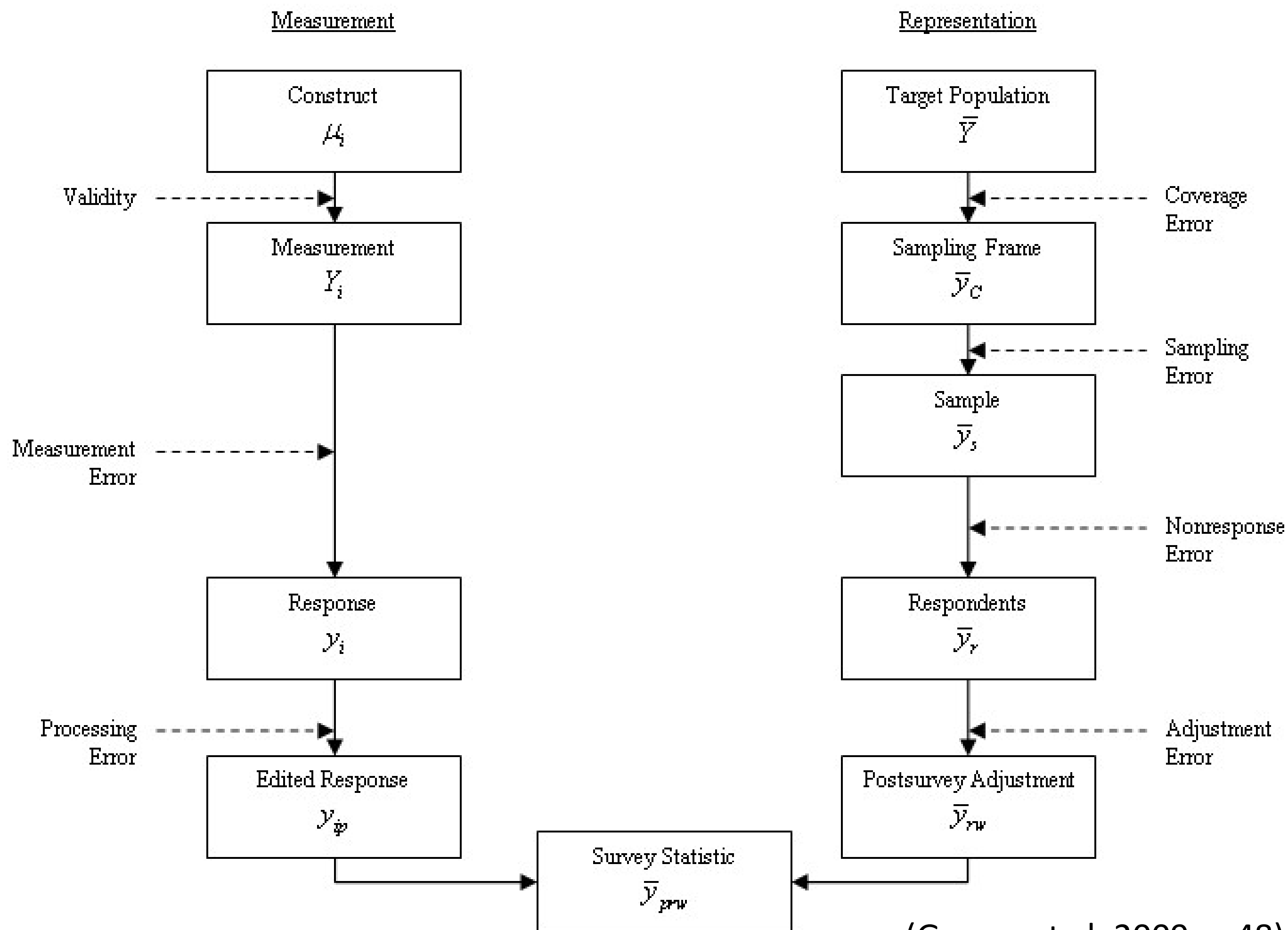
# Survey data analysis – week 14
# Data integration

Peter Lugtig – p.lugtig@uu.nl

# TSE is focused on error (accuracy + precision)




(Groves et al. 2009, p.48)

# Data quality framework of Biemer and Lyberg (2003)

- Data of high quality has....
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

*Quality is "fitness for use"*

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley & Sons.

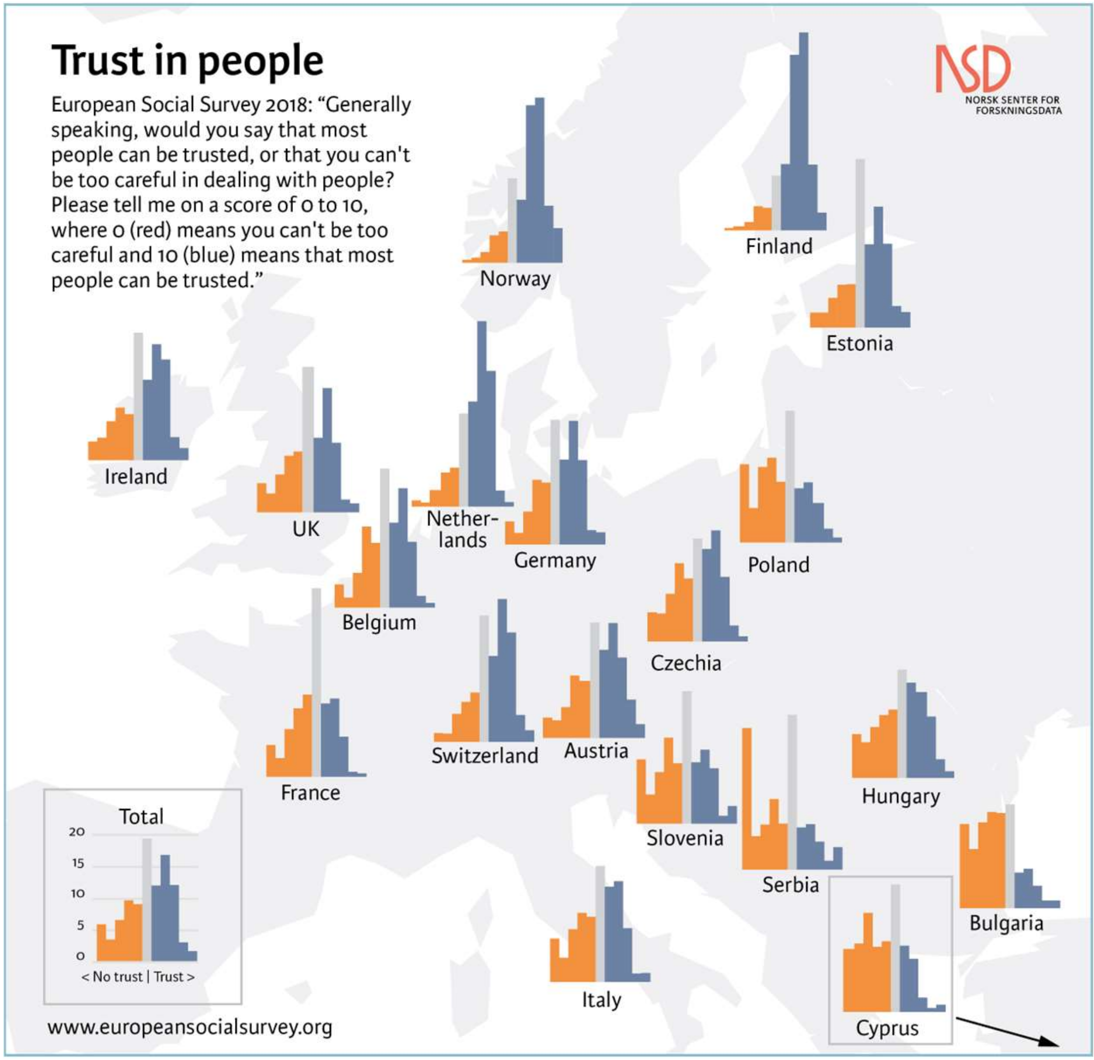
# Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

*The degree of confidence that users place in data products based on their image of the data provider.*

# Data quality framework

- Data of high quality has....
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



# Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

*There is clear data documentation (metadata) so that we understand what data is about*

B17 fx =COUNTIF(B4:B16,"yes")

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Human Trafficking Survey | | | | |
| 2 | | | | | |
| 3 | ID # | Q1 | Q2 | Q3 Routes | Q3 Origins |
| 4 | 101 | Yes | Reports | Yes | Yes |
| 5 | 102 | Yes | Reports | Yes | Yes |
| 6 | 103 | Yes | Advocacy | No | No |
| 7 | 104 | No | | | |
| 8 | 105 | Yes | Grants | Yes | Yes |
| 9 | 106 | Yes | Reports | Yes | Yes |
| 10 | 107 | Yes | Sharing | Yes | Yes |
| 11 | 108 | Yes | GIS | No | Yes |
| 12 | 109 | No | | | |
| 13 | 110 | Yes | Grants | Yes | No |
| 14 | 111 | Yes | Reports | Yes | Yes |
| 15 | 112 | Yes | Other | No | No |
| 16 | 113 | No | | | |
| 17 | | 10 | | | |

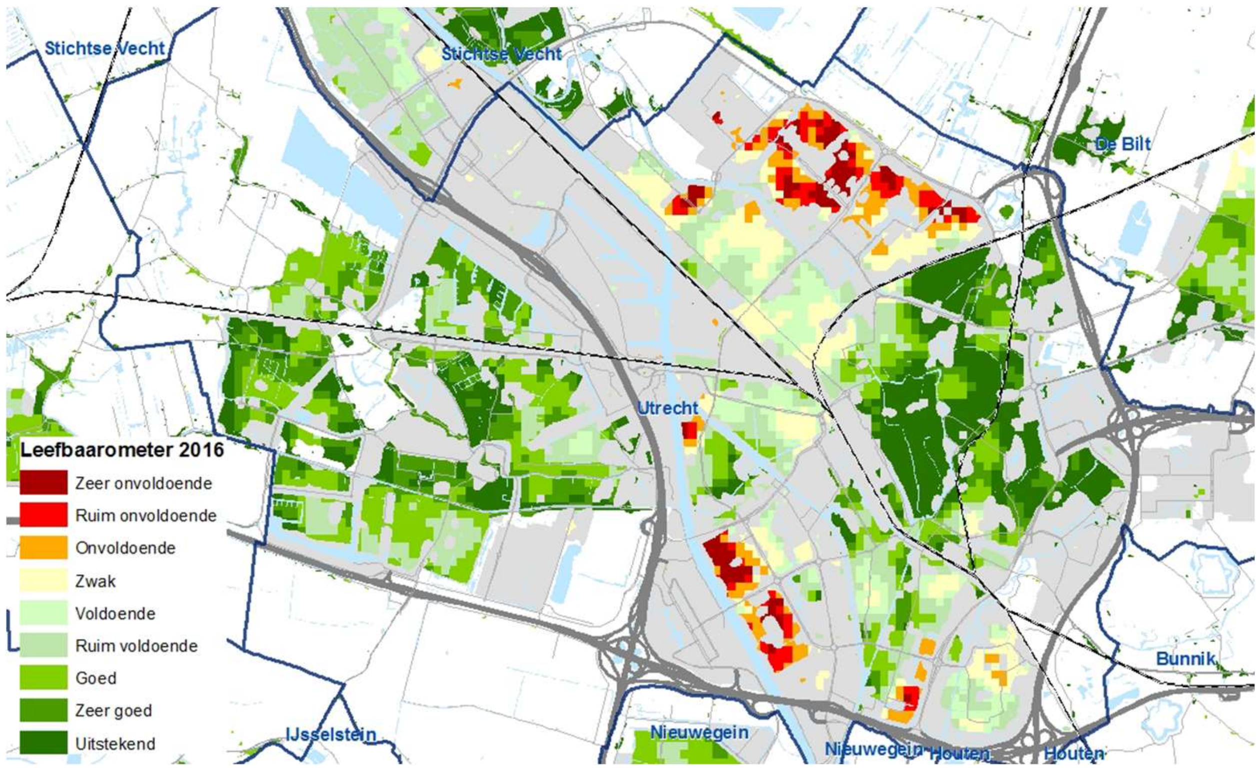
# Data quality framework

- Data of high quality has....
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence


*The Netherlands HealthRI data access process*

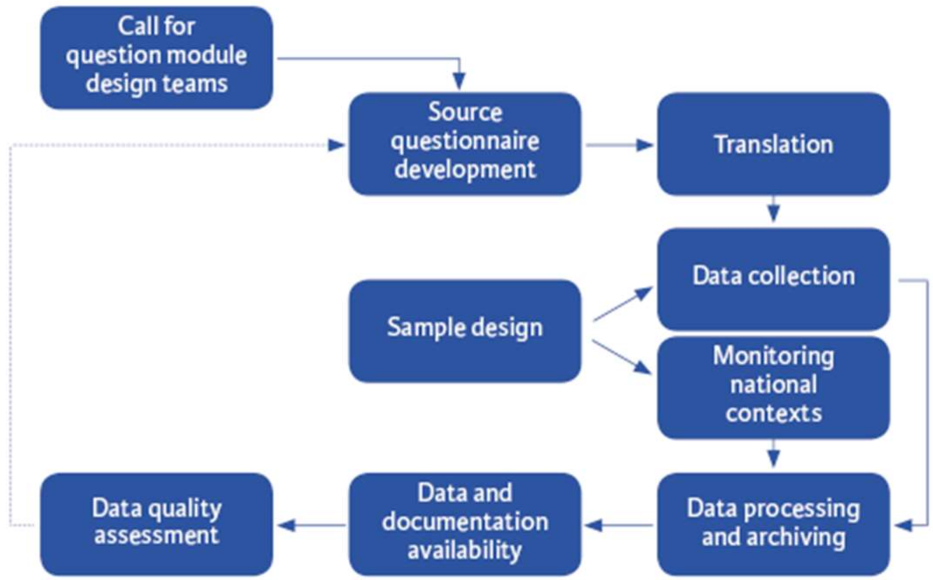
# Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



# Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



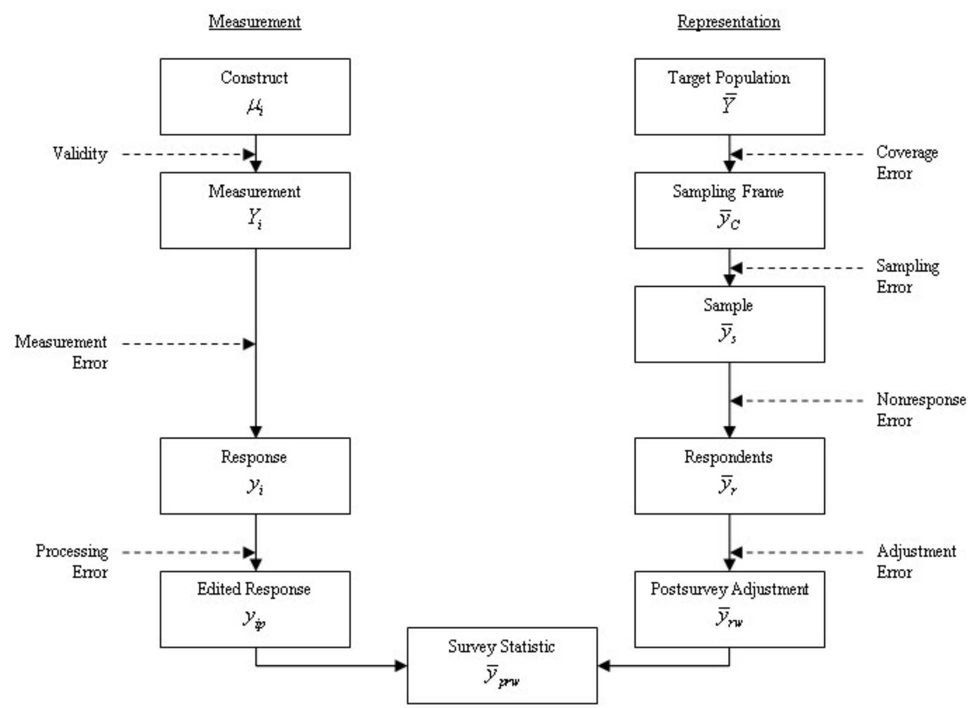

*European Social Survey methodology overview*

# Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

# Data quality framework

- Data of high quality has....
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence



# Data quality framework

- Data of high quality has….
  - Credibility
  - Comparability
  - Interpretability
  - Accessibility
  - Relevance
  - Timeliness
  - Completeness
  - Accuracy
  - Coherence

*Common definitions,classifications, and methodological standards (often over time)*

| Teen Internet Activities<br>*Do you ever...?* | *Online Teens (n=886)* |
|---|---|
| Go to websites about movies, TV shows, music groups, or sports stars | 81% |
| Get information about news and current events | 77 |
| Send or receive instant messages (IMs) | 68 |
| Watch video sharing site | 57 |
| Use an online social networking site like MySpace or Facebook | 55 |
| Get information about a college or university you are thinking of attending | 55 |
| Play computer or console games online | 49 |
| Buy things online, such as books, clothes, and music | 38 |
| Look for health, dieting, or physical fitness information | 28 |
| Download a podcast | 19 |
| Visit chatrooms | 18 |

*Source: Pew Internet & American Life Project Survey of Parents and Teens, October-November 2006. . Margin of error for teens is ±4%.*

# What has changed in the big data landscape?

- More sources
  - Administrative data
  - Survey data
  - Sensor data (phones, IoT)
  - Digital trace data
  - Organic (aka big) data
  - Non-prob surveys
- Each with it's own problems

- Not one methodology for how to do data integration
  - Approach is always statistic-specific

# The idea of data integration

**Work around a shortcoming of one source with another one**

- Credibility
- Comparability
- Interpretability
- Accessibility
- Relevance
- Timeliness
- Completeness
- Accuracy
- Coherence

- Multi-source statistics (de Waal, van Delden, Scholtus, 2020; or Zhang, 2012)

# Multi-source statistics

4 dimensions

1. Units (sample, population)
2. Measurement (same, different)
3. Time dimension (same, different)
4. Level of aggregation (micro, or macro)

- Can be combined

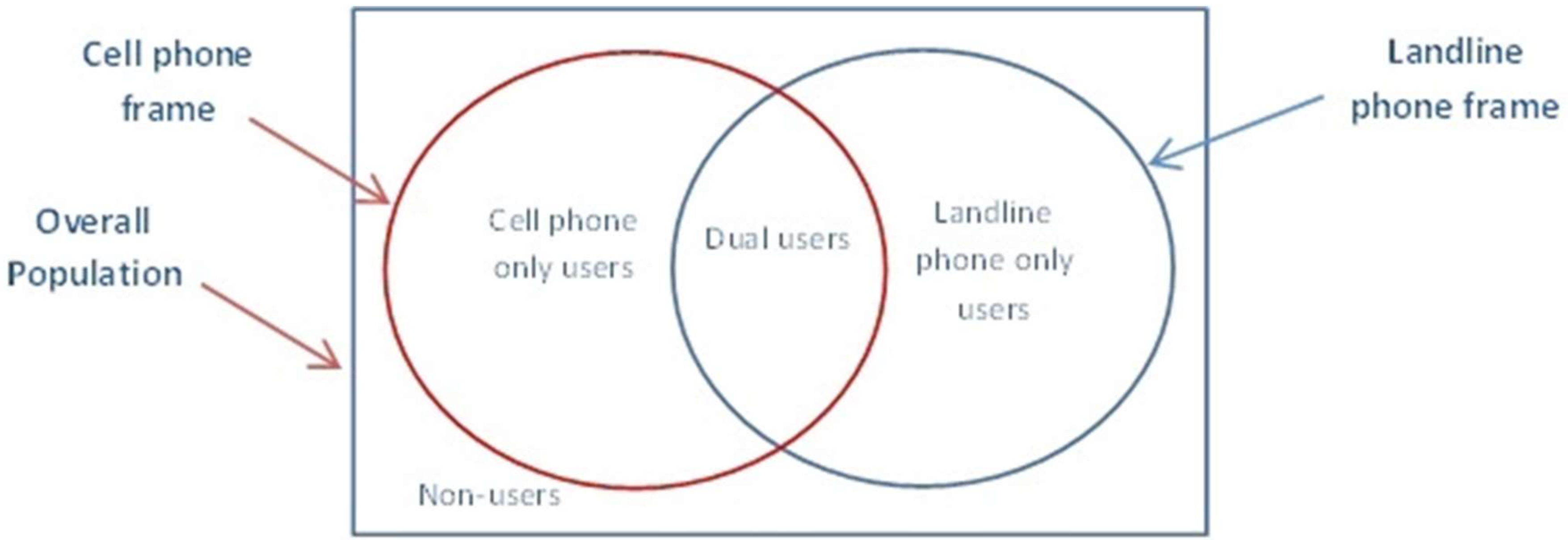
De Waal, T., van Delden, A., & Scholtus, S. (2020). Multi-source statistics: basic situations and methods. *International Statistical Review*, *88*(1), 203-228.

# A lot of examples of data integration

Question with every application:

What quality dimension are we trying to improve?

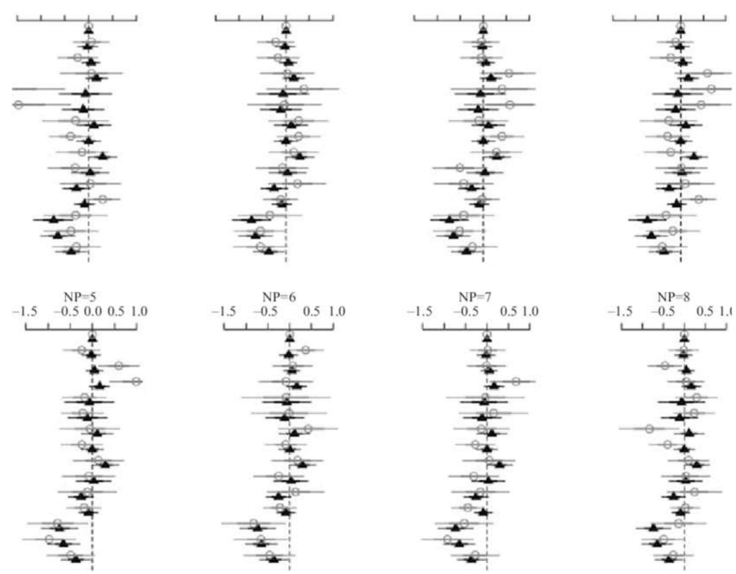
# Dimension 1:
# Multiple frames ( to cover population)




What quality dimension are we trying to improve?

# Dimensions 1,2:
# same measurements, different units

- Integrate smal probability based survey with
- Larger non-probability one
- Later guest lecture by Camilla Salvatore




What quality dimension are we trying to improve?

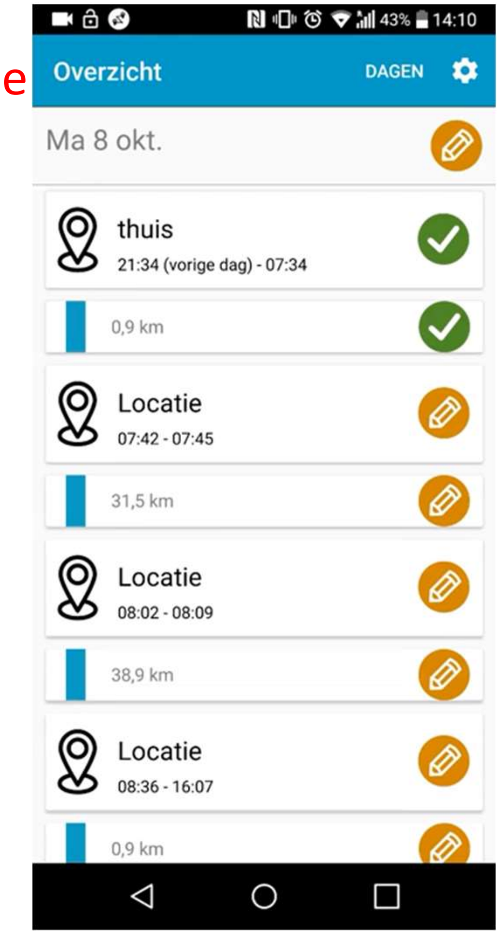
# Dimensions 1,2:
# Same units, different measurements

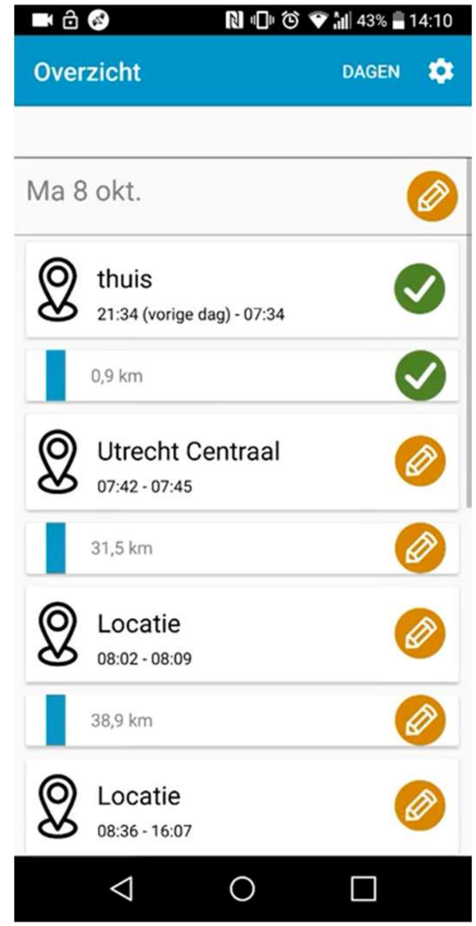
- Link 2 or more microdatasets of same individual
  - Data linkage (admin data)
  - Sampling frame information and survey data
  - Enriching surveys with administrative data

What quality dimension are we trying to improve?

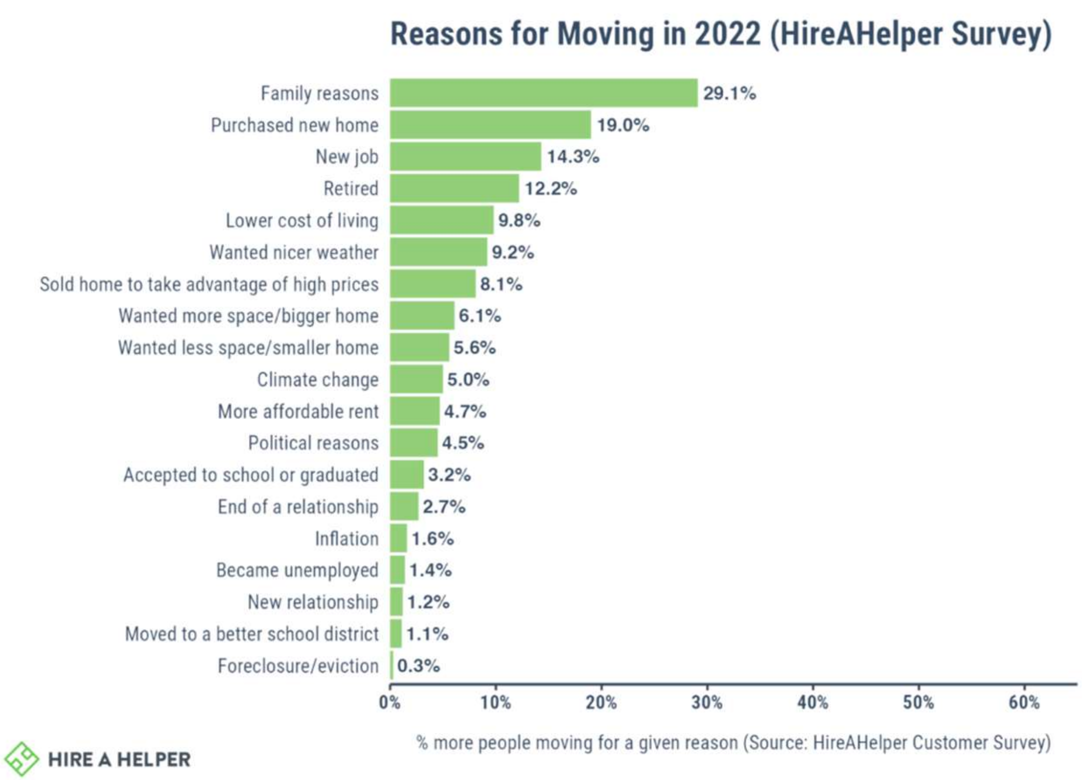
# Dimensions 1,2:
# Same units, different measurements

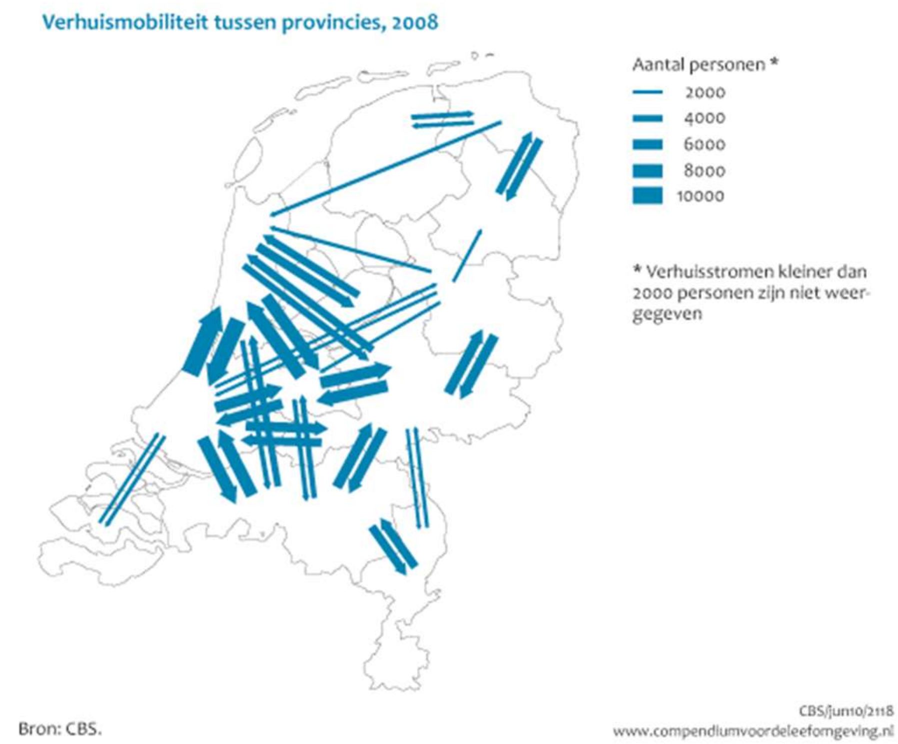
- Link 2 or more microdatasets of same individual
  - Designed big data
  - Lecture week 10



# Dimension 2,3: measurement, time



# Dimensions 2,4:
# Same measurement, different aggregation

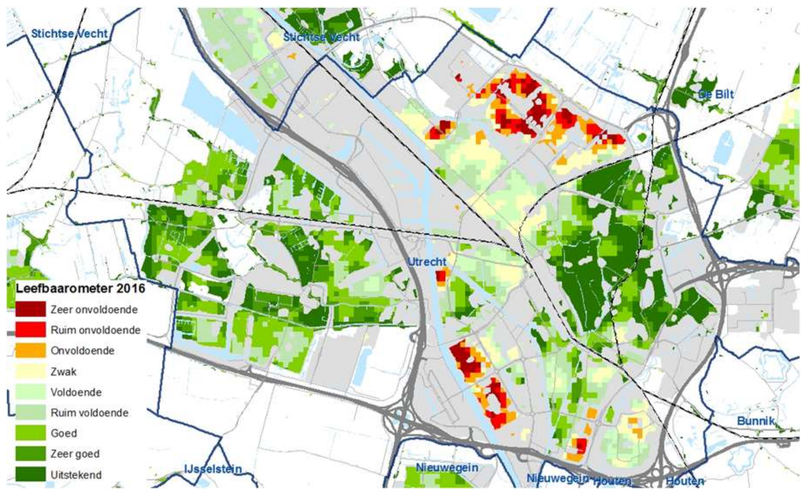
Use microdata + same statistic at aggregate levels

- Use population statistics for weighting/calibration
- Validate and asses accuracy of survey data
  - E.g. Sensitive questions

# Dimensions 3,4:
# different periods, different levels of aggregation

Use national survey data with local administrative data to make predictions at local level

- Small area estimation


Multilevel logistic regression models

$$\text{Logit}(p_{ij}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + u_j$$

Where $\beta_0$ is the 'intercept' and, $\beta_1$ to $\beta_p$ are the coefficients of the p explanatory variables

# *Data Integration and nonprobability samples: Key concepts*

**Camilla Salvatore**
c.salvatore@uu.nl

## **Revising key concepts:** Probability vs nonprobability samples

**Probability samples (PS)**

**Allow inferences to the general population**

Can you think about some good characteristics and statistical issues?

**Non-Probability samples (NPS)**

**Drawing inference is hard or not possible**

Can you think about some good characteristics and statistical issues?

## **Revising key concepts:** Probability vs nonprobability samples

### Probability samples (PS)

**Allow inferences to the general population**

- High data quality
- Rely on sampling theory
- Design/Model based inference
- Falling response rate, time-consuming, expensive

### Non-Probability samples (NPS)

**Drawing inference is hard or not possible**

- More affordable, timely, convenient, new aspects of phenomena
- No unified inferential framework
- Unknown selection mechanism:
  - Self-selection → selection bias (SB)
  - Diverse: NPS surveys, digital traces

Integrating PS and NPS data (Salvatore, 2023; Rao, 2021; Cornesse et al., 2020):

- Improve inference reducing also the costs of analysis
- Study new aspects not measured by traditional surveys

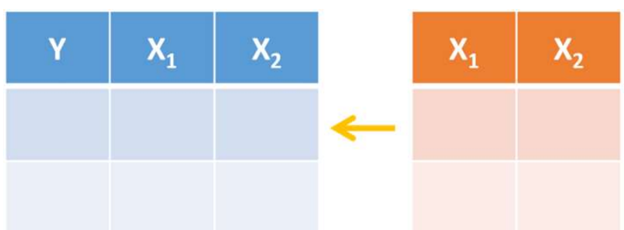
## **Revising key concepts:** Probability vs nonprobability samples

Comparing PS and NPS estimates (Pasek, 2016):

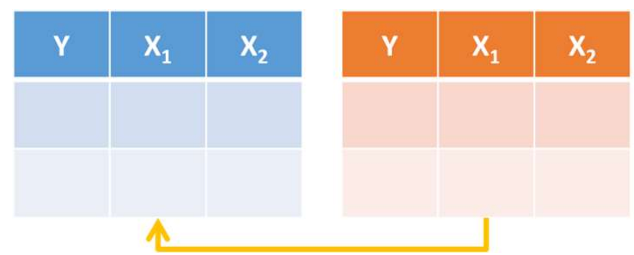
- Finite population estimates tend to be more dissimilar than correlations and regression coefficients
- No consensus about whether and in which cases differences will be notable

Two inferential approaches (Rao, 2020):

Adjusting for SB with auxiliary data




Blending PS and NPS



# Two principles of Data Integration (DI)

Any ideas? Insights from previous lectures/discussions?

Utrecht University

# Two principles of Data Integration (DI)

## 1. DI is statistics and purpose specific

**How do we integrate data?**

- Finite population and analytic inference
- Structured and unstructured data
- Composite indicators
- Variables are available in all sources or only in some
- One source is used as a supplement or to correct for selection bias
- Combining different sources to improve measurement

Utrecht University

# Two principles of Data Integration (DI)

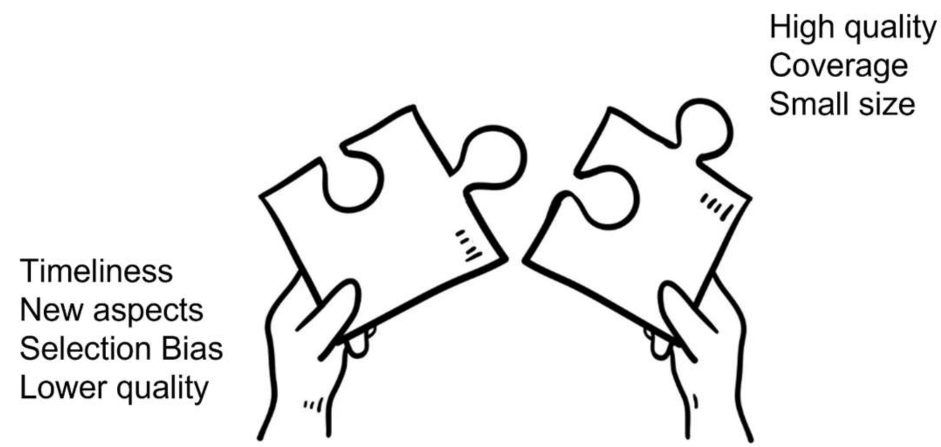
## 1. DI is statistics and purpose specific

**How do we integrate data?**

- Finite population and analytic inference
- Structured and unstructured data
- Composite indicators
- Variables are available in all sources or only in some
- One source is used as a supplement or to correct for selection bias
- Combining different sources to improve measurement

## 2. DI is a puzzle

**Why data integration?**




Utrecht University

# *Enhancing analytic inference while reducing the costs: a Bayesian data integration approach*

**Camilla Salvatore**
c.salvatore@uu.nl

# The research paper

*Journal of Survey Statistics and Methodology* (2023) **00**, 1–35

**BAYESIAN INTEGRATION OF PROBABILITY AND NONPROBABILITY SAMPLES FOR LOGISTIC REGRESSION**

CAMILLA SALVATORE *
SILVIA BIFFIGNANDI
JOSEPH W. SAKSHAUG
ARKADIUSZ WIŚNIOWSKI
BELLA STRUMINSKAYA

https://doi.org/10.1093/jssam/smad041

Utrecht University

## The context

**Problem**

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

# The context

## Problem

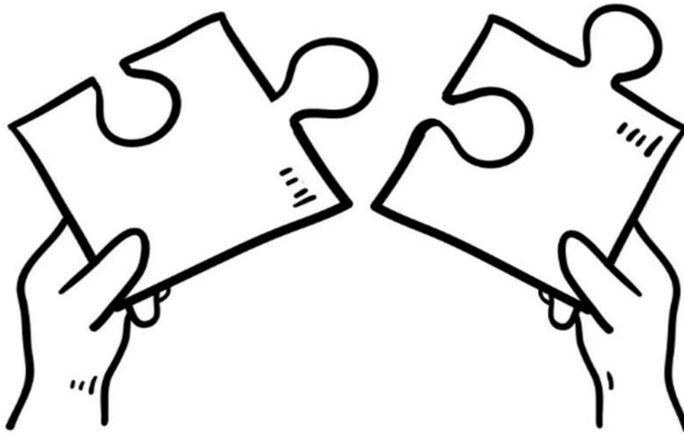
A researcher is interested in making inferences from a PS survey but cannot afford a large sample size

## Alternatives

1. Reduce the sample size: small PS → large variance but theoretically unbiased estimates
2. Opt for a NPS: bias but low variance

# Our proposal

## The data integration puzzle

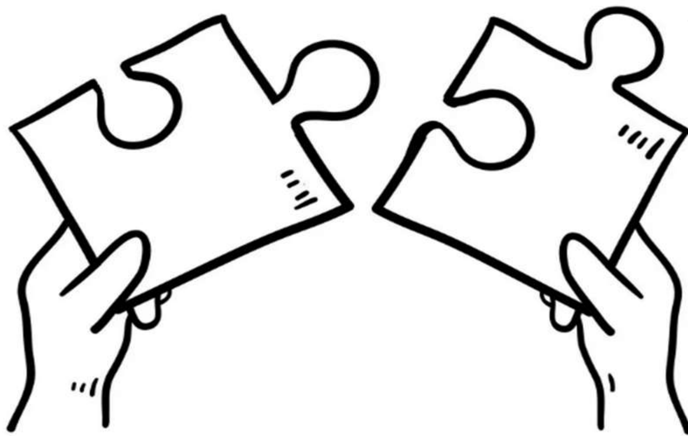


# Our proposal

**The data integration puzzle**

Small size to reduce costs

PS (high quality)
Unbiased
Large variance



NPS (lower quality)
Selection Bias
Lower variance

Utrecht University

# Our proposal

### The data integration perspective

- Integrate small PS + larger NPS
- to improve inference on logistic regression coefficients
- under the Bayesian framework
- reducing survey costs

### Inference

- Based on small PS data (unbiased, high variance)
- Incorporation of 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 costly PS only survey?

## Two aims

1. **Enhance inference (MSE)** → Simulation & Real Data Analysis →
   - *Selection scenarios & level of SB*
   - *Outcome variables*
   - *PS and NPS sizes*
   - 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** → Cost Analysis → Interactive Shiny App
   - Can we obtain at a lower cost the same MSE that we would obtain analyzing a much larger and costly PS only survey?

# What is Bayesian statistics?

**Thomas Bayes**

**Bayes' theorem**

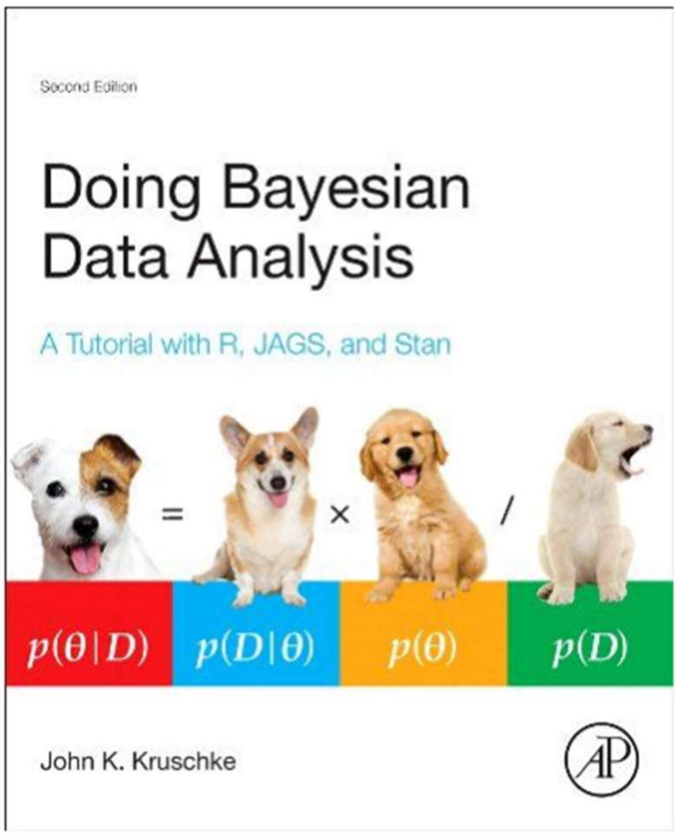
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

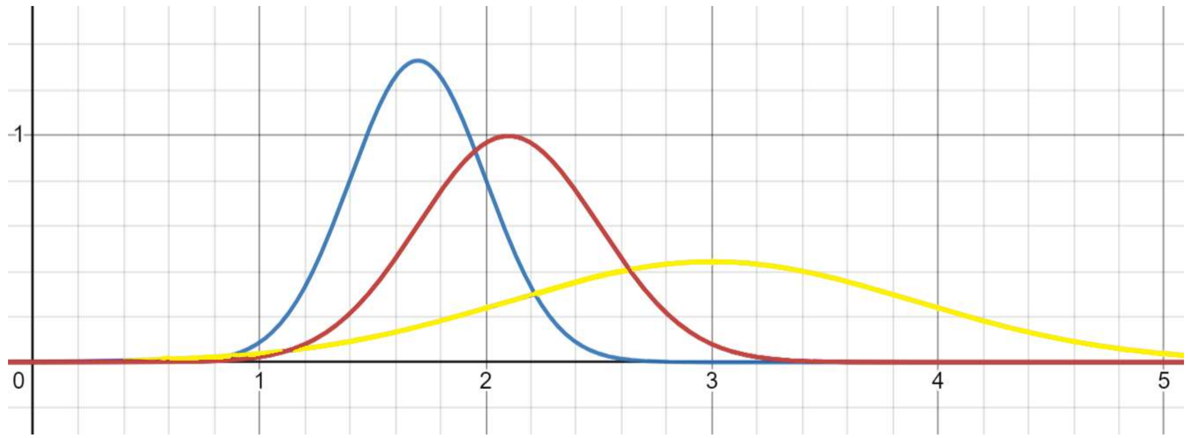
## What is Bayesian statistics?

### Some differences with the frequentist approach

- Bayesian: parameters are random variables
- Frequentist: parameters are non-random. Randomness is introduced by sampling
- In Bayesian statistics prior belief play a fundamental role (subjective approach):
    - We start with a prior belief (*prior to looking at the data*) and we update it using the data → we obtain the posterior

## What is Bayesian statistics?




$$p(\theta|x) = p(D|x) \times p(\theta)/p(D)$$

**Posterior** **Likelihood** **Prior** **Marginal**

$$\Downarrow$$

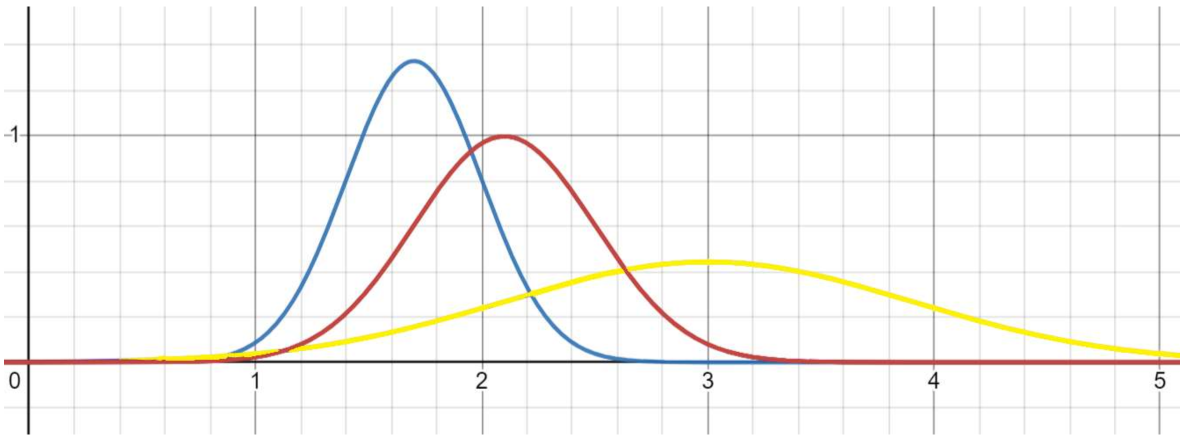
$$p(\theta|x) \propto p(D|x) \times p(\theta)$$

Utrecht University

## **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

$$p(\theta|x) \propto p(D|x) \times p(\theta)$$



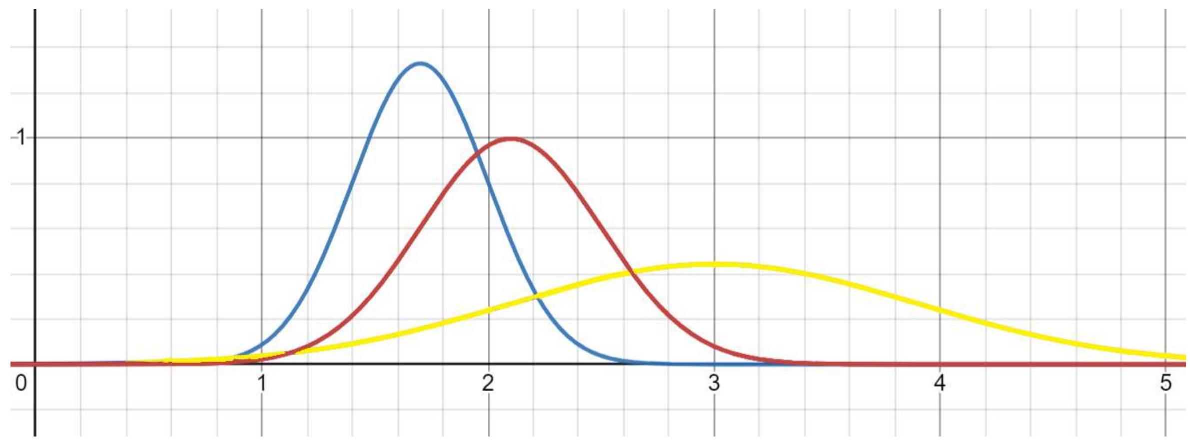
## **Why Bayesian?** (Kruschke, 2014; Gelman et al., 2013)

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

$$p(\theta|x) \propto p(D|x) \times p(\theta)$$

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

We borrow information based on the similarity between PS and NPS



# Research structure

- Background: Sakshaug et al (2019) and Wisniowski et al. (2020) papers (Continuous outcome variable)

**Part I –** Simulation study (100 repetitions)

- Different selection scenarios, prior specifications, PS and NPS sizes
- Evaluate the performance of several informative priors against a PS-ONLY one in terms of MSE

**Part II –** Real data analysis

- American Trend Panel + 9 parallel NPS surveys
- Shiny app with interactive cost analysis

Utrecht University

## Priors

**PS-ONLY** (No data integration)

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

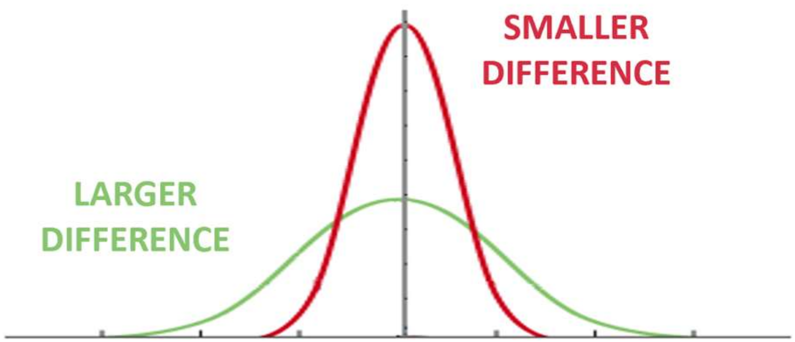
$$\beta_j \sim Student(\nu = 3, \mu = 0, s = 2.5) \quad \text{for j=0,1,2}$$

## Informative priors: integrating PS and NPS data

**Distances priors:** The influence of the prior depends on the difference between ML estimates

The Basic distance prior

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



# Informative priors: integrating PS and NPS data

**Distances priors:** The influence of the prior depends on the difference between ML estimates

The Distance Log Prior

$$\beta_j \sim \mathcal{N}\left(\widehat{\beta}_{j\text{NPS}}, \sqrt{\frac{1}{\log(n_{\text{NPS}})} \cdot \max\left(\left(\widehat{\beta}_{j\text{PS}} - \widehat{\beta}_{j\text{NPS}}\right)^2, \widehat{\sigma}^2_{\beta_j\text{NPS}}\right)}\right).$$

The Distance Log 10 Prior (wider distribution)

$$\beta_j \sim \mathcal{N}\left(\widehat{\beta}_{j\text{NPS}}, \sqrt{\frac{1}{\log_{10}(n_{\text{NPS}})} \cdot \max\left(\left(\widehat{\beta}_{j\text{PS}} - \widehat{\beta}_{j\text{NPS}}\right)^2, \widehat{\sigma}^2_{\beta_j\text{NPS}}\right)}\right).$$

**Mixed distance priors:** Baseline prior for $\beta_0$ and distances priors for other coefficients

Utrecht University

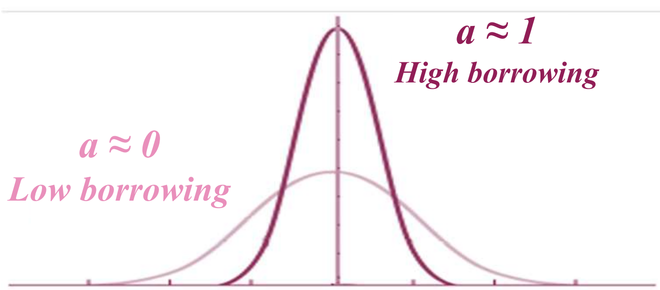
# Informative priors: integrating PS and NPS data

**Power prior** (Ibrahim et al., 2000)

$$\pi(\beta, a|D_{NP}) \propto L(\beta|D_{NP})^{a}\pi_0(\beta)$$

*Power prior* *Likelihood NPS* *Baseline prior*

**Likelihood NPS**




**How much do we borrow from NPS?**

The power parameter "a":

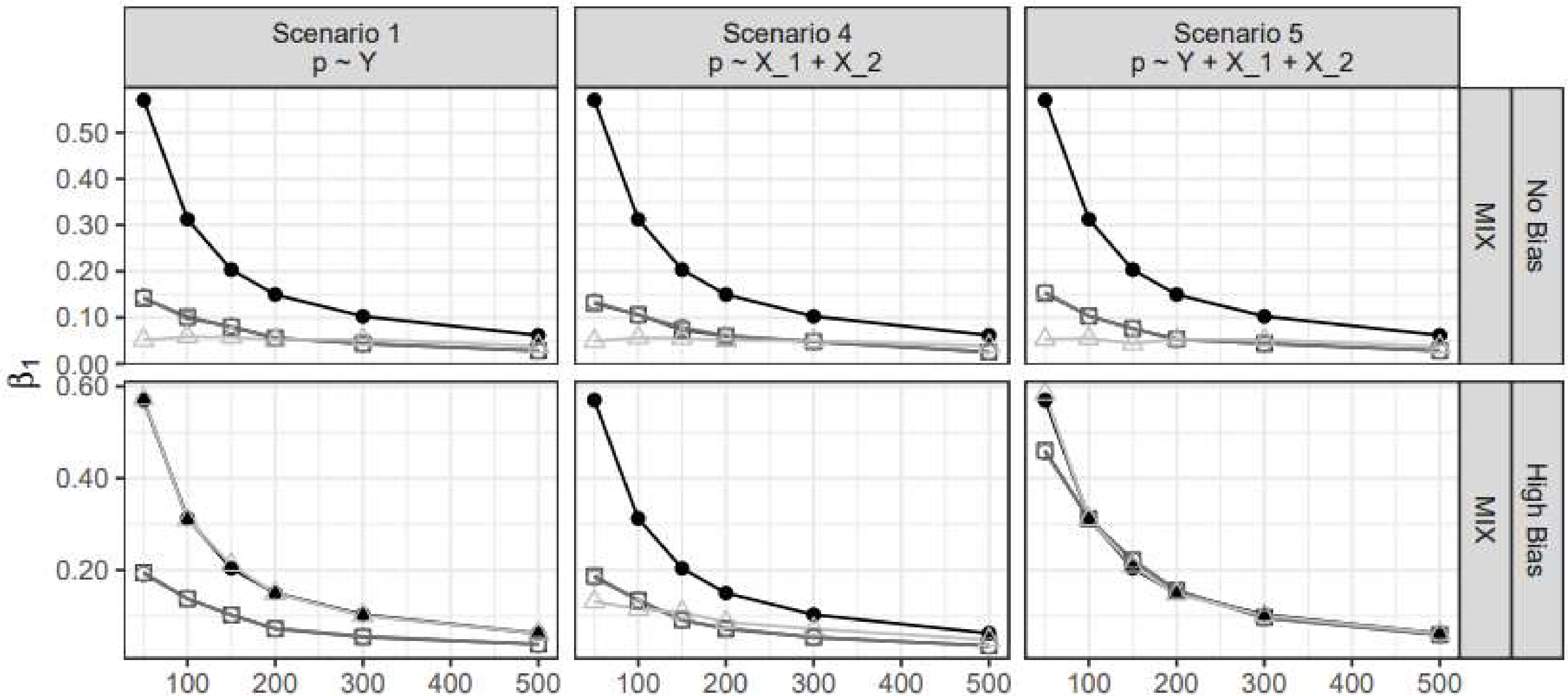
1 = full borrowing

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 between $\beta_P$ and $\beta_{NP}$ )

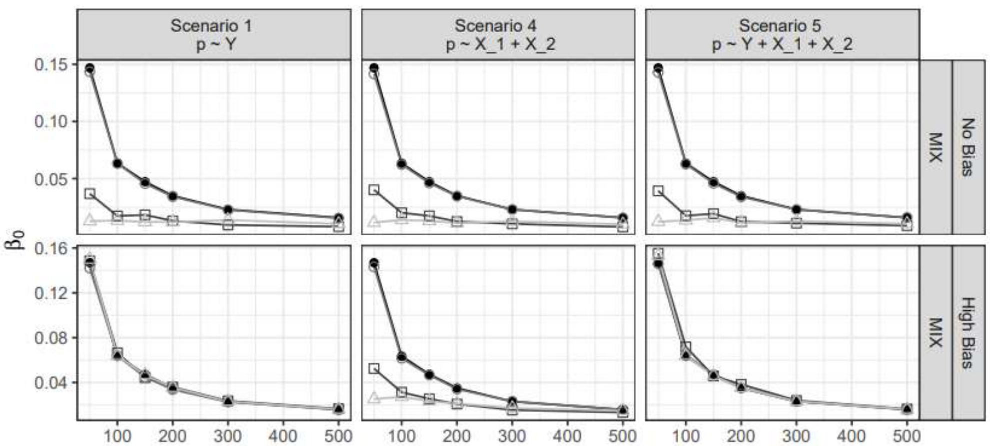
Utrecht University

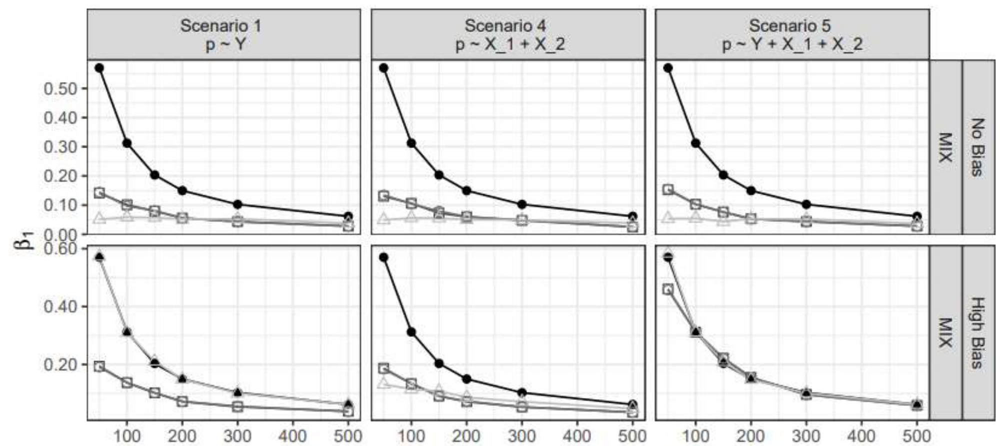
## Results: selected cases – Beta1

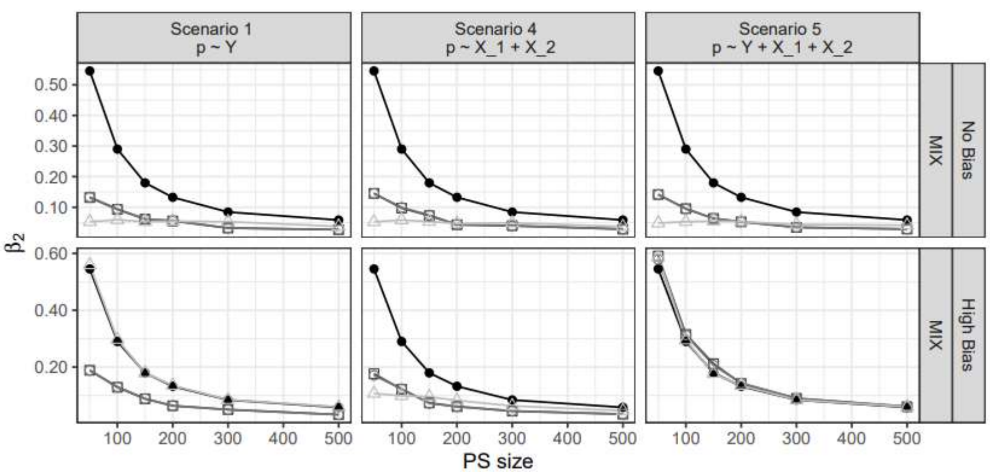


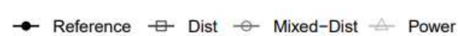

Utrecht University

# Results: selected cases




## Median MSE across 100 repetitions:

- Worst case-scenario: INF prior perform similarly to PS-only prior (bias protection mechanism)
- Low SB and small PS: large improvements in MSE

# Application: the data

**PS data** – American Trends Panel (ATP)

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

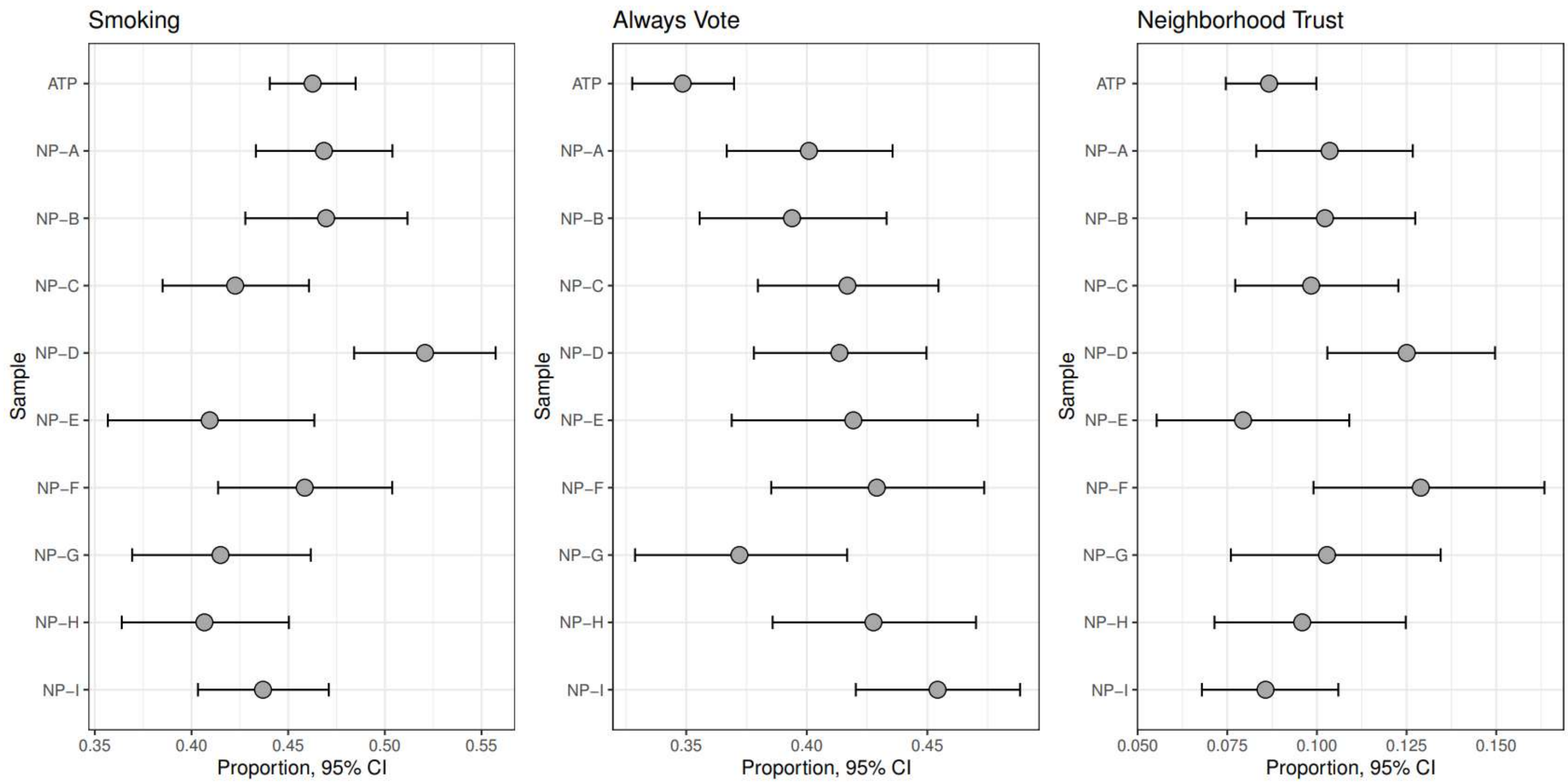
**NPS data** - 9 parallel online NPS from different vendors

- Vendors implemented quota sampling with different quota variables (demographic vs webographic)
- Sample size of about 1000 respondents

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

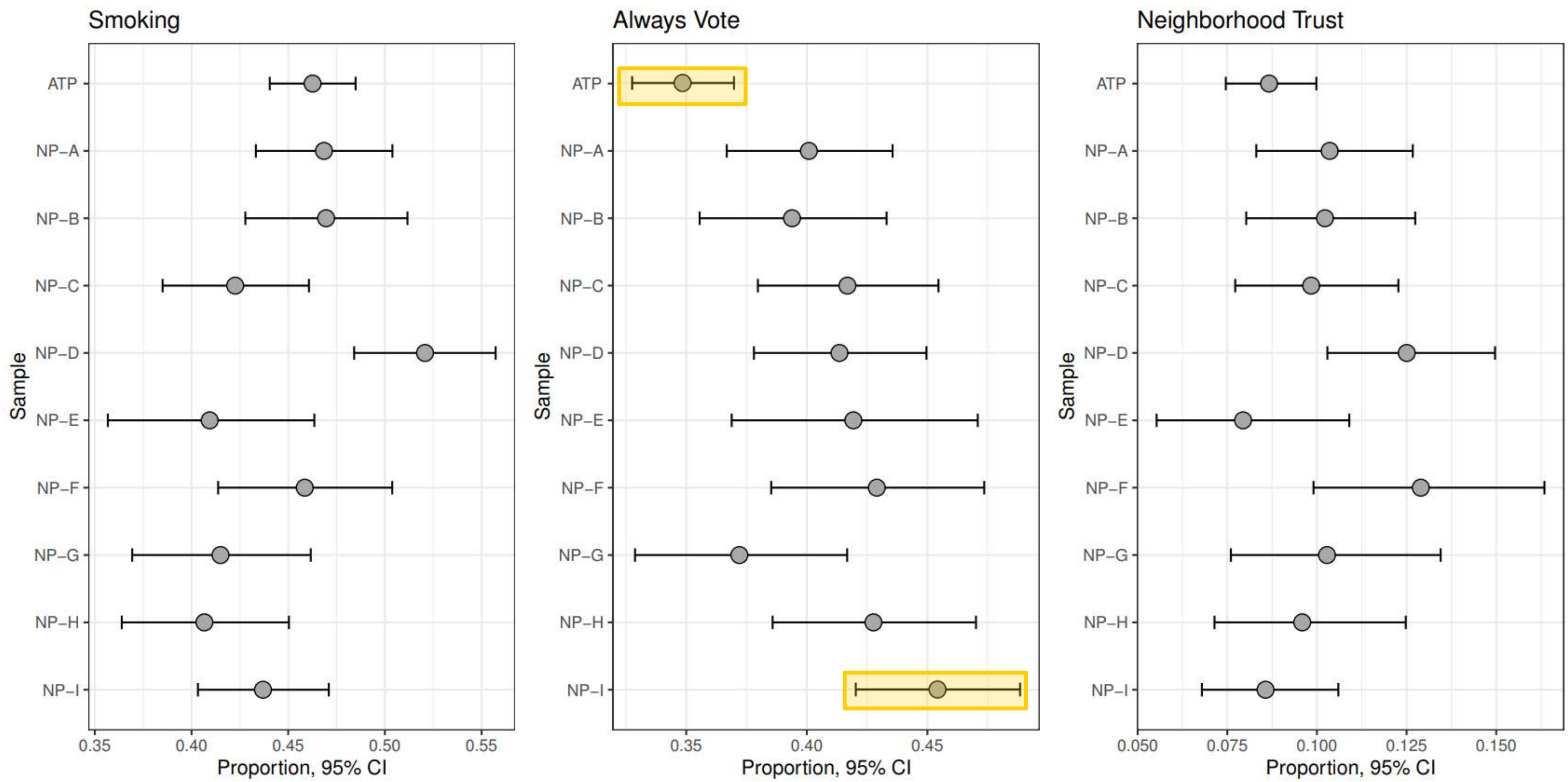
**Covariates**: Age, gender, education, survey weight

# Comparing proportions

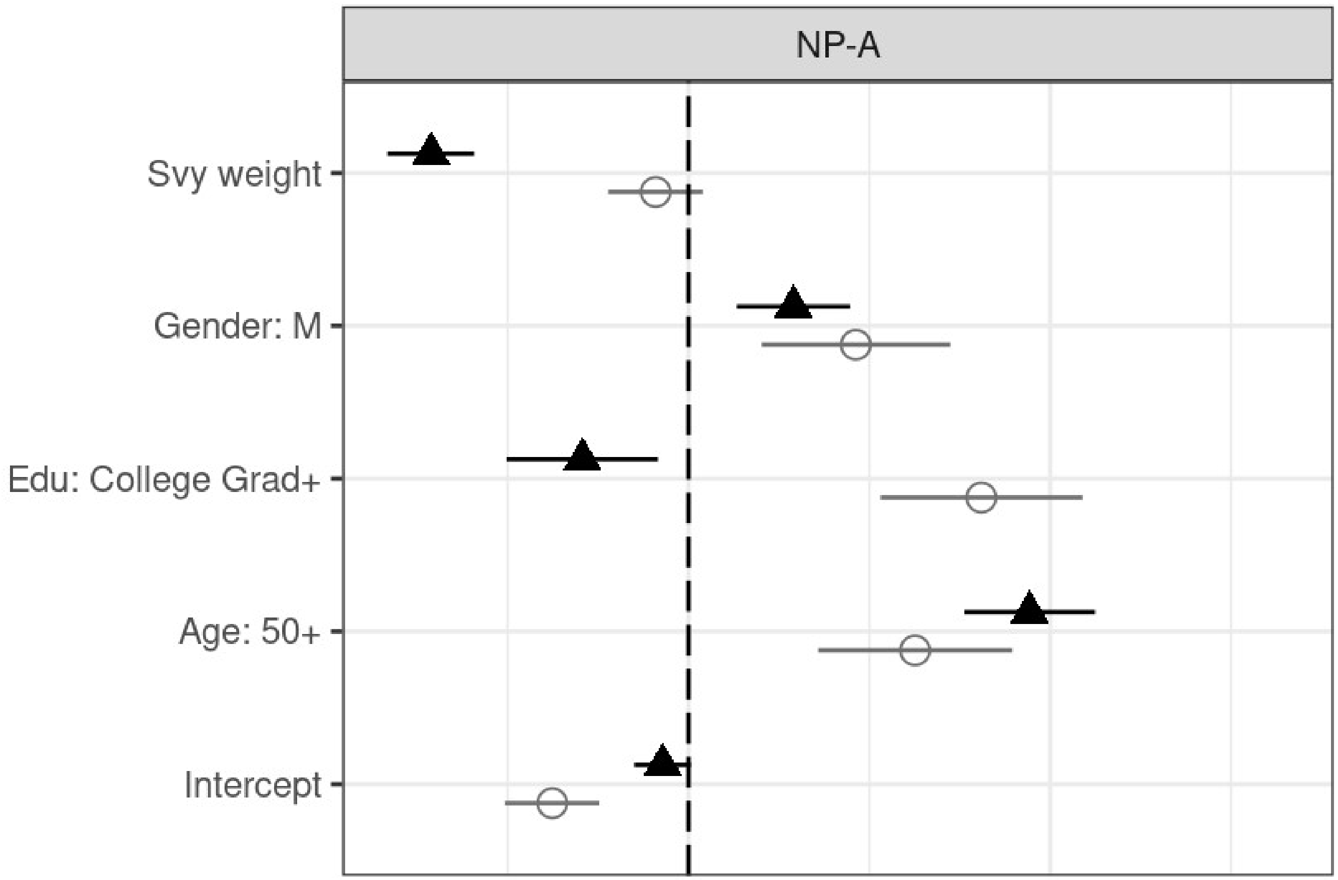



Utrecht University

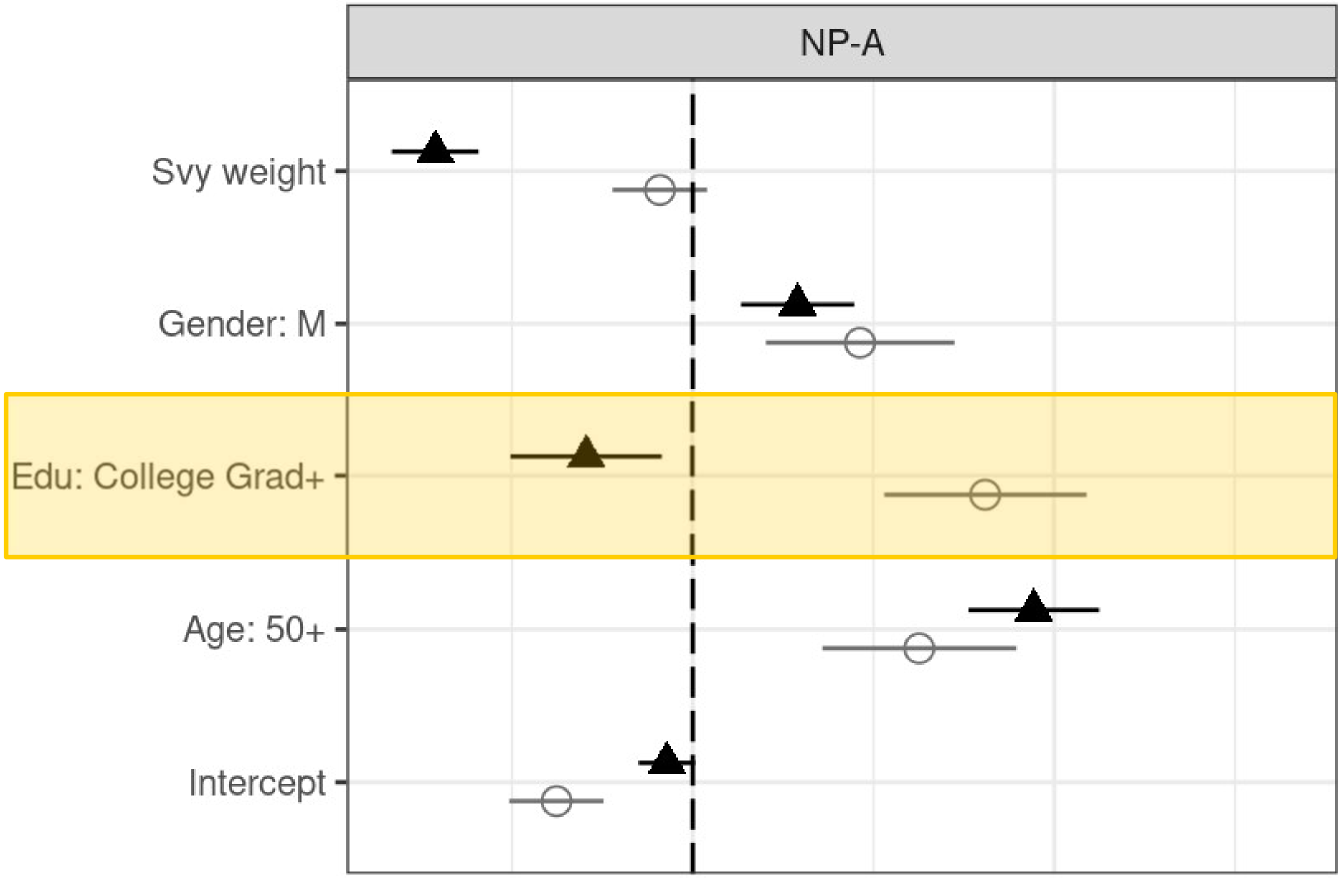
# Comparing proportions



## Comparing coefficients: an example with *Always vote*

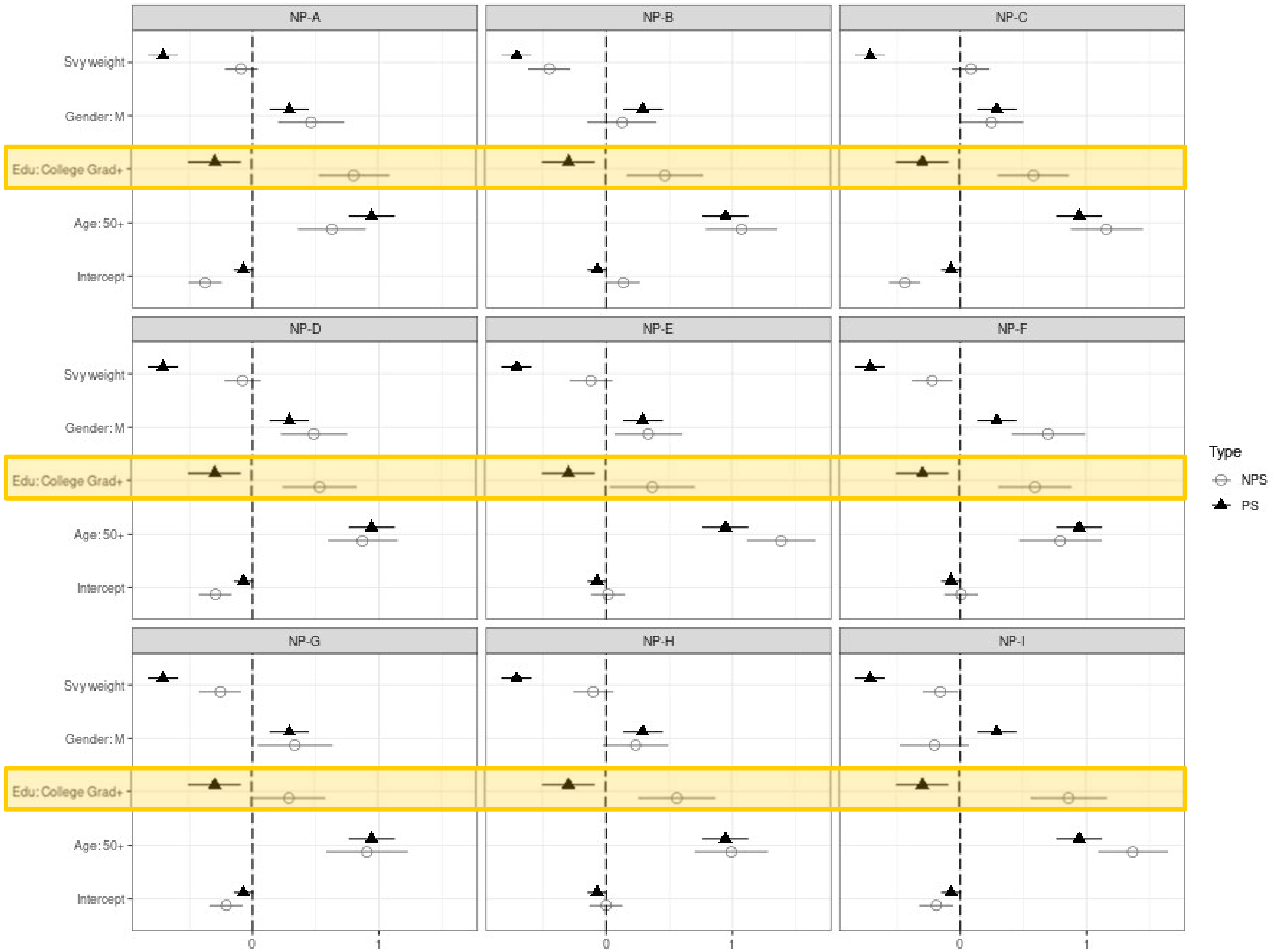


## Comparing coefficients: an example with *Always vote*



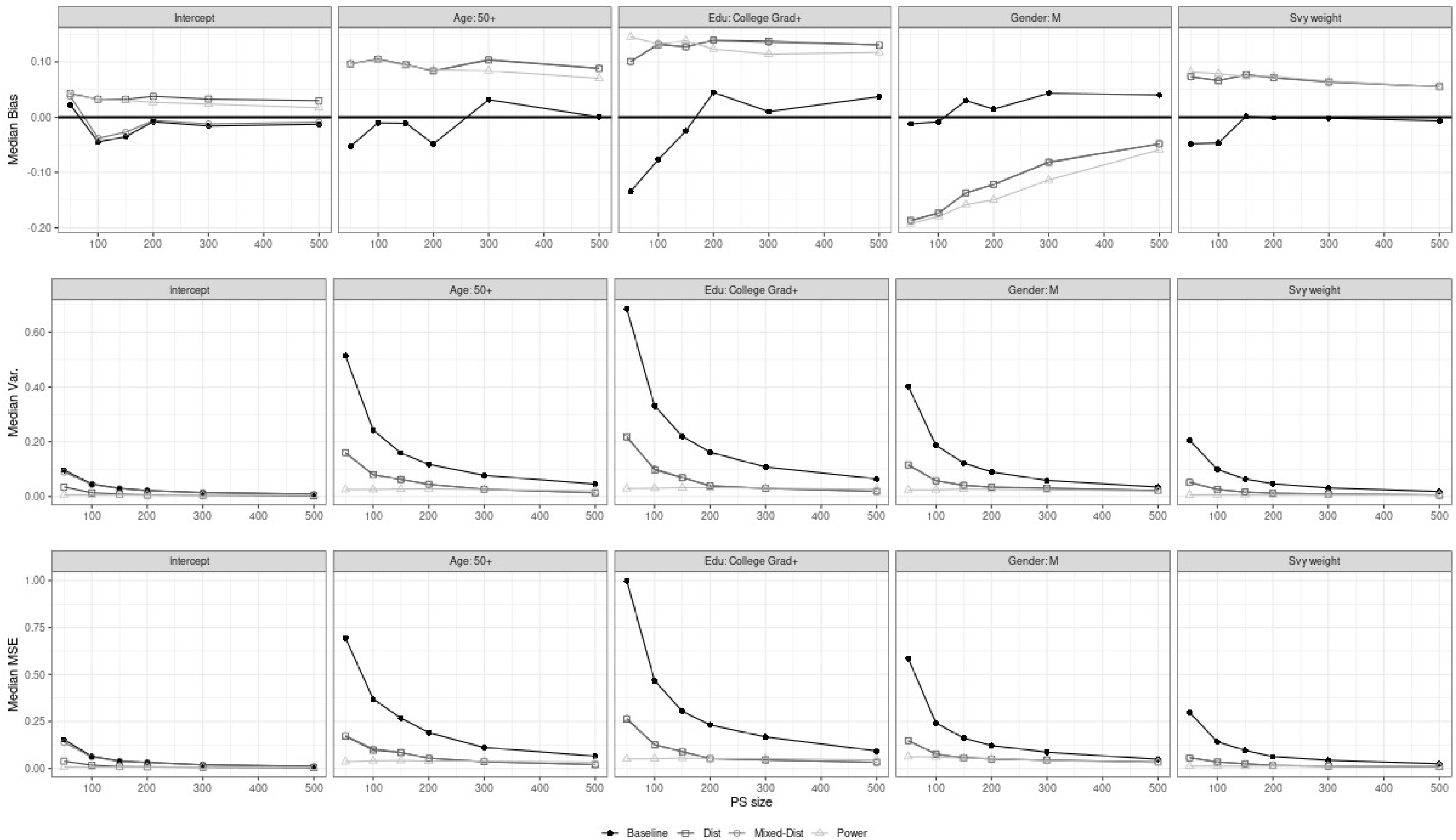

Utrecht University

# Comparing coefficients: an example with *Always vote*



# Results: an example with *Smoking*

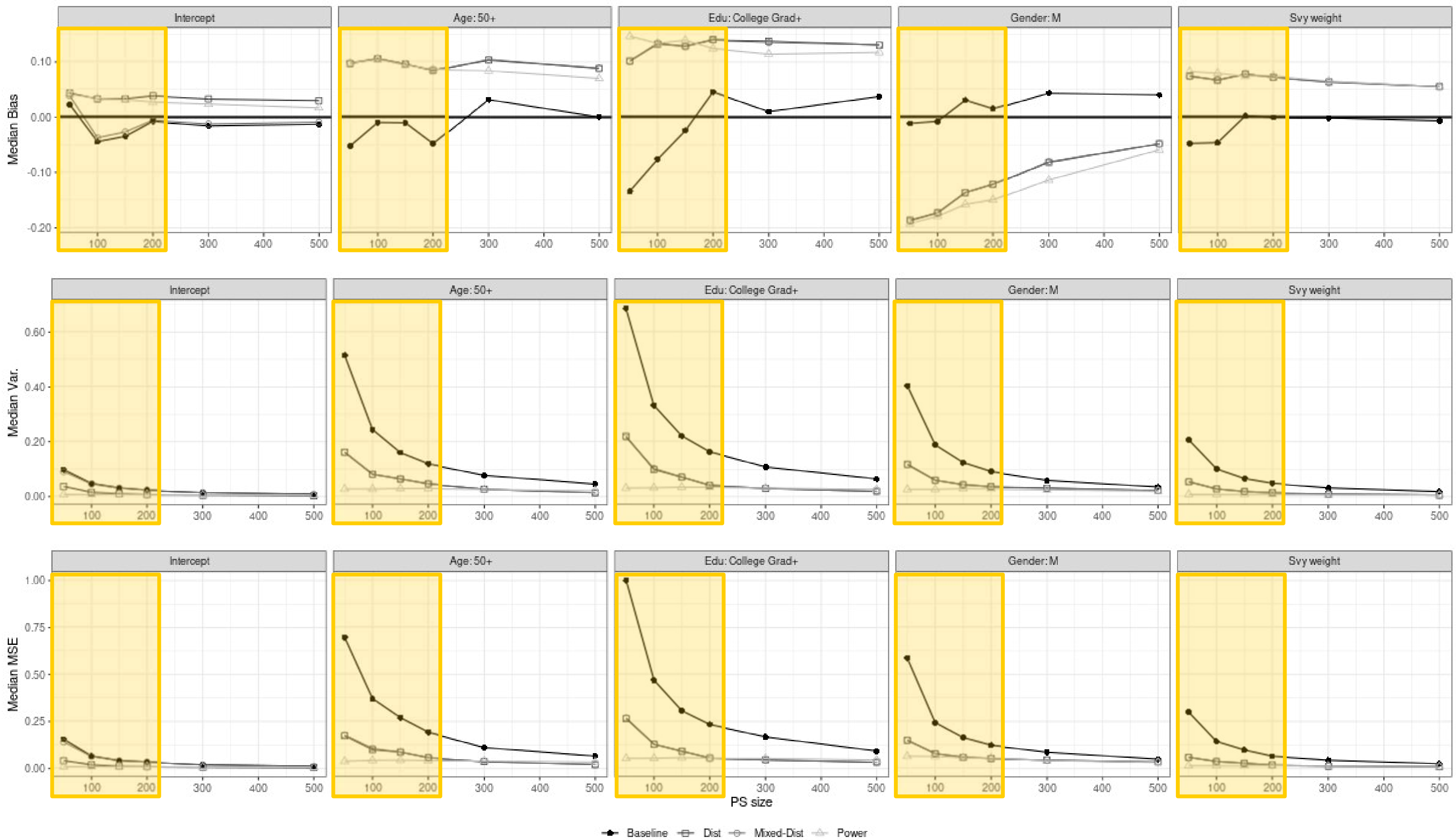
Reduction in MSE is driven by a reduction in the variability




Utrecht University

# Results: an example with *Smoking*

Reduction in MSE is driven by a reduction in the variability




Utrecht University

# Cost Analysis

**Results**

- <u>PS costs ≥ 3 times NPS costs</u>: best performing INF priors yield significant cost savings ≈ 70%
- <u>PS costs = 2 times NPS costs</u>: cost savings are marginal or negative

**Interactive Analysis:** Shiny App

https://bayesdataintegration.shinyapps.io/shiny_bayes_data_integration/

Utrecht University

## Take aways

- Survey researchers face budgetary and time constraints → fielding large size PS is difficult
- Small PS yield large variances for survey estimates
- Our approach offers a practical solution to improve analytic inference (reduced variances and MSEs) while lowering survey costs
- Shiny App: facilitate researchers interested in designing and integrating parallel PS and NPS

Some thoughts:

1. Do you think integrating various data types will become crucial for surveys research (& Official Stat.)?
2. What do you think about integrating surveys and digital trace (big) data?

Utrecht University

## Group exercise (40 min)

- 25 min group work + 5 min discussion per group
- 6 groups:
  - Group 1: Smoking
  - Group 2: Always vote
  - Group 3: Volunteering
  - Group 4: Neighborhood Safety
  - Group 5: Healthcare coverage
  - Group 6: Neighborhood Trust
- Shortly present your findings to others
- Points of discussion in the next slides (you can decide to address them all or only the most relevant for your case)

## Group exercise (40 min)

- Open the Shiny App and go to Real Data Analysis
- In Data/Additional plots look at the variables of interest:
  - Are there differences across NPS? Which NPS do you think is better?
  - Are PS and NPS estimates similar?
- In Data/Results look at the variables of interest:
  - Is there a bias/variance trade off?
  - Are there differences across NPS? Which NPS you think is better?
  - Which prior works better and under which conditions (ex. PS or NPS size)?
- In Data/Cost Analysis/Max savings look at the variables of interest (use the search tool)
  - Which is the best NPS in terms of savings if the PS cost is 6 times the NPS cost (the default in the app)?
  - Does the result change according to the PS size?

# References


- Salvatore, C., Biffignandi, S., Sakshaug, J. W., Wiśniowski, A., & Struminskaya, B. (2023). Bayesian Integration of Probability and Nonprobability Samples for Logistic Regression. *Journal of Survey Statistics and Methodology,* https://doi.org/10.1093/jssam/smad041
- Salvatore, C. (2023). Inference with non-probability samples and survey data integration: a science mapping study. *Metron*, 1-25. https://doi.org/10.1007/s40300-023-00243-6
- Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., & Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, *35*(3), 653-681.
- Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, *8*(1), 120-147.
- Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., ... & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4-36.
- Beaumont, J. F., & Rao, J. N. K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? Surv. *Stat*, *83*, 11-22.
- Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, *83*, 242-272.
- Pasek, J. (2016). When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence. *International Journal of Public Opinion Research*, *28*(2), 269-291.
- Ibrahim, J. G., & Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Science*, 46-60.

# References


- Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley & Sons.
- De Waal, T., van Delden, A., & Scholtus, S. (2020). Multi-source statistics: basic situations and methods. *International Statistical Review*, *88*(1), 203-228.
- llic, G., Schouten, J.G., Lugtig, P., Mulder, J., Streefkerk, M., Kumar, and P. Höcük, S. (2022). [Pictures instead of survey questions: An experimental investigation of the feasibility of using pictures in a housing survey](). JRSS:A.
- McCool, D., Schouten, J.G. & Lugtig, P. (2021). An app-assisted travel survey in official statistics. Possibilities and challenges. *Journal of Official Statistics*
- van Delden, A., Scholtus, S., de Waal, T., & Csorba, I. (2023). Methods for estimating the quality of multisource statistics. *Advances in Business Statistics, Methods and Data Collection*, 781-804.
- Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, *8*(1), 120-147.