Correcting for Nonignorable Nonresponse Bias in Ordinal Observational Survey Data

Jozef Michal Mintál[1], Lukáš Lafférs[2], Ivan Sutoris[3]

[1]Matej Bel University, Research and Innovation Center
[2]Matej Bel University, Dept. of Mathematics
[2]NHH, Dept. of Economics
[3]NBS

SEAM conference 2025

Motivation

Motivation

- Survey sample data often not representative of general population.
- We cannot sample from the general population - difficult.

- Even if we could, how about the non-response?

- Cannot be ignored.
- Non-response rates easily $\sim 50\%$

- We are interested in ordinal data.
- These are very common.

”How satisfied are you with life?”

- Extremely satisfied
- Very satisfied
- Moderately satisfied
- Slightly satisfied
- Not satisfied at all

”National economy has gotten better or worse?”

- Gotten much better
- Gotten somewhat better
- Stayed about the same
- Gotten somewhat worse
- Gotten much worse

”Do you favor or oppose death penalty”

- Favor strongly
- Favor not strongly
- Oppose not strongly
- Oppose strongly

”How willing should US be to use military force to solve international problems?”

- Extremely willing
- Very willing
- Moderately willing
- A little willing
- Not at all willing

We would like to have a model that allows for

- survey sample weighting
- estimation of relationship between outcomes and response and thus modeling non-response selection bias
- the use of covariates to model outcomes and responses

Peress (2010): ☑ ☑ ☑

Peress, Michael. "Correcting for survey nonresponse using variable response propensity." Journal of the American Statistical Association 105.492 (2010): 1418-1430.

But also

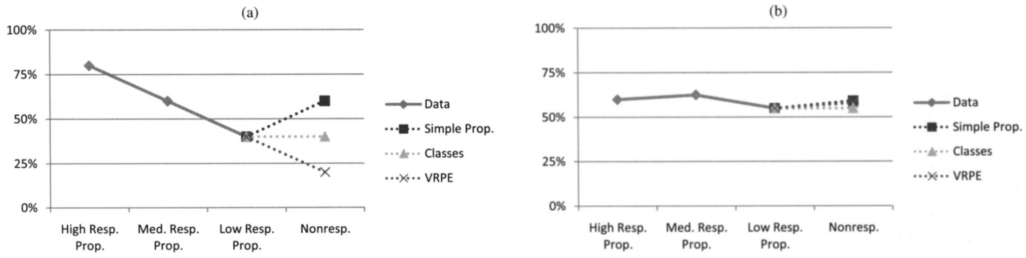
- can handle ordinal data

Peress (2010): ☑ ☑ ☑ ☒

This paper: ☑ ☑ ☑ ☑

Main idea is that we extrapolate from low-propensity respondents to $\rightarrow$ non-respondents.

- No matter what we do, we have to extrapolate somehow.



Peress (2010), p.1421

Literature

- extension of variable response propensity estimator (VRPE) of Peress (2010)
- Heckman (1979) - sample selection models
- continuum of resistance models - Fillion (1975), Drew and Fuller (1980)
- classes models - O'Neil (1979)
- missing data problem - Rosenbaum and Rubin (1983)
- Behaghel et al. (2015): bounds in the spirit of Lee (2009)

Model

Model with Gaussian errors ε_n and η_n

Outcome model

$$y_n \in \{1, 2, 3, ..., Y\}$$

$$y_n^* = \alpha^T x_n + \varepsilon_n$$

$$y_n = \begin{cases} 1 & \text{if } y_n^* \leq \gamma_1 \\ 2 & \text{if } y_n^* \in (\gamma_1, \gamma_2] \\ 3 & \text{if } y_n^* \in (\gamma_2, \gamma_3] \\ \vdots & \\ Y & \text{if } y_n^* > \gamma_{Y-1}. \end{cases}$$

Response model

$$r_n \in \{1, 2, 3, ..., R\}$$

$$r_n^* = \beta^T z_n + \eta_n$$

$$r_n = \begin{cases} 1 & \text{if } r_n^* \leq \theta_1 \\ 2 & \text{if } r_n^* \in (\theta_1, \theta_2] \\ 3 & \text{if } r_n^* \in (\theta_2, \theta_3] \\ \vdots & \\ R & \text{if } r_n^* > (\theta_{R-1}, \theta_R] \\ R+1 & \text{if } r_n^* > \theta_R. \end{cases}$$

$$\text{corr}(\varepsilon_n, \eta_n) = \rho$$

Non-respondents

Outcome model

$$y_n \in \{1, 2, 3, ..., Y\}$$

$$y_n^* = \alpha^T x_n + \varepsilon_n$$

$$y_n = \begin{cases} 1 & \text{if } y_n^* \leq \gamma_1 \\ 2 & \text{if } y_n^* \in (\gamma_1, \gamma_2] \\ 3 & \text{if } y_n^* \in (\gamma_2, \gamma_3] \\ \vdots & \\ Y & \text{if } y_n^* > \gamma_{Y-1}. \end{cases}$$

Response model

$$r_n \in \{1, 2, 3, ..., R\}$$

$$r_n^* = \beta^T z_n + \eta_n$$

$$r_n = \begin{cases} 1 & \text{if } r_n^* \leq \theta_1 \\ 2 & \text{if } r_n^* \in (\theta_1, \theta_2] \\ 3 & \text{if } r_n^* \in (\theta_2, \theta_3] \\ \vdots & \\ R & \text{if } r_n^* > (\theta_{R-1}, \theta_R] \\ R+1 & \text{if } r_n^* > \theta_R. \end{cases}$$

$$\text{corr}(\varepsilon_n, \eta_n) = \rho$$

Parameters $(\alpha, \beta, \gamma, \theta, \rho)$

Outcome model

$$y_n \in \{1, 2, 3, ..., Y\}$$

$$y_n^* = \alpha^T x_n + \varepsilon_n$$

$$y_n = \begin{cases} 1 & \text{if } y_n^* \leq \gamma_1 \\ 2 & \text{if } y_n^* \in (\gamma_1, \gamma_2] \\ 3 & \text{if } y_n^* \in (\gamma_2, \gamma_3] \\ \vdots & \\ Y & \text{if } y_n^* > \gamma_{Y-1}. \end{cases}$$

Response model

$$r_n \in \{1, 2, 3, ..., R\}$$

$$r_n^* = \beta^T z_n + \eta_n$$

$$r_n = \begin{cases} 1 & \text{if } r_n^* \leq \theta_1 \\ 2 & \text{if } r_n^* \in (\theta_1, \theta_2] \\ 3 & \text{if } r_n^* \in (\theta_2, \theta_3] \\ \vdots & \\ R & \text{if } r_n^* > (\theta_{R-1}, \theta_R] \\ R+1 & \text{if } r_n^* > \theta_R. \end{cases}$$

$$\text{corr}(\varepsilon_n, \eta_n) = \rho$$

Data (y_n, r_n, x_n, z_n)

Outcome model

$$y_n \in \{1, 2, 3, ..., Y\}$$

$$y_n^* = \alpha^T x_n + \varepsilon_n$$

$$y_n = \begin{cases} 1 & \text{if } y_n^* \leq \gamma_1, \\ 2 & \text{if } y_n^* \in (\gamma_1, \gamma_2] \\ 3 & \text{if } y_n^* \in (\gamma_2, \gamma_3] \\ \vdots & \\ Y & \text{if } y_n^* > \gamma_{Y-1}. \end{cases}$$

Response model

$$r_n \in \{1, 2, 3, ..., R\}$$

$$r_n^* = \beta^T z_n + \eta_n$$

$$r_n = \begin{cases} 1 & \text{if } r_n^* \leq \theta_1, \\ 2 & \text{if } r_n^* \in (\theta_1, \theta_2] \\ 3 & \text{if } r_n^* \in (\theta_2, \theta_3] \\ \vdots & \\ R & \text{if } r_n^* > (\theta_{R-1}, \theta_R] \\ R+1 & \text{if } r_n^* > \theta_R. \end{cases}$$

$$\text{corr}(\varepsilon_n, \eta_n) = \rho$$

Log-Likelihood

$$\log L(\alpha, \beta, \gamma, \theta, \rho | y_n, r_n, x_n, z_n)$$

$$=$$

$$\sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{y=1}^{Y} I\{r_n = r, y_n = y\} \times$$

$$\times \log \int I\{\gamma_{y-1} \leq \alpha^T x_n + \varepsilon \leq \gamma_y, \theta_{r-1} \leq \beta^T z_n + \eta \leq \theta_r\} \; \phi(\varepsilon, \eta) \; \mathrm{d}\varepsilon \; \mathrm{d}\eta$$

$$+$$

$$N_{miss} \cdot \log \sum_{k=1}^{K} p_k^z \int I\{\beta^T z_k + \eta \geq \theta_R\} \; \phi(\eta) \mathrm{d} \; \eta$$

$$\log L(\alpha, \beta, \gamma, \theta, \rho | y_n, r_n, x_n, z_n)$$

$$=$$

$$\sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{y=1}^{Y} I\{r_n = r, y_n = y\} \times$$

$$\times \log \int I\{\gamma_{y-1} \leq \alpha^T x_n + \varepsilon \leq \gamma_y, \theta_{r-1} \leq \beta^T z_n + \eta \leq \theta_r\} \underbrace{\phi(\varepsilon, \eta)}_{\rho \text{ is here}} \, \mathrm{d}\varepsilon \, \mathrm{d}\eta$$

$$+$$

$$\underbrace{N_{miss} \cdot \log \sum_{k=1}^{K} p_k^z \int I\{\beta^T z_k + \eta \geq \theta_R\} \, \phi(\eta) \mathrm{d} \, \eta}_{\text{non-respondents}}$$

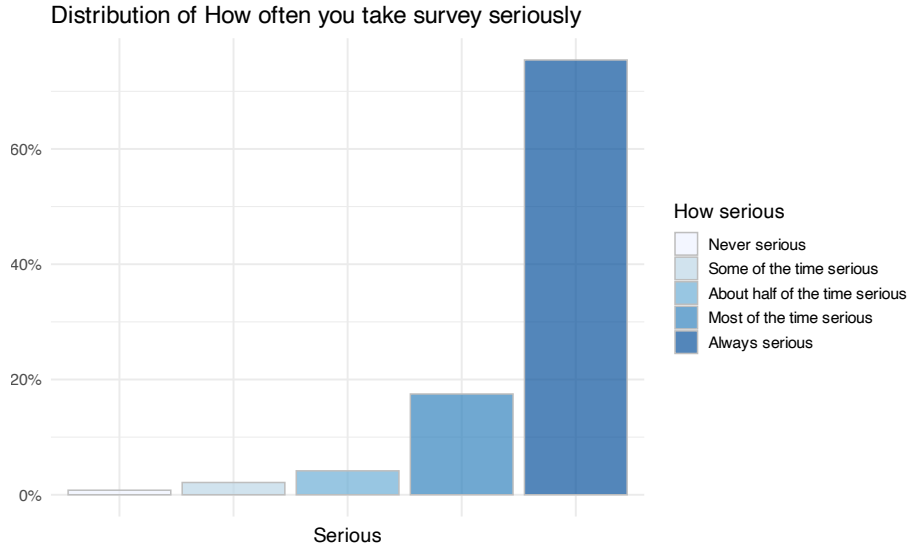
data, parameters, outcome error, response error, non-respondents, weights

Illustration

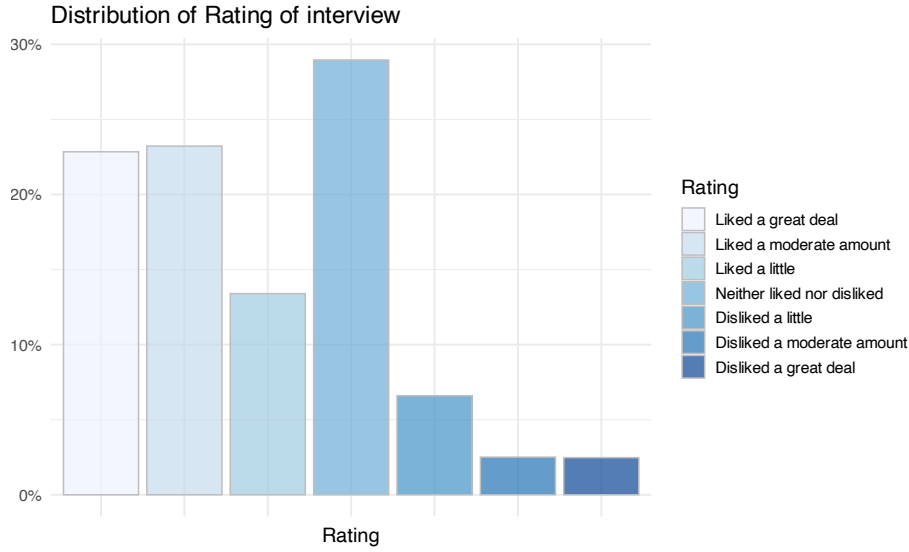
American National Election Studies data

- Published Feb 2025
- $\sim$ 3000 obs: face-to-face, web, paper
- $\sim$ 50% non-response
- response variables: rate interviewer, rate interview, do you take survey seriously
- covariates: married, gender, race, education
- outcomes: ordinal data (various questions related to politics, values etc.)

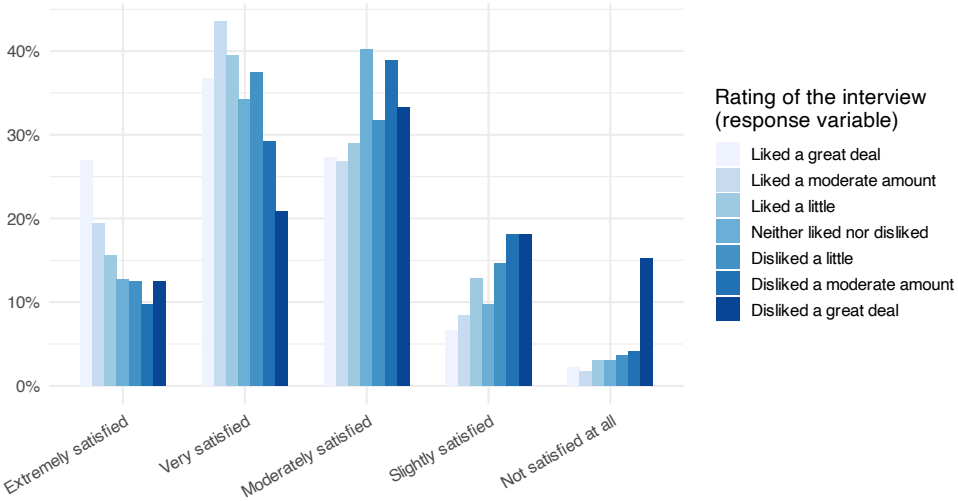
Response measure: !!! Little variability !!! ☒



Response measure: Fine. ☑

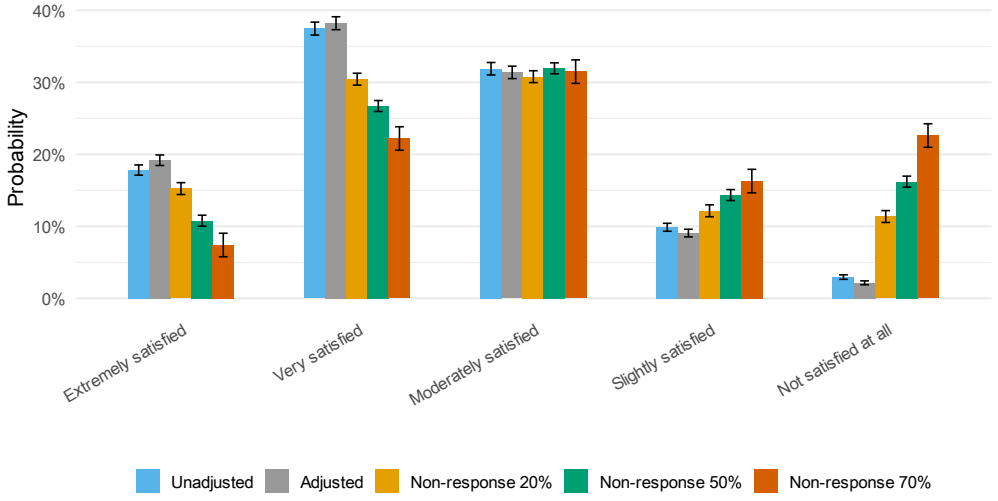


How satisfied are you with life?

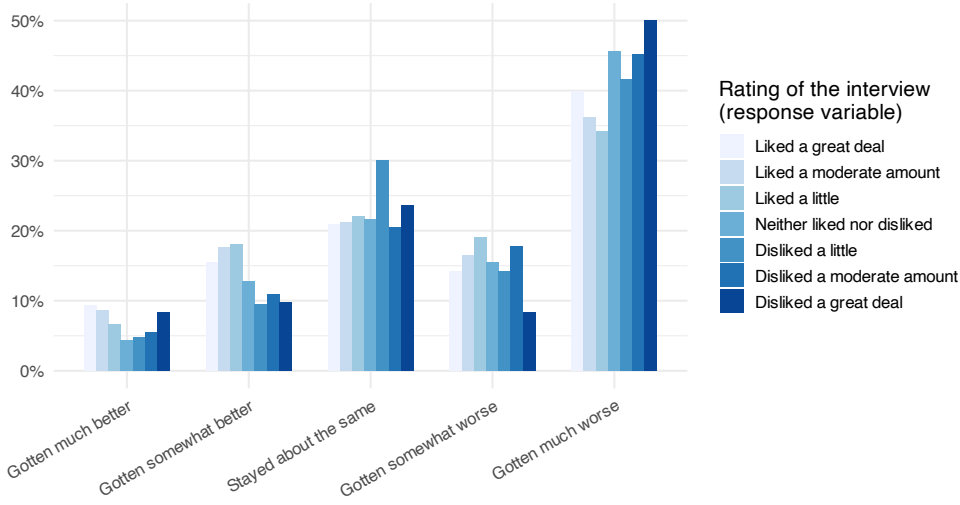


How satisfied are you with life?

(ρ = 0.414, ρ = 0.491, ρ = 0.548)

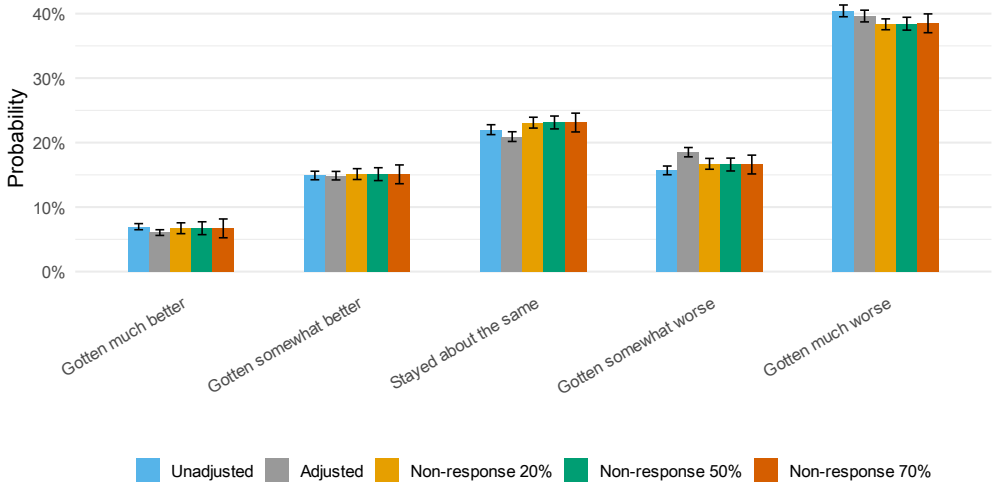


National economy has gotten better or worse?

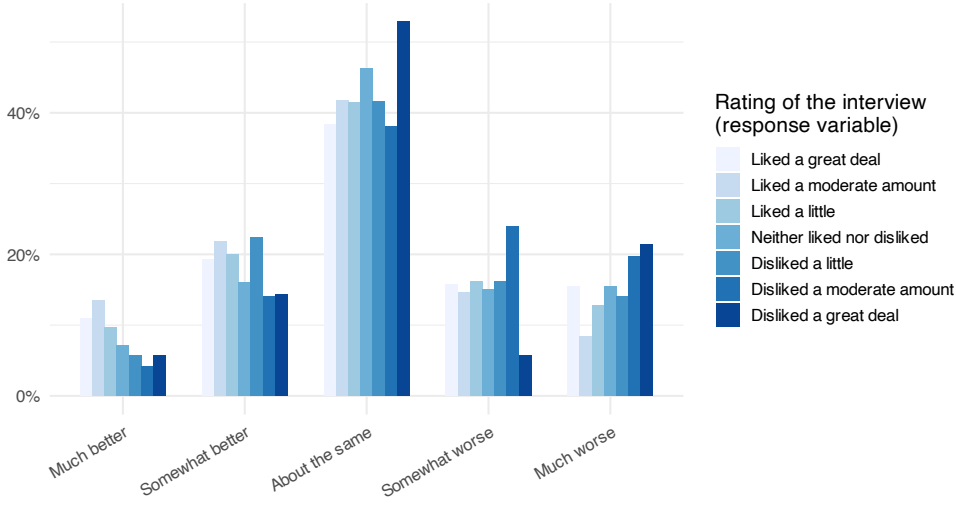


National economy has gotten better or worse?

(ρ = -0.008, ρ = 0.001, ρ = 0.002)

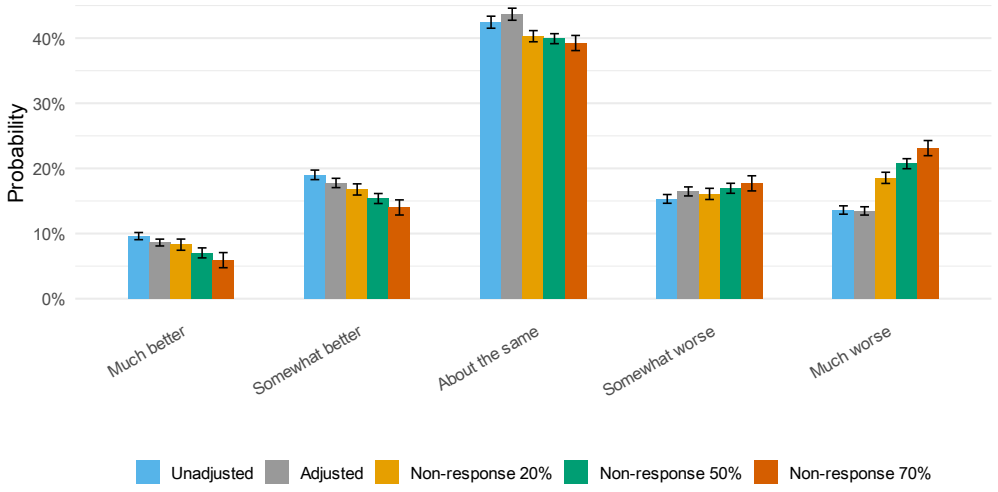


Unemployment is better or worse than last year?

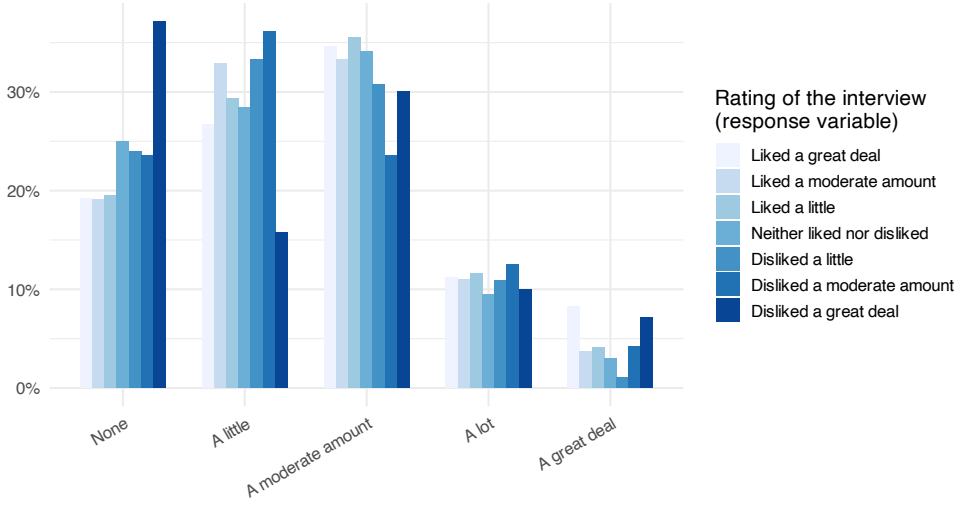


Unemployment is better or worse than last year?

(ρ = 0.151, ρ = 0.186, ρ = 0.211)

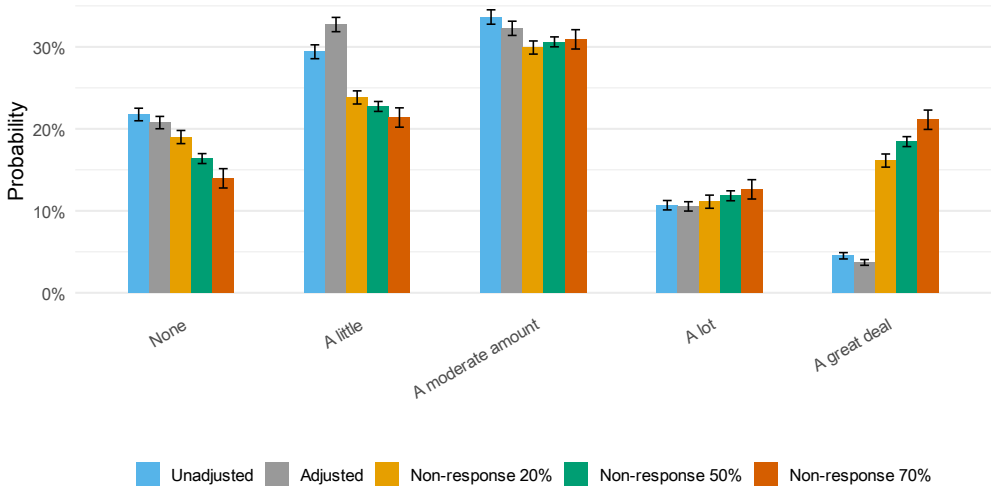


How much trust and confidence do you have in news?



How much trust and confidence do you have in news?

(ρ = 0.198, ρ = 0.224, ρ = 0.253)



Conclusion

Conclusion

What we have:

- extension of Peress (2010) for ordinal outcome variables
- that is: parametric model for outcome and response that may reduce non-response bias
- derived likelihood and standard errors
- empirical illustration on American National Election Studies data (Feb 2025)
- `R code` of the implementation

What is left to do (?)

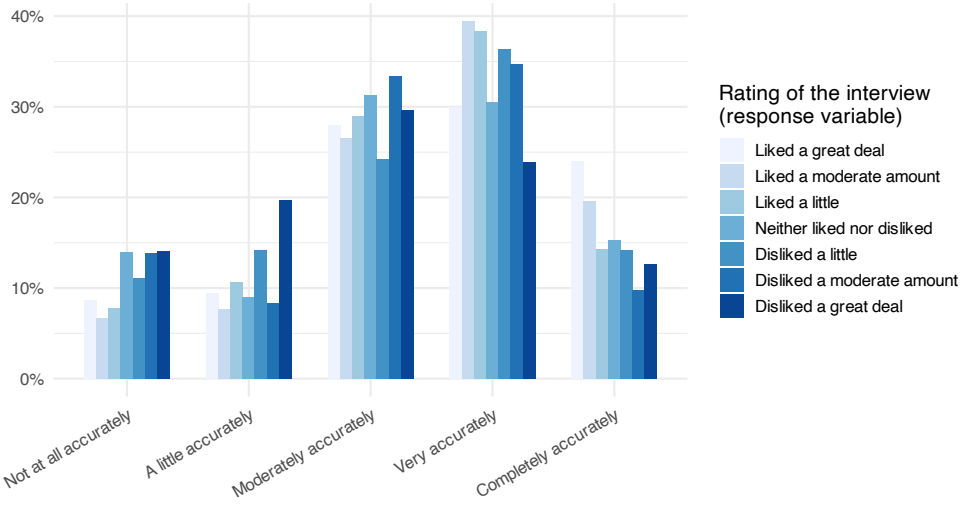
- simulations
- other measures for response propensity
- performance benchmark
- marketing

Thank you.

www.lukaslaffers.com

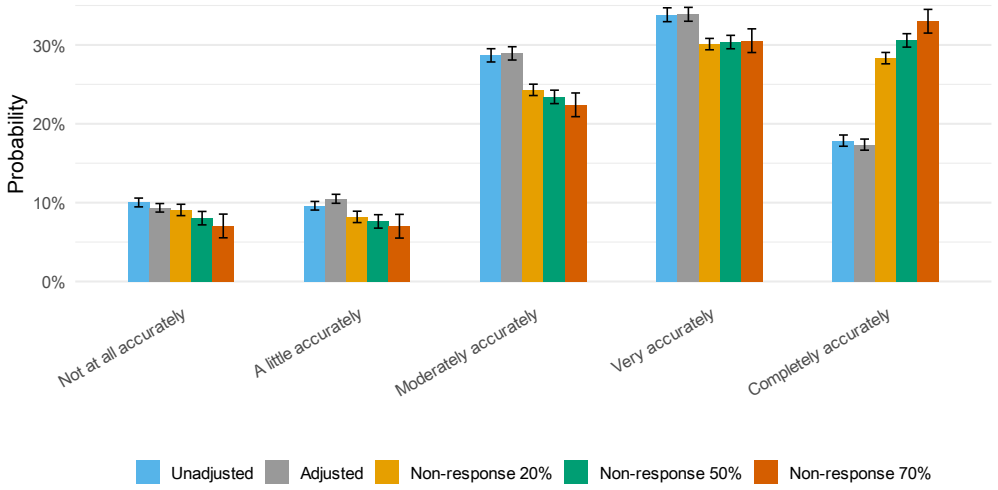
Additional figures

How accurately do you think the votes will be counted?

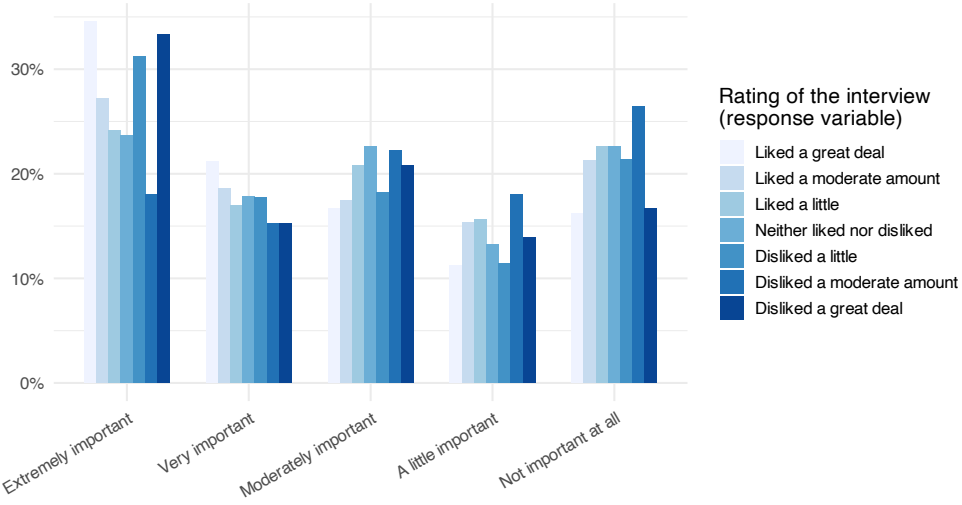


How accurately do you think the votes will be counted?

(ρ = 0.135, ρ = 0.15, ρ = 0.17)

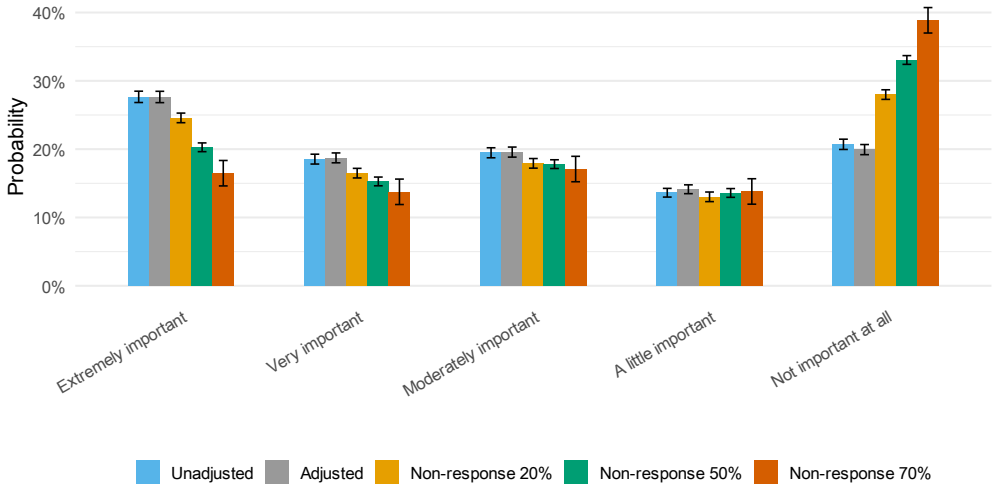


Is religion an important part of your life?

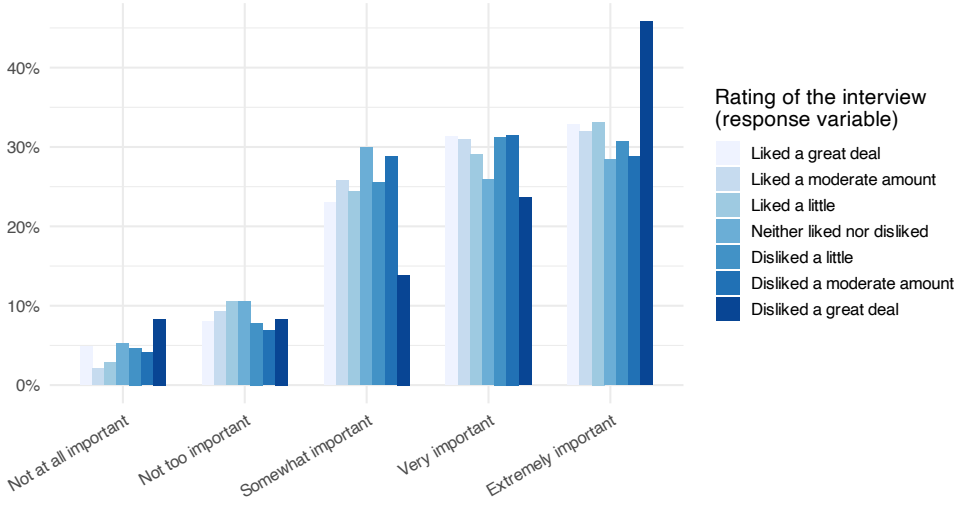


Is religion an important part of your life?

(ρ = 0.257, ρ = 0.316, ρ = 0.363)



Importance of abortion issue.



Importance of abortion issue.

(ρ = 0.072, ρ = 0.076, ρ = 0.085)