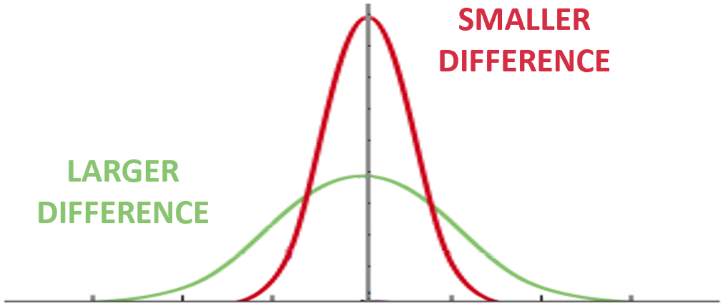
$$\beta_j \sim Student(\nu = 3, \mu = 0, s = 2.5) \quad \text{for j=0,1,2}$$

Informative priors: integrating PS and NPS data

Distance priors: The influence of the prior depends on the difference between ML estimates in both PS and NPS surveys

Example: the **basic distance prior**

$$\beta_j \sim \mathcal{N}\left(\widehat{\beta_{NP}}, \left|\widehat{\beta_P} - \widehat{\beta_{NP}}\right|\right)$$



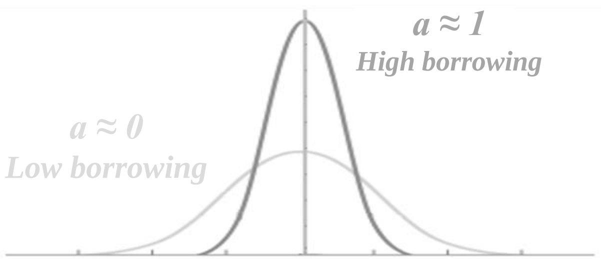
Mixed distance priors: Baseline prior for β_0 and distances priors for other coefficients

Informative priors: integrating PS and NPS data

Power prior (Ibrahim et al., 2000)

$$\underbrace{\pi(\beta, a|D_{NP})}_{\text{Power prior}} \propto \underbrace{L(\beta|D_{NP})^{a}}_{\text{Likelihood NPS}} \underbrace{\pi_0(\beta)}_{\text{Baseline prior}}$$

Likelihood NPS



How much do we borrow from NPS?

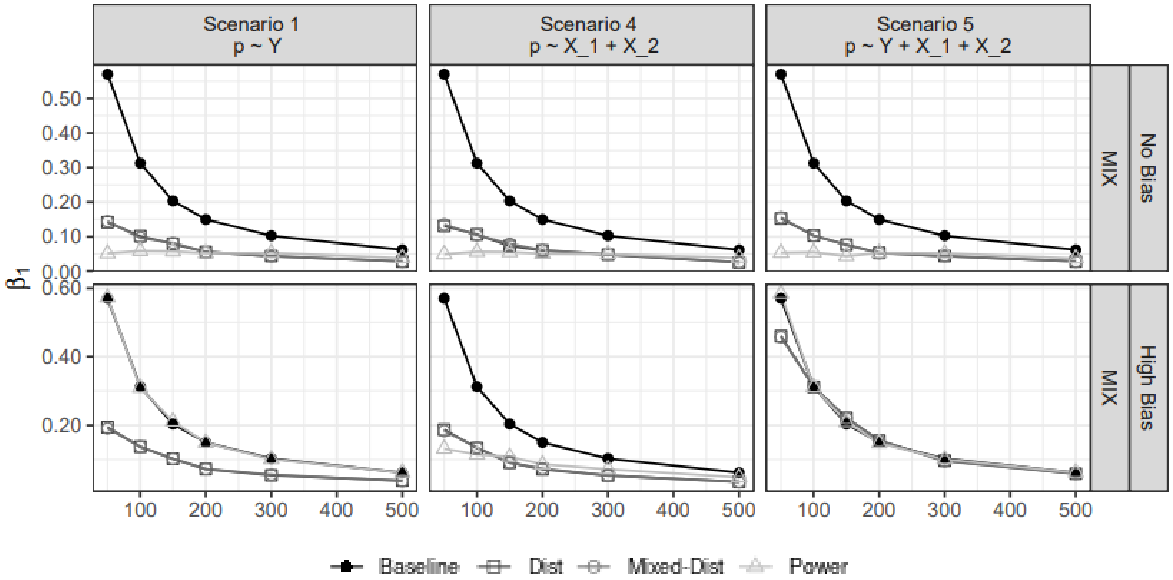
The power parameter "a":

1 = full borrowing

0 = no borrowing

- We select it dynamically based on the similarity between PS and NPS
- We are working on different measures but for now:
- It is the p-value of the Hotelling t-test for the difference betweer β_P and β_{NP})

MSE Results: selected cases



Median MSE across 100 repetitions:

- Low selection bias and small PS: large improvements in MSE
- High selection bias: INF prior performs similarly to baseline prior

9

Application: American Trends Panel

PS data – American Trends Panel (ATP)

- Pew Research Center's nationally representative online survey panel
- Sample size: 3000 units → PS $\in$ (N=50, 100, 150, 200, 500)

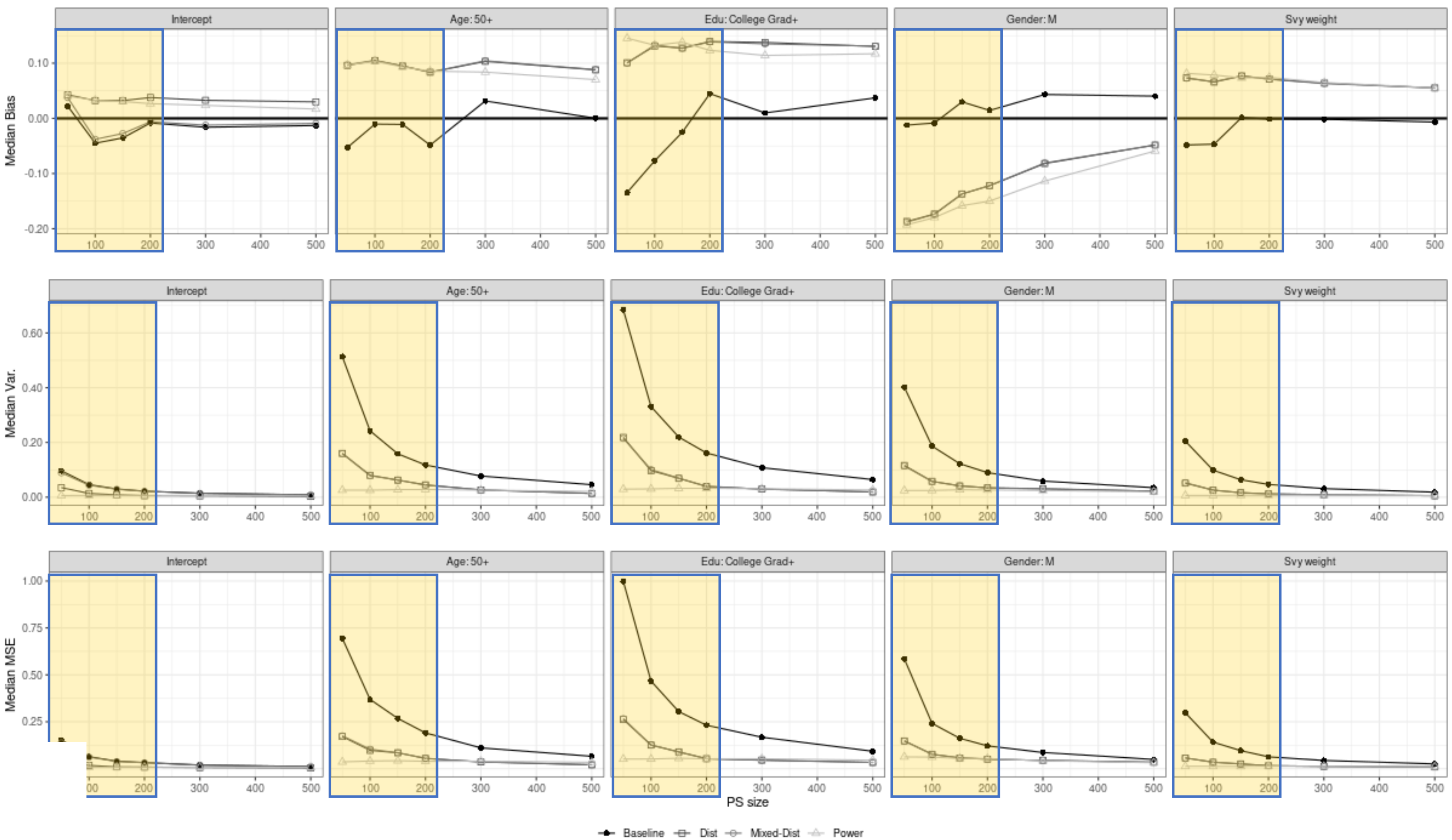
NPS data - 9 parallel online NPS from different vendors

- Vendors implemented quota sampling with different quota variables
- Sample size of about 1000 respondents

Outcome variables: Smoking, Always vote, Neighborhood Trust, Neighborhood Safety, Healthcare coverage, Volunteering

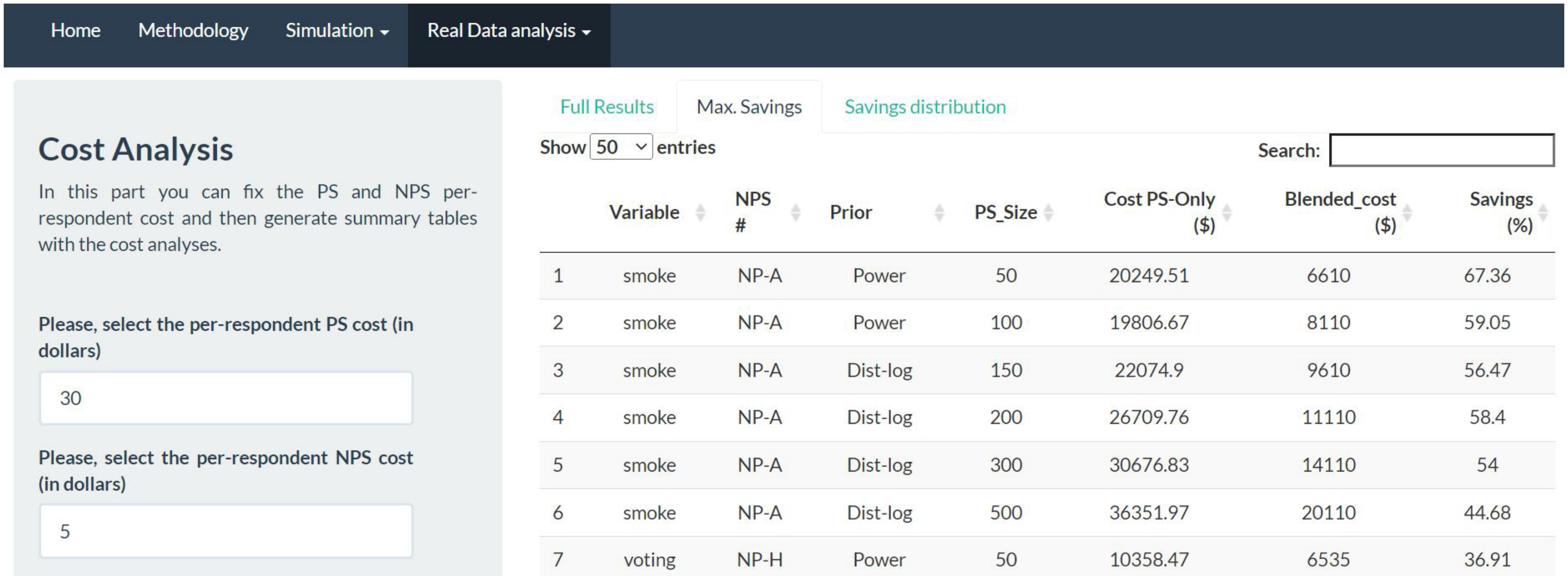
Covariates: Age, gender, education, survey weight

Results: Bias, Variance, MSE for *Current Smoking Status*



Reduction in MSE is mainly driven by a reduction in variance

Interactive Cost Analysis: Shiny App



	Variable	NPS #	Prior	PS_Size	Cost PS-Only ($)	Blended_cost ($)	Savings (%)
1	smoke	NP-A	Power	50	20249.51	6610	67.36
2	smoke	NP-A	Power	100	19806.67	8110	59.05
3	smoke	NP-A	Dist-log	150	22074.9	9610	56.47
4	smoke	NP-A	Dist-log	200	26709.76	11110	58.4
5	smoke	NP-A	Dist-log	300	30676.83	14110	54
6	smoke	NP-A	Dist-log	500	36351.97	20110	44.68
7	voting	NP-H	Power	50	10358.47	6535	36.91

- Cost savings of up to 67% achieved for some priors

Salvatore et al. 2024

Conclusions

- Growing interest in methods and applications of data integration for both survey methodological and substantive research
  - More special issues forthcoming (JOS, JRSS-A)
- Obtaining consent from respondents is important from legal and ethical standpoint
  - Challenge lies in ensuring respondents are sufficiently informed about linkage process
- Harnessing the full potential of administrative covariate information may improve upon current NR adjustments
- The combination of probability and (less expensive) non-probability samples can improve estimation efficiency and reduce costs

Thank you for your attention

Slides and references available upon request