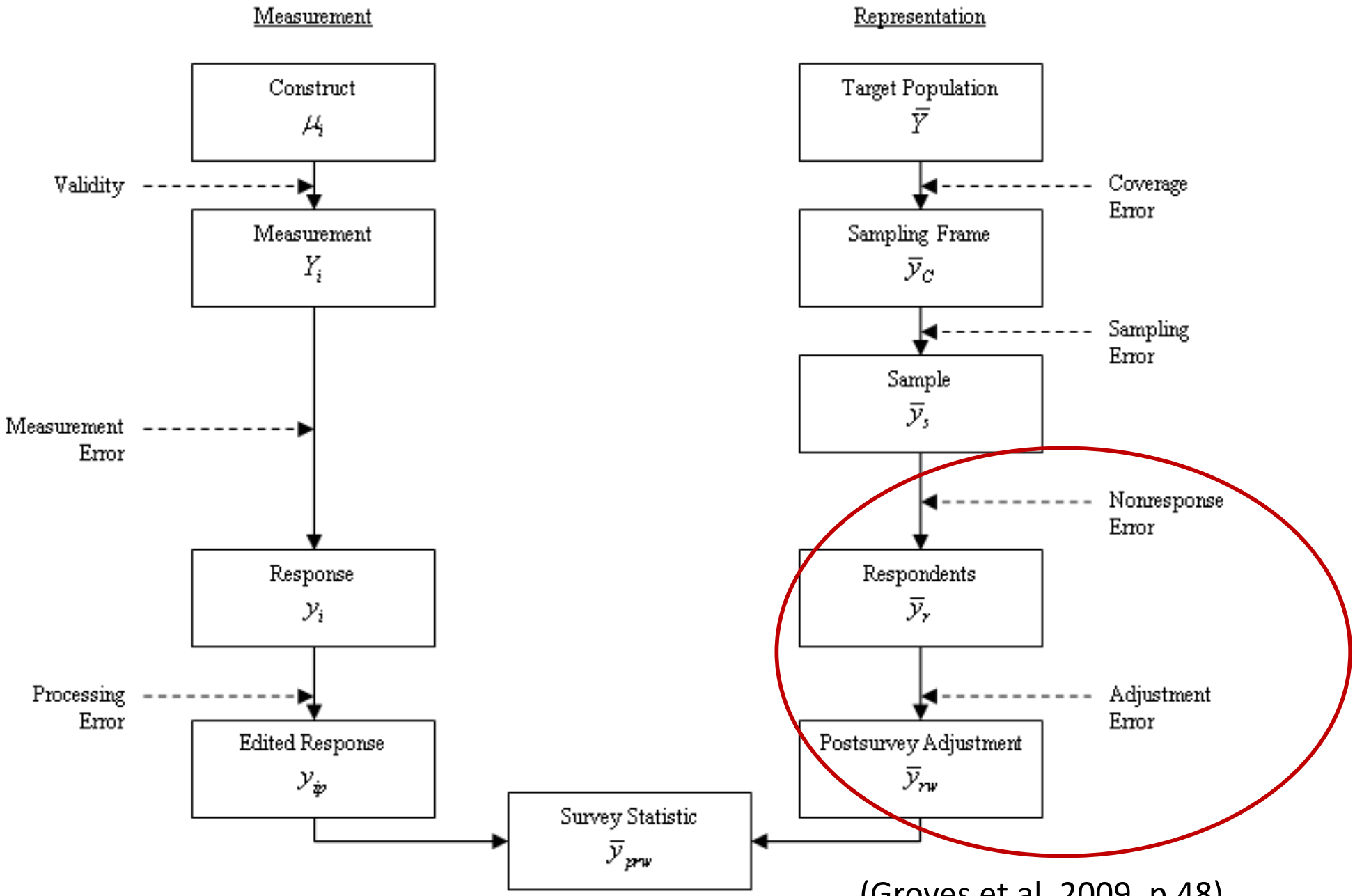
Survey data analysis
Week 45:
“Designing weights”

Literature today

- Exercise last week
- Bias-variance trade-off
  - Adjustment error
- Paradata
  - Short lecture
  - How to design and select covariates for weights?
- Issues in weighting
  - Brick (2013)
  - Class discussion
- Exercise on calibration, raking

Total Survey Error (TSE) Framework



(Groves et al. 2009, p.48)

Exercise nonresponse weights

- Discussion
  - Is weighting succesful?
  - What happens to bias?
    - MAR or MNAR?
  - What happens to standard errors

Bias-variance trade-off

Bias

- Sample mean $\overline{y} = \frac{1}{n}\sum^{n} y_i$
- Weighted mean $\overline{y_w} = \frac{1}{n}\sum^{n} w_i y_i$

Variance

- If there is no correlation between survey weights and the characteristic to be estimated, maximum increase in variance of the mean is (Kish, 1965):

$$UWE = 1 + cv^2(w_i)$$

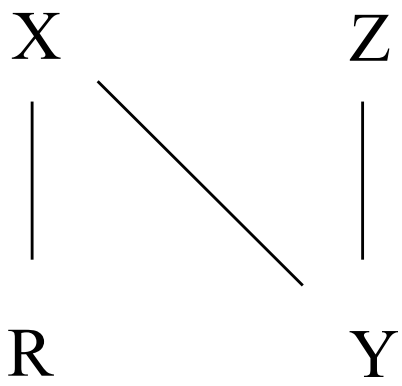
$$cv = standard\ deviation(w_i)/mean(w_i)$$

Succesful NR weighting

– When (NR) weighting is succesful

- Little and Vartivarian (2005)

	X -> R weak	X -> R strong
X-> Y weak	Nothing happens	Still bias Variance inflation
X -> Y strong	Bias somewhat reduced	Bias reduced Variance inflation



MAR

Bias-variance trade-off(2)

- To avoid the inflation of variance
  - Outlier trimming
  - Trimming weights >3 to 3
  - Propensity score weighting
    - Use 5 bands of propensity scores

Designing weights

- Sample level information
  - Location, gender, age, ….
  - ?
- Population level information
  - LOTS of potential variables
  - Gender, age, education, ethnicity, region, income
  - Membership of union, newspaper readership, politically active, visited the Efteling, etc,

Designing weights

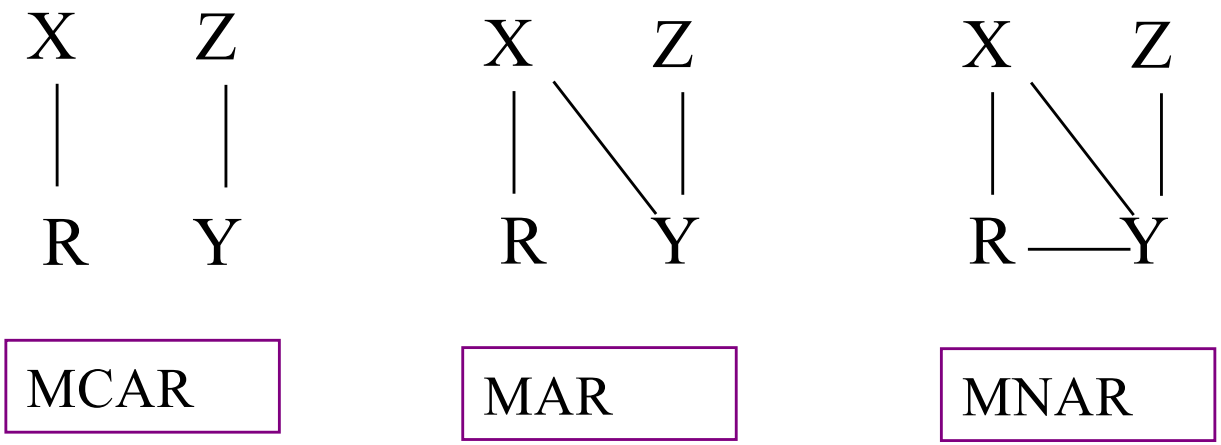
- Sample level information
  - Location, gender, age, ….
  - ? ---- paradata
- Population level information
  - LOTS of potential variables
  - Gender, age, education, ethnicity, region, income
  - Membership of union, newspaper readership, politically active, visited the Efteling, etc,

Paradata

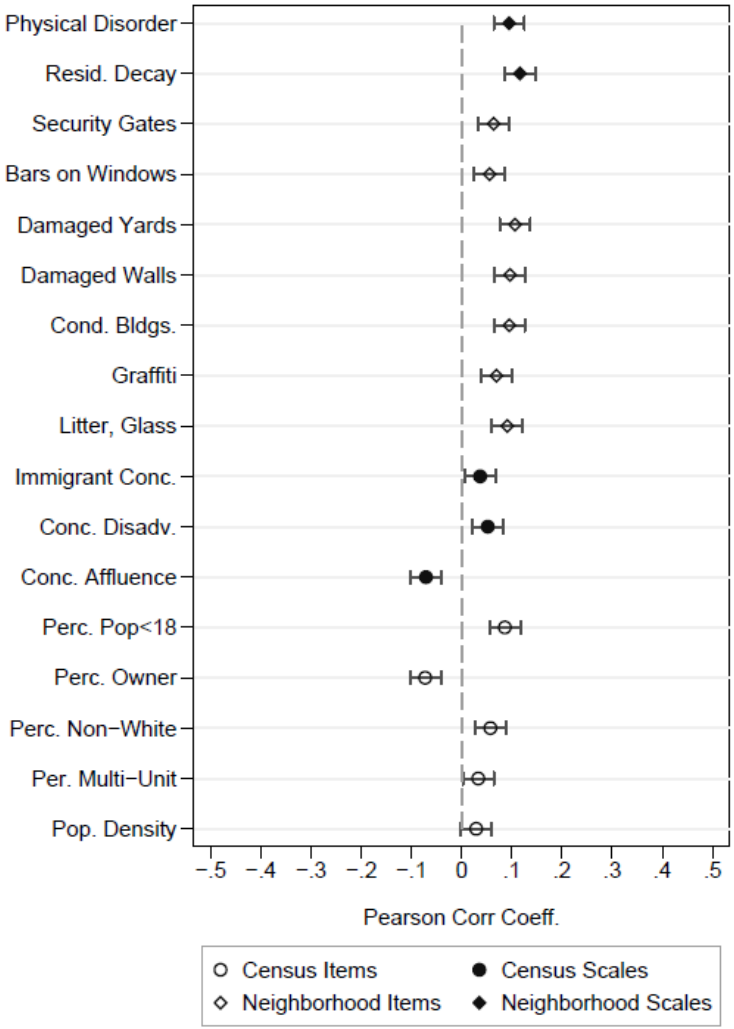
- By-product of doing research
- Surveys
  - F2f: Interviewer observations
    - Or Recordings
  - CATI: Call record data
  - WEB: Browser data, Response timings, section timings, evaluation questions, etc.

Paradata – so what?

- Interviewer observations
  - Useful in nonresponse corrections
    - Strong link X -> Y



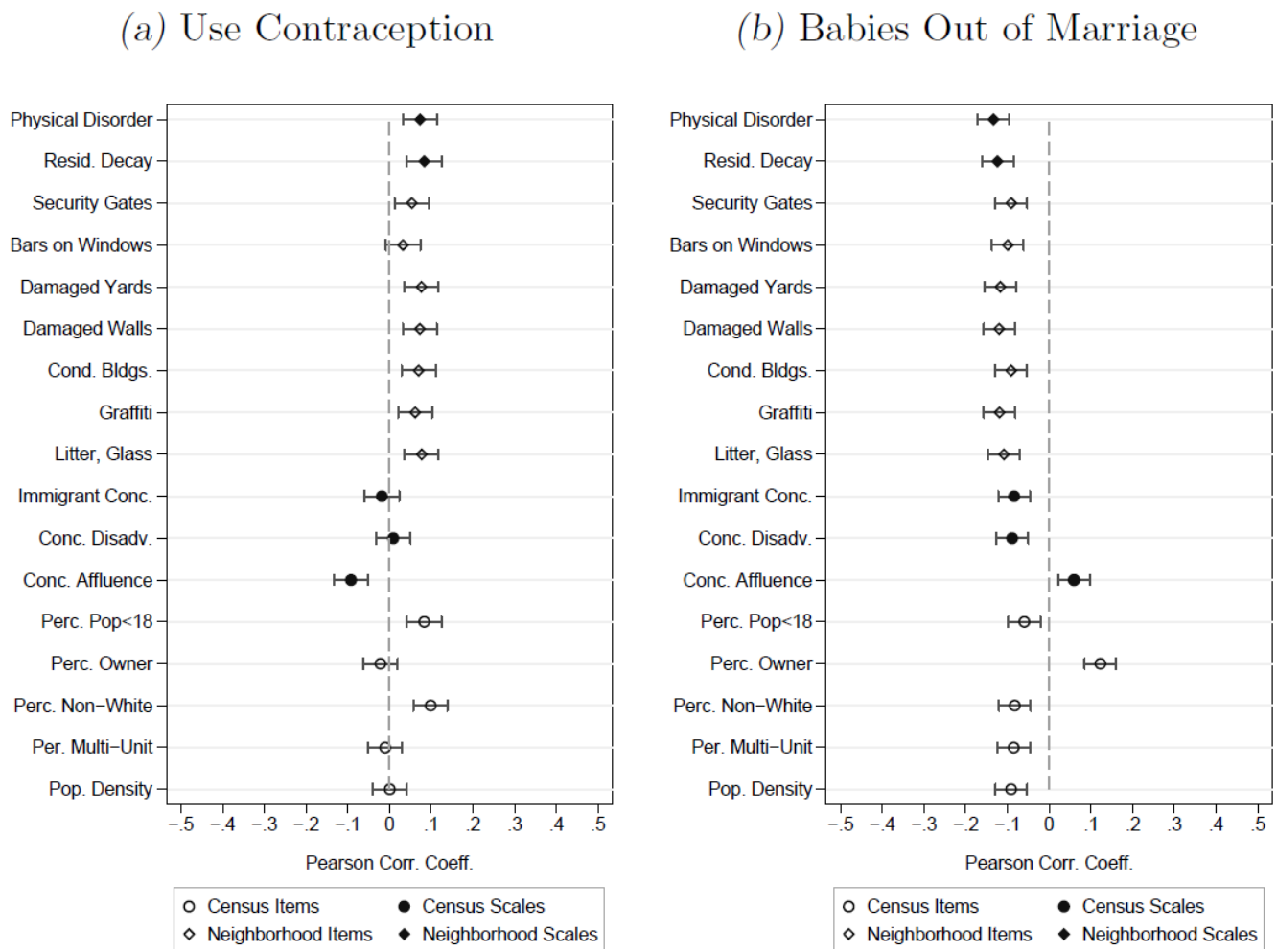
Interviewer observations example Casas-Cordero (2010)



Relations X -> R.
Bivariate (logistic) relations rather weak,
… but strong link with Y?

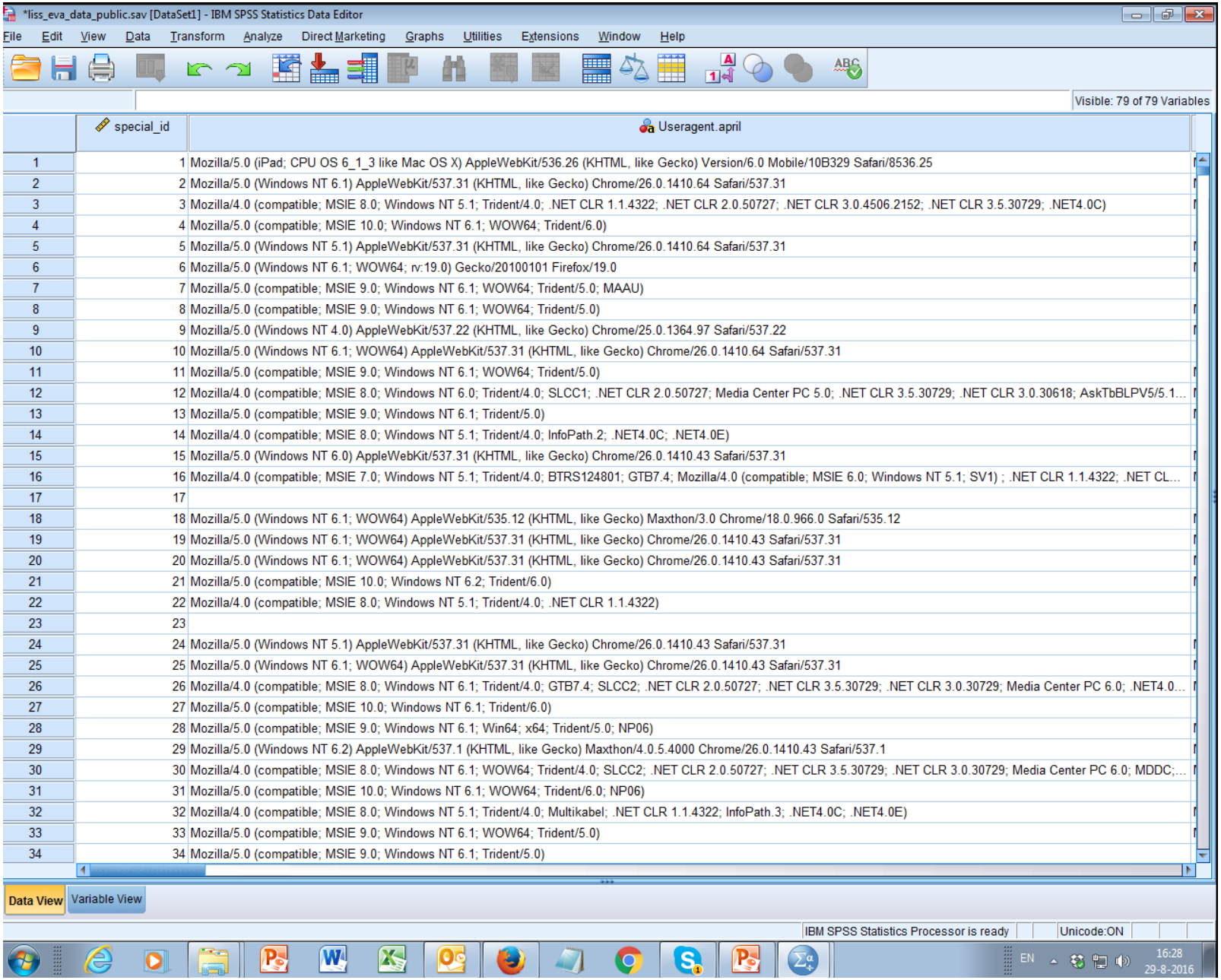
- Figure taken from Casas-Cordero (2010)

Interviewer observations (2)

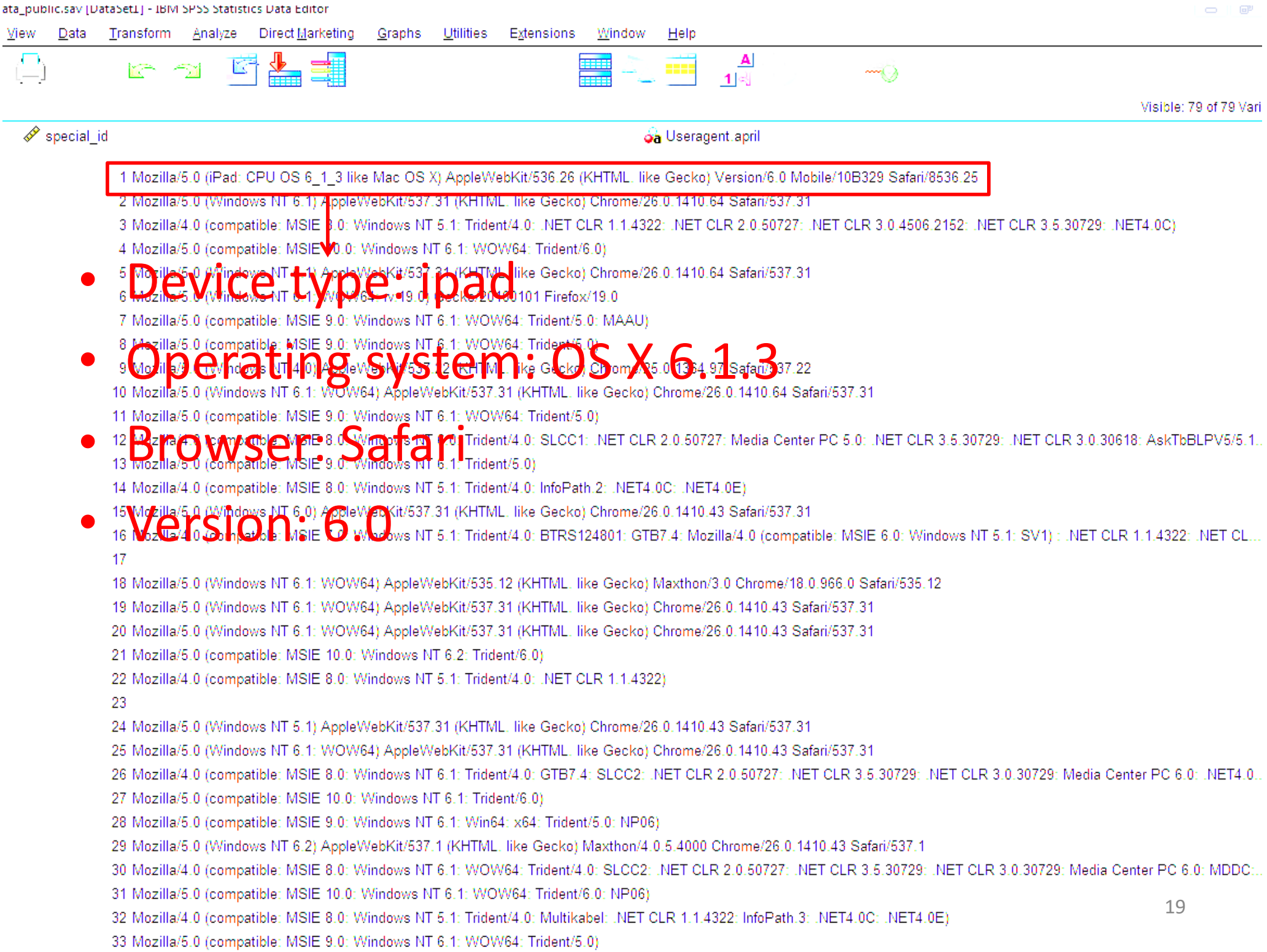


- Relation X-> Y. Figure taken from Casas-Cordero (2010)

Paradata – web



User Agent strings:

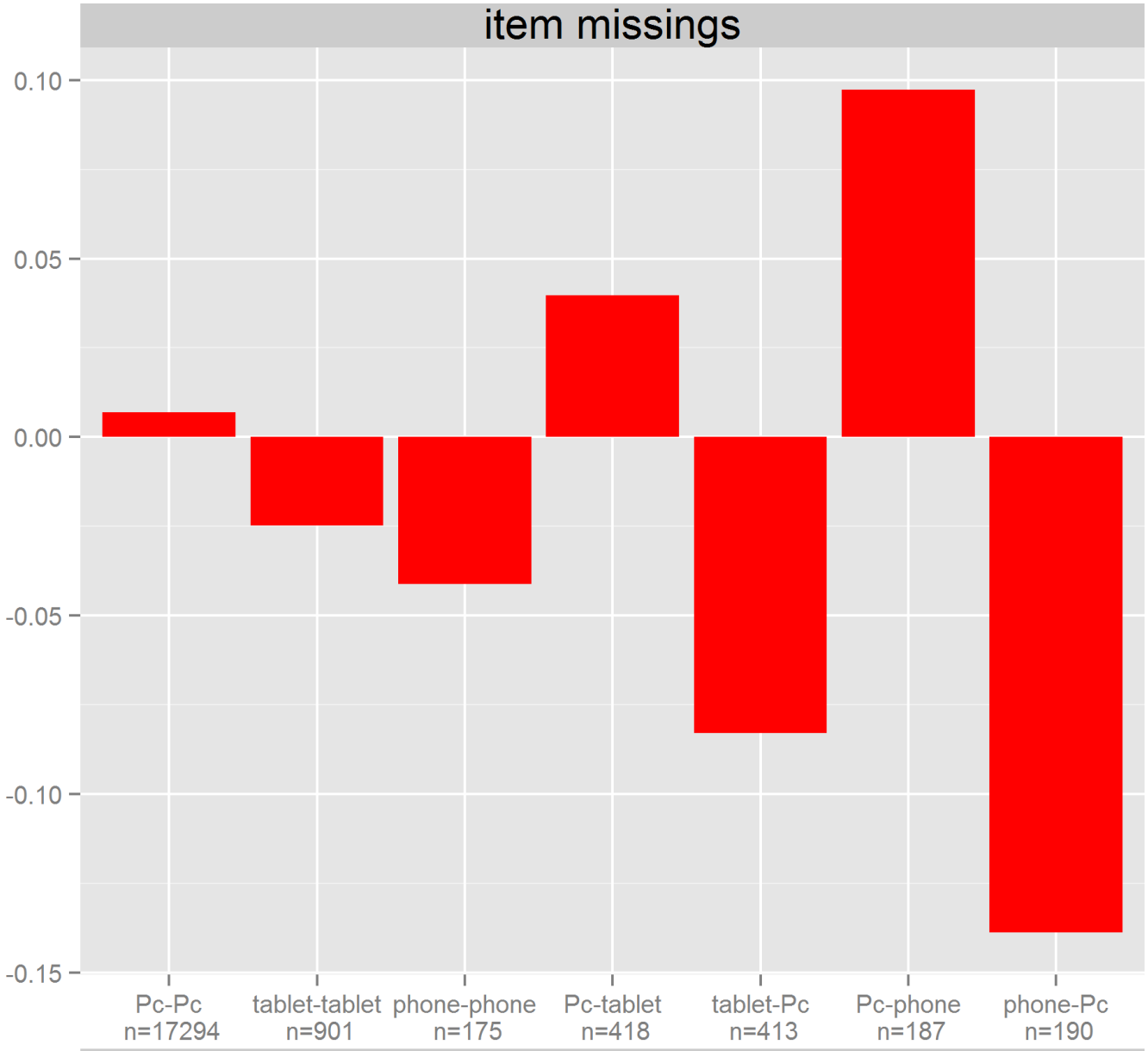


- Device type: ipad
- Operating system: OS X 6.1.3
- Browser: Safari
- Version: 6.0

User agent strings
Example by Lugtig and Toepoel (2015)

% Device used in month (t)		% Device used in May						
		PC	Tablet	Phone	No participation		N	% of wave respondents
April	PC	77.4	1.1	0.4	21.1		4966	90.2
	Tablet	19.3	24.4	0.2	56.1		435	7.8
	Phone	33.9	4.6	30.3	31.2		109	2.0
	No participation	37.6	2.1	0.8	59.4		715	-

User agent strings
Example by Lugtig and Toepoel (2)



Paradata +++

- Taken from Toepoel and Lugtig (2014)
  - more in week “designed big data”

How to use paradata

- Scheduling what to do next -> reduce costs
  - Call scheduling, interviewer visits
- Planning next wave of data collection
  - Longitudinal, or repeated surveys
- For corrections!
  - Use interviewer observations to weight (or impute)
  - Use type of nonresponse and make separate models.
    - Noncontact <- age + #hh members + work
    - Refusal <- level of education + gender

Brick (2013)

- Review of weighting approaches
  - Propensity score models
    - Sampling frame data is limited
  - Population information is limited
    - Other weighting models
- Can we use more information?
  - Section 7: paradata?

Brick (2013)

- Review of weighting approaches
- RHG – response homogeneity groups
  - Groups of propensity score models
- Responsive design models
  - Adjust fieldwork efforts so that $P_{respond}$ = equal
  - Var ($p_{respond}$) = 0

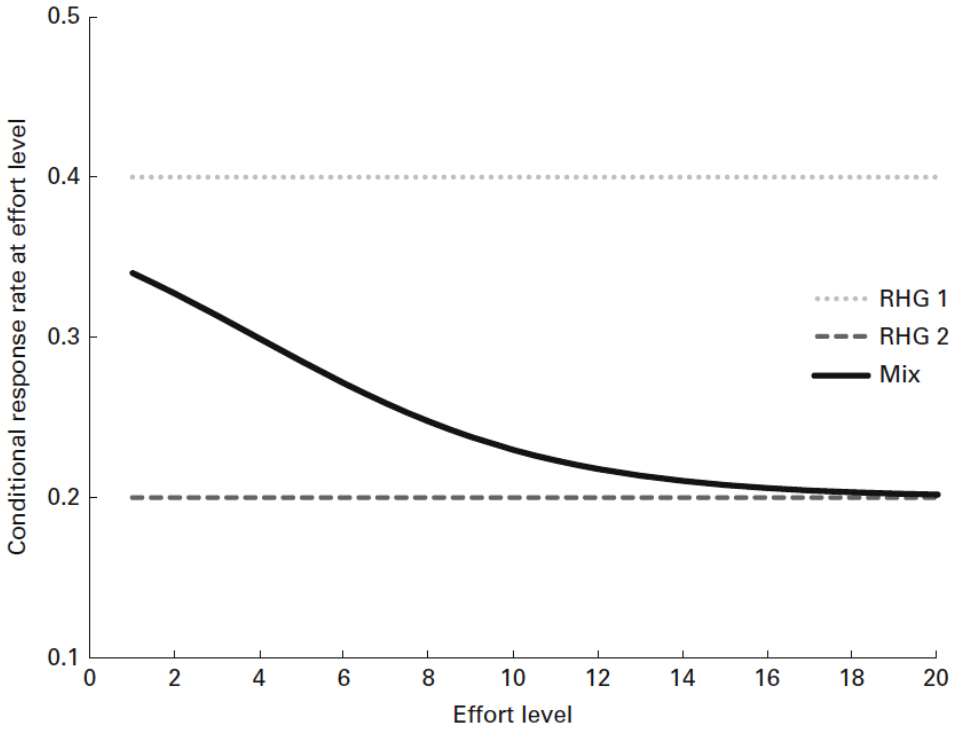


Fig. 1. Observed response propensities for a sample composed of two RHGs

Designing weights in R

1. Design your study such that
- You get a rich frame (paradata)
- You ask about known population statistics that predict Y

2. Find variables (x) that predict both R and Y
- propensity score weighting
- Poststratify
- calibrate (linear weighting)
- Rake

Class exercise

- Website of Statistics Netherlands: statline
  - https://opendata.cbs.nl/statline/#/CBS/en/
- Find auxiliary variables for your scenario
  - Should predict your Y
  - … and nonresponse
- 15 minutes in groups of 4

How to use variables in R…

- ‘survey’ library
- Imagine: auxiliary data for sex, age
  - PostStratify(design= svy.unweighted, strata = agegender, population=agegender.dist
  - Calibrate(design=svy.unweighted, formula = ~age+gender , population=c(sex.dist,age.dist)
  - rake(design = svy.unweighted, sample.margins = list(~sex, ~age), population.margins = list(sex.dist, age.dist))

Next week

- Finish exercise on creating weights
- Two weeks on imputation (by Stef van Buuren)
  - readings

Now: work on exercise

- Exercise on creating your own weights
  - Poststratification
  - Propensity weighting
  - Raking

- All with the survey package