Multiverse Analysis

Reinhard Schunck
Nora Huth-Stöckle

BERGISCHE
UNIVERSITÄT
WUPPERTAL

0. Disclaimer

- What not to expect: Tutorial
- Instead: Overview

- Suggested readings to start with:
  - Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208-1214.
  - Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.
  - Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*(1), 3-40.
- Stata resources:
  - mrobust
  - multivrs
  - https://hbs-rcs.github.io/post/specification-curve-analysis/
  - Do-files from Uri Simonsohn
  - …

1. Introduction

1. Introduction
2. Data-Analytic Decisions
3. Example
4. Results
5. Discussion

1. Introduction

"Anyone using a parametric statistical technique for long enough (and it does not take very long) will recognize the difficulty of choosing which of hundreds of possible regressions to present in a written work. This choice is difficult, fraught with ethical and methodological dilemmas, and not covered in any serious way in classical statistics texts. Parametric methods merely assume that we know the correct specification. In practice, the 'correct' specification is chosen after looking at the estimates, and so it is never clear to a reader whether an article is a true test of a hypothesis, in the sense that the author was vulnerable to being proved wrong, or whether the article is merely a proof of the existence of at least one specification consistent with the author's favored hypothesis." (Ho et al. 2007)

1. Introduction

- Crisis in science

  … published findings cannot be replicated

1. Introduction

- Crisis in science (Young 2018)

1. Model uncertainty

2. Lack of transparency

1. Introduction

- Crisis in science (Young 2018)

1. Model uncertainty:
   How robust are our results?

2. Lack of transparency:
   How can we increase transparency in (statistical) research?

1. Introduction

- One answer: Multiverse / specification curve / multimodel analysis
  (Ho et al. 2007; Simonsohn et al. 2020; Steegen et al. 2016; Young 2015)

- Related: many analysts approach
  (Breznau et al. 2022; Silberzahn et al. 2018)

1. Introduction

Instead of presenting a single analysis (and an arbitrary selection of robustness checks):

- Present “*all*” valid and non-redundant specifications.
- Make the data-analytic decisions transparent.

1. Introduction

- … not really new:

I Just Ran Two Million Regressions

Xavier X. Sala-I-Martin

The American Economic Review, Vol. 87, No. 2, Papers and Proceedings of the Hundred and Fourth Annual Meeting of the American Economic Association (May, 1997), pp. 178-183 (6 pages)

https://www.jstor.org/stable/2950909

- and its methodological or philosophy of science background even less so.

2. Data-Analytic Decisions

- The problem: unsystematic approach to decisions in the research process.
  - Weak link between theory and design / empirical specification.
  - Numerous ways to approach a research question.
  - Myriad *data-analytic decisions*, e.g.,
    - Design: experimental / observational
    - Population / sampling
    - Operationalization
    - Data preparation
    - Covariate selection
    - Model selection
    - …

2. Data-Analytic Decisions

- Some decisions are unambiguous (directly derivable from RQ / theory).

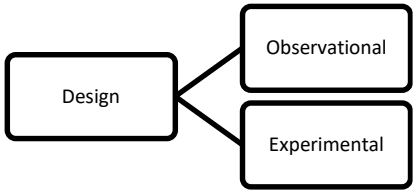
- Some decisions are arbitrary:

  → *Researcher's degrees of freedom* (Simmons et al. 2011)

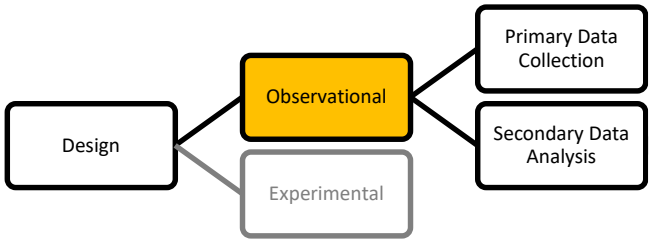
  → *Garden of forking paths* (Gelman & Loken 2014)

2. Data-Analytic Decisions

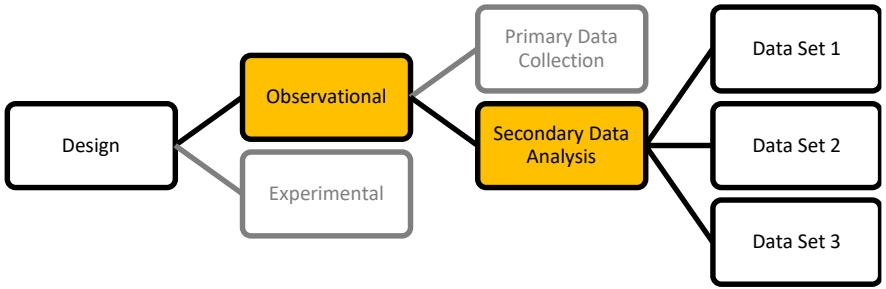
- An (incomplete) illustration:



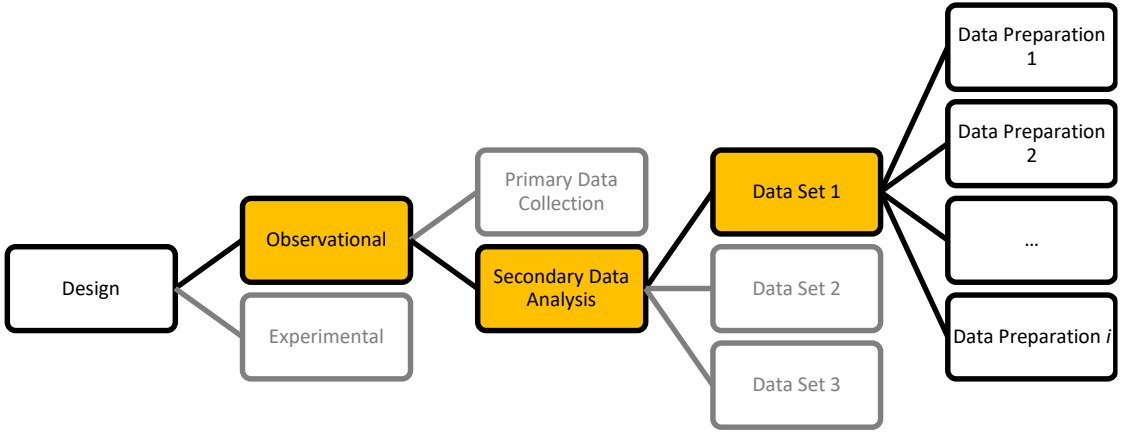
2. Data-Analytic Decisions



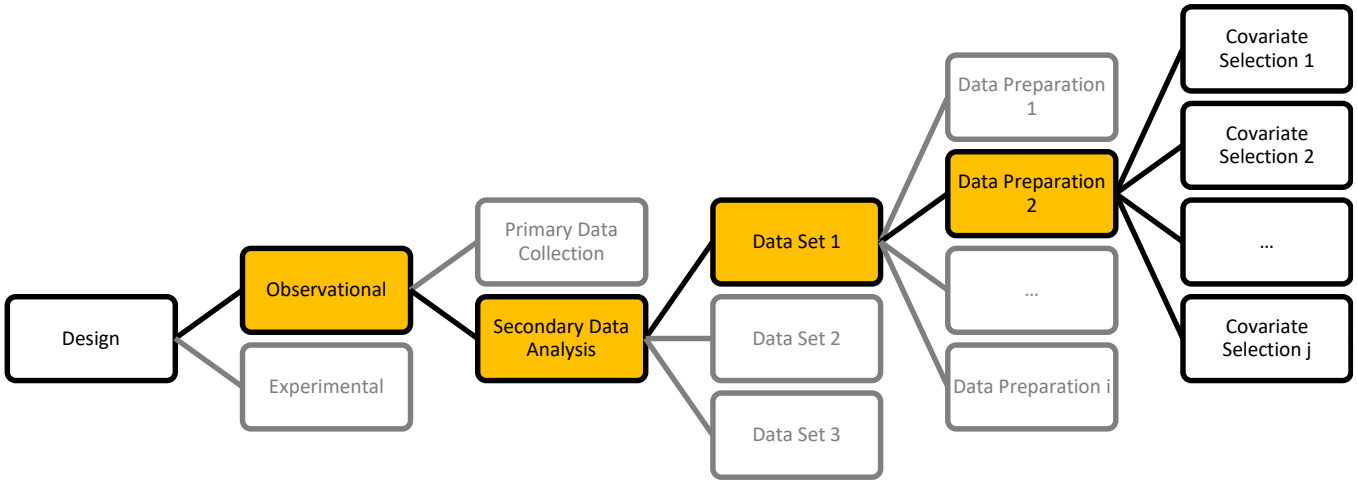
2. Data-Analytic Decisions



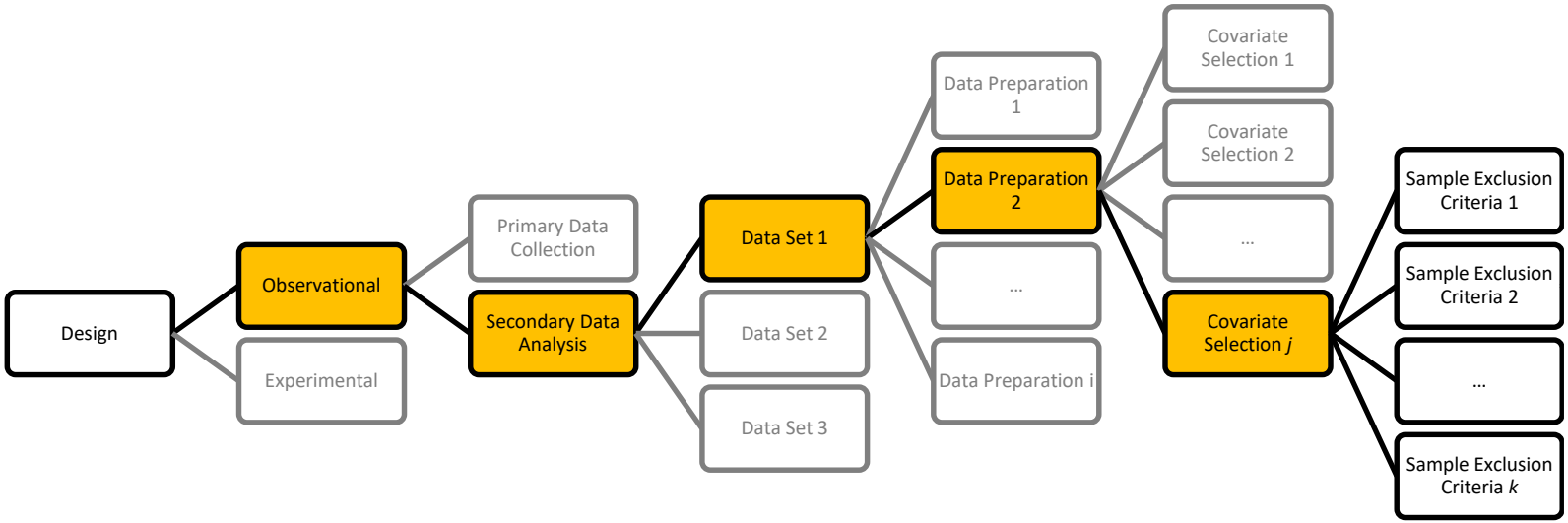
2. Data-Analytic Decisions



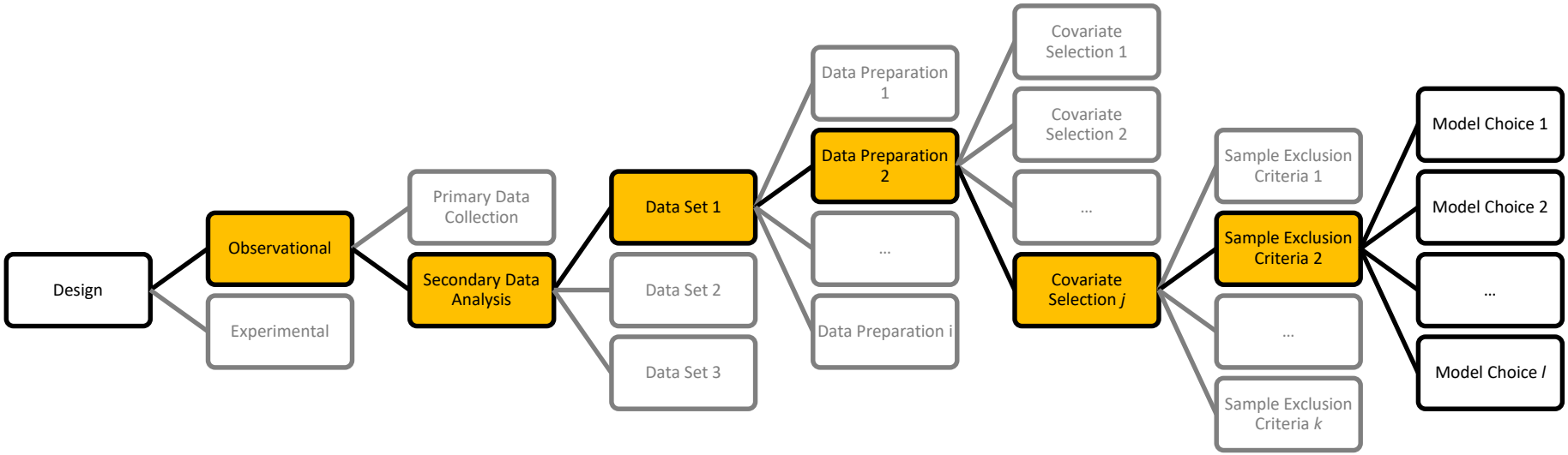
2. Data-Analytic Decisions



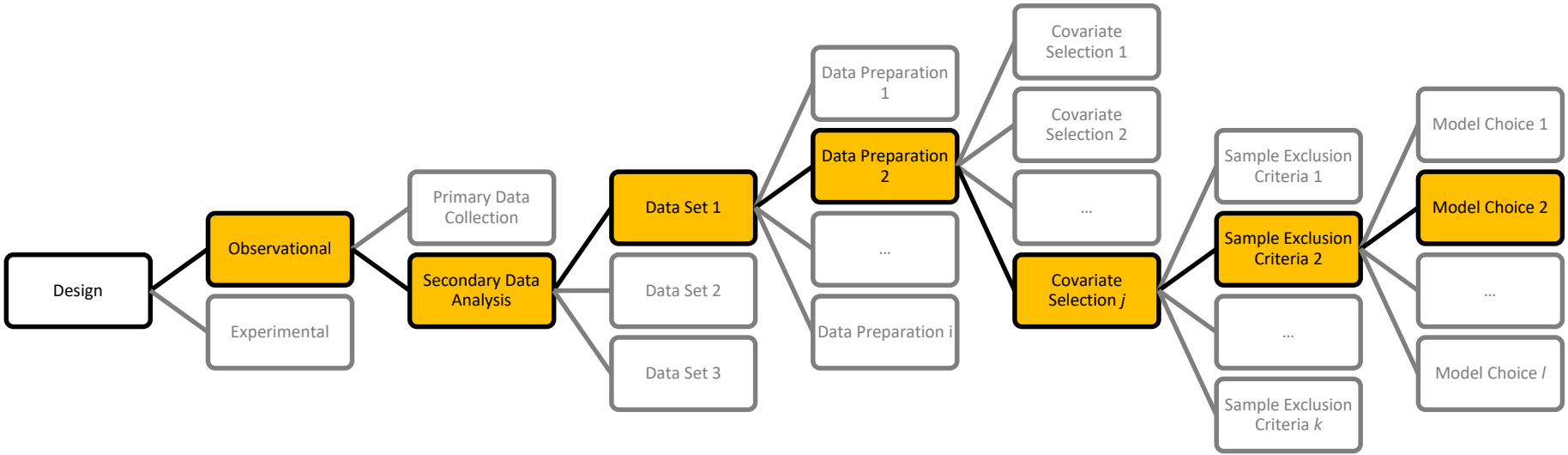
2. Data-Analytic Decisions



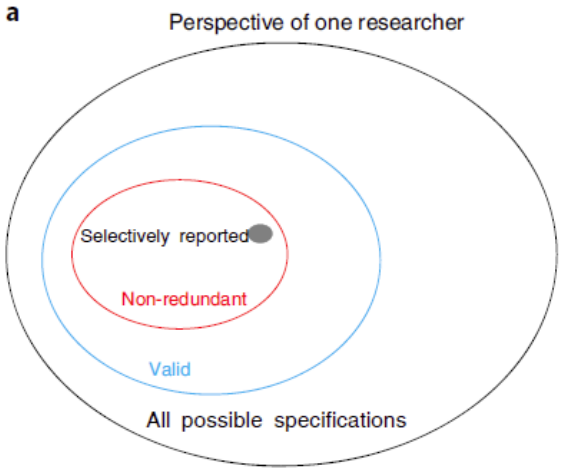
2. Data-Analytic Decisions



2. Data-Analytic Decisions

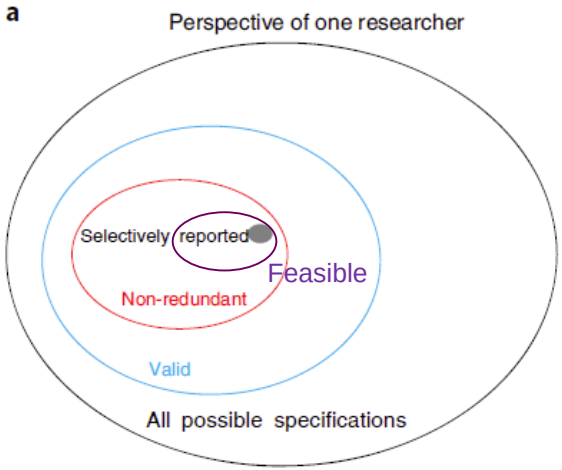


2. Data-Analytic Decisions



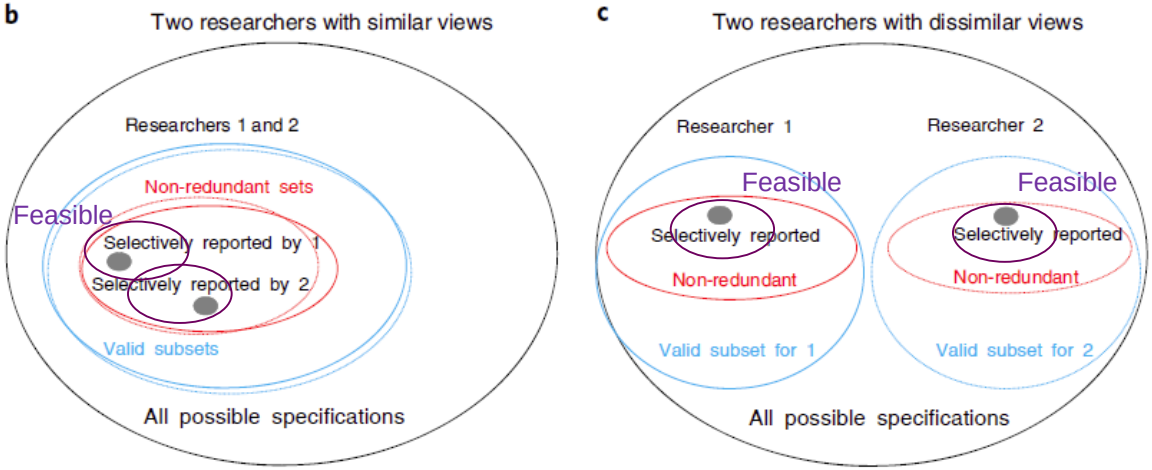
(modified from Simonsohn et al. 2020)

2. Data-Analytic Decisions



(modified from Simonsohn et al. 2020)

2. Data-Analytic Decisions



(modified from Simonsohn et al. 2020)

3. Example

- Pretty Integrated project (DFG 447581390)
- RQ: Do returns to physical attractiveness in the German labor market vary by ethnicity and gender?

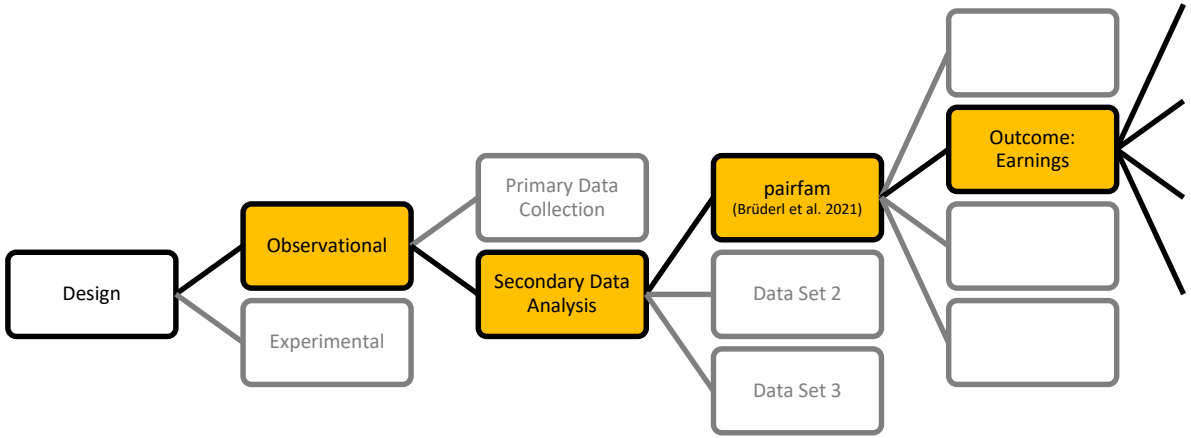
3. Example

- Pretty Integrated project (DFG 447581390)
- RQ: Do returns to physical attractiveness in the German labor market vary by ethnicity and gender?

- Disclaimer I: only part of the multiverse.
- Disclaimer II: only part of the results.

3. Example

- Pretty Integrated project (DFG 447581390)
- RQ: Do returns to physical attractiveness in the German labor market vary by ethnicity and gender?



3. Example

- Data-analytic decisions:

1. Outcome: Earnings
    1. Gross hourly earnings (preferable, only collected biennially)
    2. Net hourly earnings (annually)
    3. Log (Mincer-type equation, production function: multiplicative) (Gelman & Hill 2006)
    4. Untransformed
    5. Trimmed (implausible values)
    6. Not trimmed

3. Example

- Data-analytic decisions:

2. Treatment / independent variable: interviewer rated physical attractiveness
    1. Continuous (7-point scale)
    2. Binary (very attractive vs. rest)

3. Example

- Data-analytic decisions:

3. Group membership
    1. Migration background (generational status)
    2. Ethnicity (German, Ethnic-Germen, Turkish, interethnic, other)

3. Example

- Data-analytic decisions:

4. Exclusion criteria I: Problematic “interviewers”
    1. Exclude interviewers with too little variance in attractiveness ratings
    2. Include

3. Example

- Data-analytic decisions:

5. Exclusion criteria II: Refreshment sample
    1. Exclude refreshment sample (Covid pandemic, mixed mode)
    2. Include

3. Example

- Data-analytic decisions:

6. Exclusion criteria III: Age
    1. Exclude resp. < 18
    2. Include

3. Example

- Data-analytic decisions:

7. Additional covariates
    1. Exclude
    2. Include

    – Personality: confounder / mediator
    – Proxy for cog. ability: measurement error independent and nondifferential (Hernan & Cole 2009)

3. Example

- Data-analytic decisions:

7. Additional covariates
    1. Exclude
    2. Include

    – Personality: confounder / mediator
    – Proxy for cog. ability: measurement error independent and nondifferential (Hernan & Cole 2009)

    – Disclaimer: models are probably still misspecified

3. Example

- Data-analytic decisions:

8. Modelling: Regression models (other approaches possible)
   - Multilevel data: Interviewers – respondents – occasions
   - Not all combinations possible:

3. Example

- Data-analytic decisions:

8. Modelling: Regression models (other approaches possible)
   - Multilevel data: Interviewers – respondents – occasions
   - Not all combinations possible:
     1. Interviewer and respondents random effects, cluster robust s.e.
        - respondent f.e. not possible
        - weighting not possible

3. Example

- Data-analytic decisions:

8. Modelling: Regression models (other approaches possible)
   - Multilevel data: Interviewers – respondents – occasions
   - Not all combinations possible:
     1. Interviewer and respondents random effects, cluster robust s.e.
     2. Interviewer fixed effects, three-way cluster robust s.e.
        | weighted (calibrate design weights)
        1. IPW: female labor force participation (Hernan & Robins 2020)
        2. No IPW

3. Example

- Data-analytic decisions:

8. Missing values
   - Listwise deletion (computational time: “fast” 30 hours)
   - …

3. Example

- Data-analytic decisions:
  - Female respondents: 1536 specifications
    - 768 migration background
    - 768 ethnicity
  - Male respondents: 1024 specifications
    - 512 migration background
    - 512 ethnicity

3. Example

- Regression models – marginal effects:

$$y_{ijt} = \beta_1 + \beta_2 Attr_{ij} + \beta_3 Mig_{ij} + \beta_4 Attr_{ij} Mig_{ij} + \varepsilon_{ijt}$$

$$\frac{\partial y_{ijt}}{\partial Attr_{ij}} \left[\beta_1 + \beta_2 Attr_{ij} + \beta_3 Mig_{ij} + \beta_4 Attr_{ij} Mig_{ij} + \varepsilon_i\right] = \beta_2 + \beta_4 Mig_{ij}$$

4. Results

- Presenting the results (exemplary)

4. Results

- Presenting the results
  - Summary statistics (Simonsohn et al. 2020; Young & Holsteen 2017)
  - Influence analysis (Young & Holsteen 2017)
  - Specification curves (Simonsohn et al. 2020)
  - Inference … (Simonsohn et al. 2020)

4. Results

- Presenting the results
  - Summary statistics (Simonsohn et al. 2020; Young & Holsteen 2017)
    - Median effect size
    - Median p-value
    - Sign stability
    - …

4. Results

- Presenting the results
  - Summary statistics (Simonsohn et al. 2020; Young & Holsteen 2017)
    - Median effect size
    - Median p-value
    - Sign stability
    - Robustness ratio (Young & Holsteen 2017)

$$rr = \frac{b_{preferred}}{\sigma_T} \text{ or } rr = \frac{\bar{\bar{b}}}{\sigma_T}$$

$$\sigma_T = \sqrt{\frac{1}{KJ}\sum_{k=1}^{K}\sum_{j=1}^{J}(b_{kj} - \bar{\bar{b}})}$$

with $\{S_1, \ldots, S_k\}$ samples and $\{M_1, \ldots, M_j\}$ models

4. Results

- Presenting the results
  - Summary statistics (Simonsohn et al. 2020; Young & Holsteen 2017)
    - Median effect size
    - Median p-value
    - Sign stability
    - Robustness ratio (Young & Holsteen 2017)
      - Constructed analogous to t-statistic
      - But: underlying statistical properties unknown; depend on specified model space.

4. Results

Women: Median effect sizes and p-values (768 specifications)

		Median b
Log hourly earnings	W/o migration background	0.03
	Ethnic-German (Aussiedler)	-0.03
	Interethnic (“Half-German”)	0.01
	Turkish	0.09
	Other non-German	-0.03

4. Results

Women: Median effect sizes and p-values (768 specifications)

		Median b
Log hourly earnings	W/o migration background	0.03
	Ethnic-German (Aussiedler)	-0.03
	Interethnic (“Half-German”)	0.01
	Turkish	0.09
	Other non-German	-0.03

b = marginal effect
Log scale: 0.03 ~ 3% higher earnings

4. Results

Women: Median effect sizes and p-values (768 specifications)

		Median b	Median $p-value$	Sign stability
Log hourly earnings	W/o migration background	0.03	0.00	1.00
	Ethnic-German (Aussiedler)	-0.03	0.37	0.88
	Interethnic (“Half-German”)	0.01	0.51	0.65
	Turkish	0.09	0.06	1.00
	Other non-German	-0.03	0.33	0.80

4. Results

Women: Median effect sizes and p-values (768 specifications)

		Median b	Median $p-value$	Sign stability	Robustness ratio
Log hourly earnings	W/o migration background	0.03	0.00	1.00	1.46
	Ethnic-German (Aussiedler)	-0.03	0.37	0.88	-0.85
	Interethnic (“Half-German”)	0.01	0.51	0.65	0.41
	Turkish	0.09	0.06	1.00	1.46
	Other non-German	-0.03	0.33	0.80	-0.71

4. Results

Women: Median effect sizes and p-values (768 specifications)

		Median b	Median $p-value$	Sign stability	Robustness ratio
Log hourly earnings	W/o migration background	0.03	0.00	1.00	1.46
	Ethnic-German (Aussiedler)	-0.03	0.37	0.88	-0.85
	Interethnic (“Half-German”)	0.01	0.51	0.65	0.41
	Turkish	0.09	0.06	1.00	1.46
	Other non-German	-0.03	0.33	0.80	-0.71
Hourly earnings	W/o migration background	0.28	0.03	1.00	1.47
	Ethnic-German (Aussiedler)	-0.08	0.50	0.57	-0.27
	Interethnic (“Half-German”)	0.06	0.52	0.55	0.08
	Turkish	0.98	0.06	1.00	1.34
	Other non-German	-0.20	0.59	0.76	-0.26

4. Results

Women: Median effect sizes and p-values (768 specifications)

		Median b	Median $p-value$	Sign stability	Robustness ratio
Log hourly earnings	W/o migration background	0.03	0.00	1.00	1.46
	Ethnic-German (Aussiedler)	-0.03	0.37	0.88	-0.85
	Interethnic (“Half-German”)	0.01	0.51	0.65	0.41
	Turkish	0.09	0.06	1.00	1.46
	Other non-German	-0.03	0.33	0.80	-0.71
Hourly earnings	W/o migration background	0.28	0.03	1.00	1.47
	Ethnic-German (Aussiedler)	-0.08	0.50	0.57	-0.27
	Interethnic (“Half-German”)	0.06	0.52	0.55	0.08
	Turkish	0.98	0.06	1.00	1.34
	Other non-German	-0.20	0.59	0.76	-0.26

- Evidence for positive association b/w physical attractiveness and earnings for women w/o migration background.
- Weak evidence for positive association for women of Turkish origin.

4. Results

Men: Median effect sizes and p-values (768 specifications)

		Median b	Median $p-value$	Sign stability	Robustness ratio
Log hourly earnings	W/o migration background	0.04	0.00	1.00	1.99
	Ethnic-German (Aussiedler)	0.06	0.06	1.00	1.40
	Interethnic (“Half-German”)	-0.04	0.28	0.97	-1.01
	Turkish	-0.00	0.64	0.57	-0.30
	Other non-German	0.03	0.38	0.86	0.93
Hourly earnings	W/o migration background	0.48	0.00	1.00	1.51
	Ethnic-German (Aussiedler)	0.72	0.12	1.00	0.96
	Interethnic (“Half-German”)	-0.65	0.15	1.00	-1.08
	Turkish	-0.35	0.57	0.77	-0.81
	Other non-German	0.16	0.43	0.72	0.58

4. Results

Men: Median effect sizes and p-values (768 specifications)

		Median b	Median $p-value$	Sign stability	Robustness ratio
Log hourly earnings	W/o migration background	0.04	0.00	1.00	1.99
	Ethnic-German (Aussiedler)	0.06	0.06	1.00	1.40
	Interethnic (“Half-German”)	-0.04	0.28	0.97	-1.01
	Turkish	-0.00	0.64	0.57	-0.30
	Other non-German	0.03	0.38	0.86	0.93
Hourly earnings	W/o migration background	0.48	0.00	1.00	1.51
	Ethnic-German (Aussiedler)	0.72	0.12	1.00	0.96
	Interethnic (“Half-German”)	-0.65	0.15	1.00	-1.08
	Turkish	-0.35	0.57	0.77	-0.81
	Other non-German	0.16	0.43	0.72	0.58

- Robust evidence for positive association b/w physical attractiveness and earnings in men w/o migration background.
- Weak evidence for positive association in Ethnic-German men.

4. Results

- Presenting the results
  - Summary statistics (Simonsohn et al. 2020; Young & Holsteen 2017)
  - Influence analysis (Young & Holsteen 2017)
    1. How many model assumptions can be relaxed without overturning the conclusion from that estimate?
    2. Which model assumptions are most critical to the results

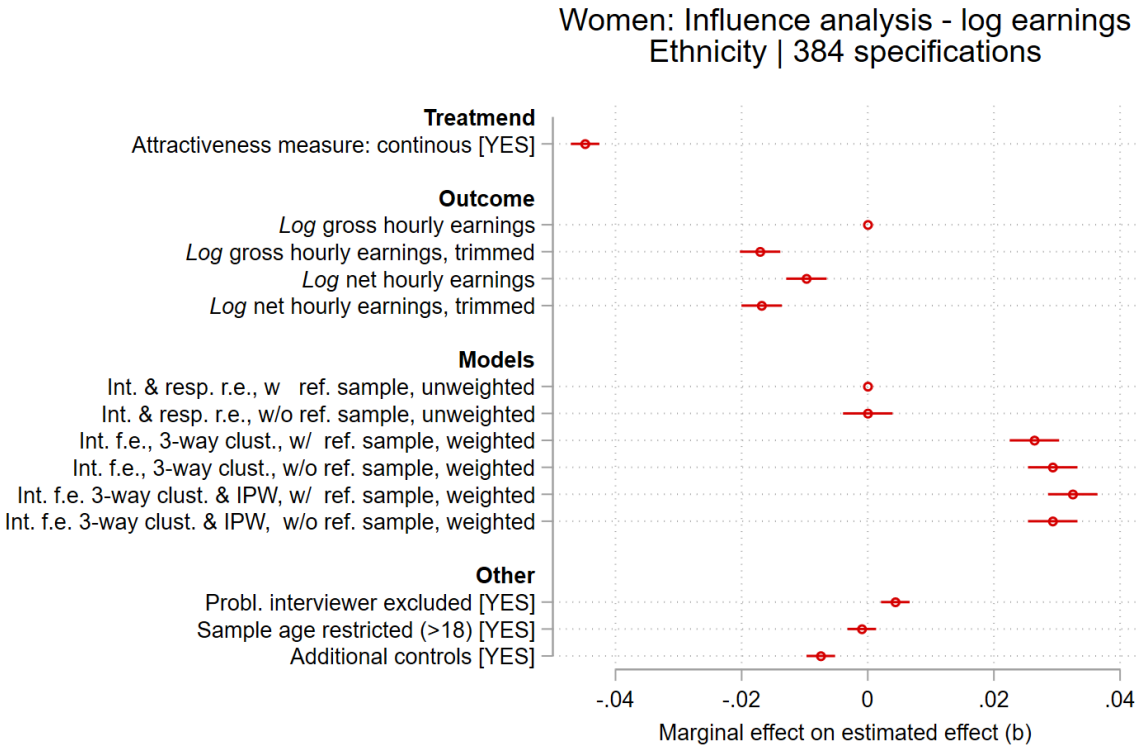
4. Results

- Presenting the results
  - Summary statistics (Simonsohn et al. 2020; Young & Holsteen 2017)
  - Influence analysis (Young & Holsteen 2017)
    1. How many model assumptions can be relaxed without overturning the conclusion from that estimate?
    2. Which model assumptions are most critical to the results

Importantly:

Does not tell us which assumptions are correct!
Only points to those which are critical to the findings.

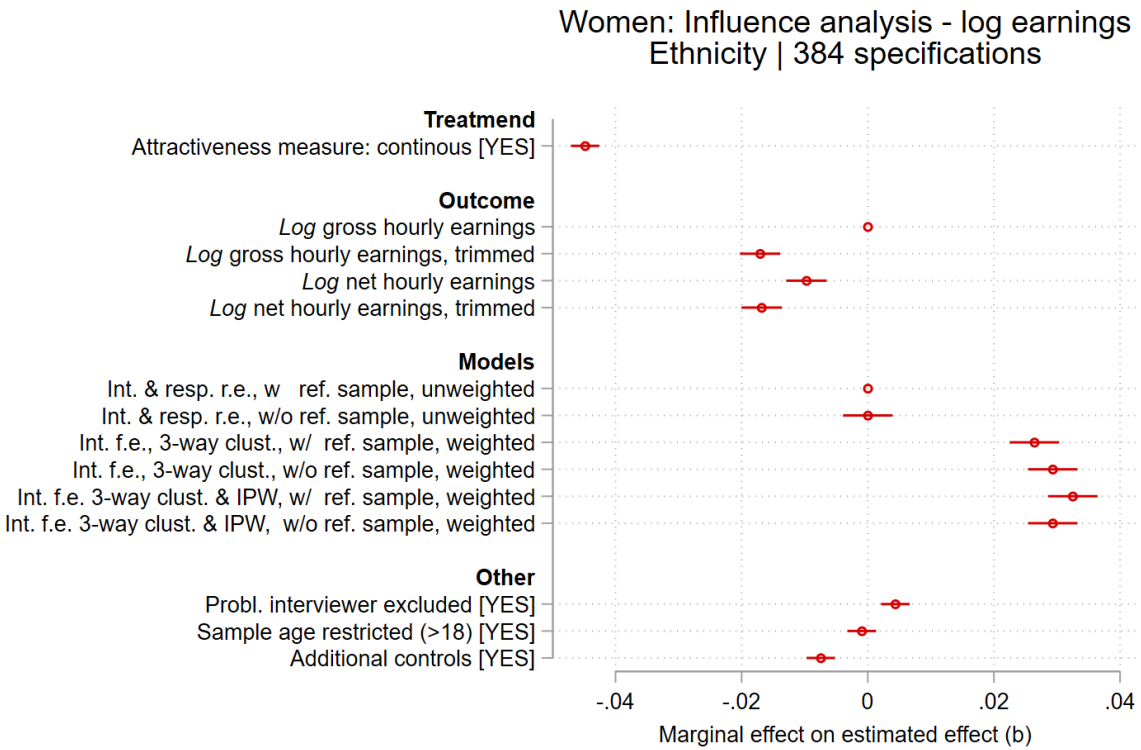
4. Results



Outcome = b
Covariates: specifications

Disclaimer: Ignore CIs!

4. Results

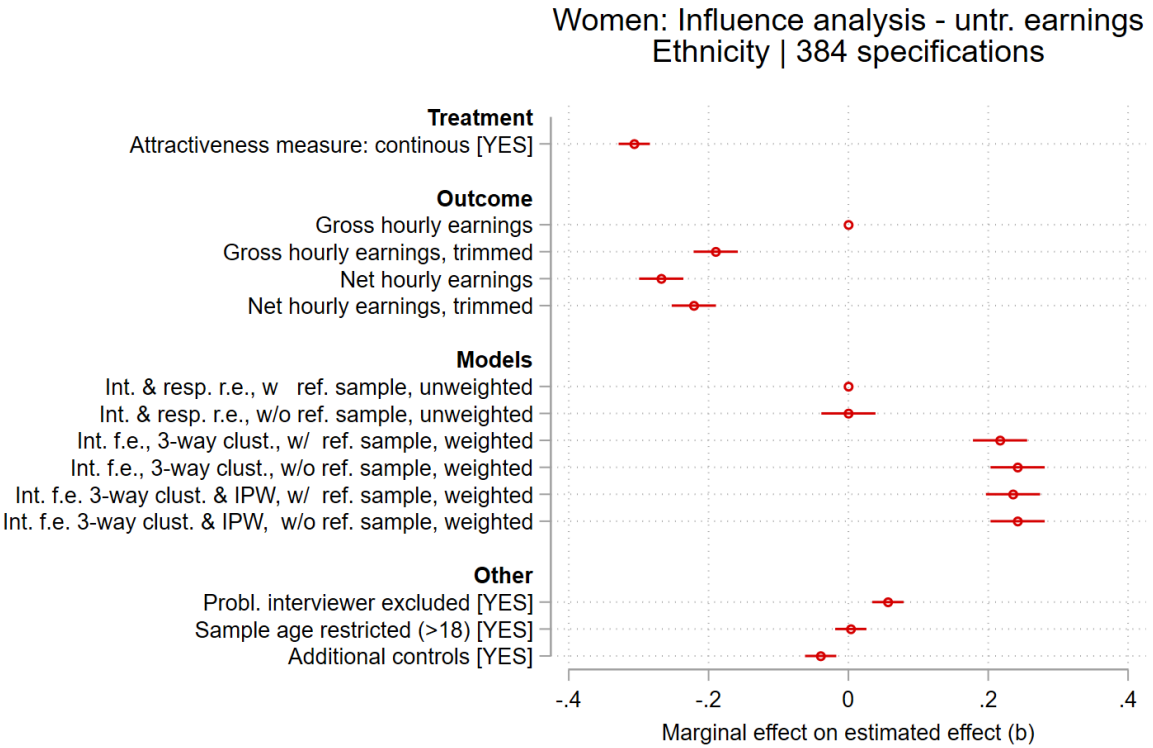


Outcome = b
Covariates: specifications

Trimming makes a difference

Model choice has a large impact:
Weighting?

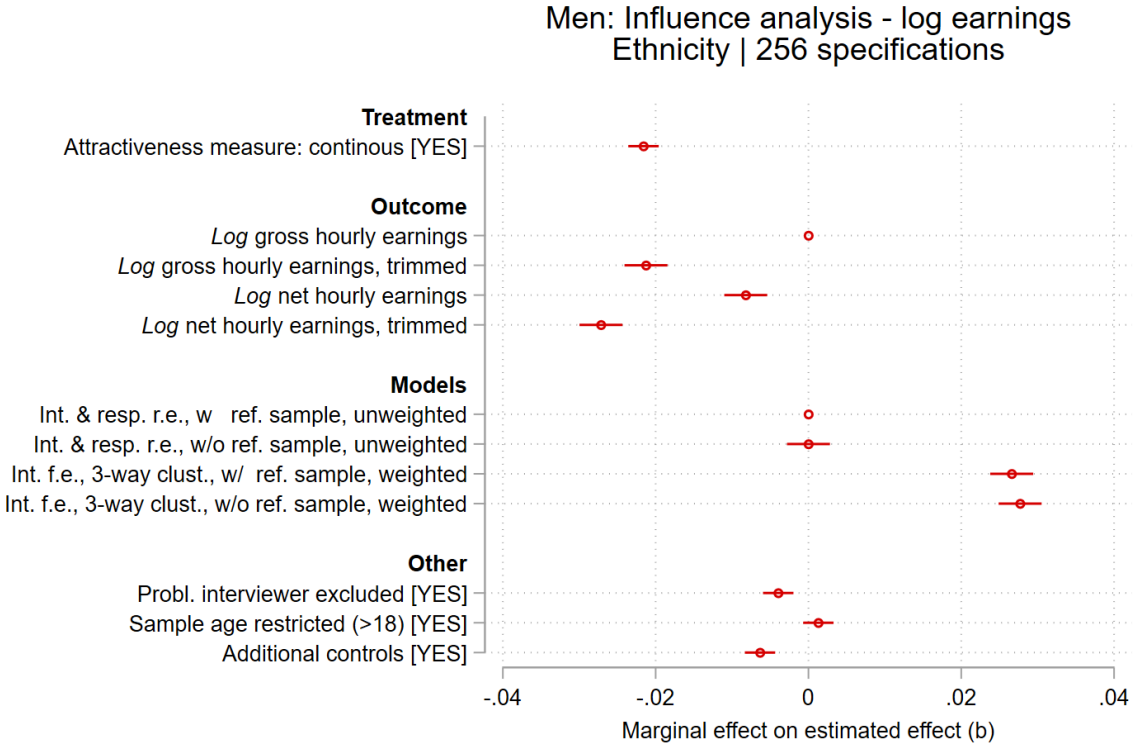
4. Results



Outcome = b
Covariates: specifications

Similar picture for untr. earnings

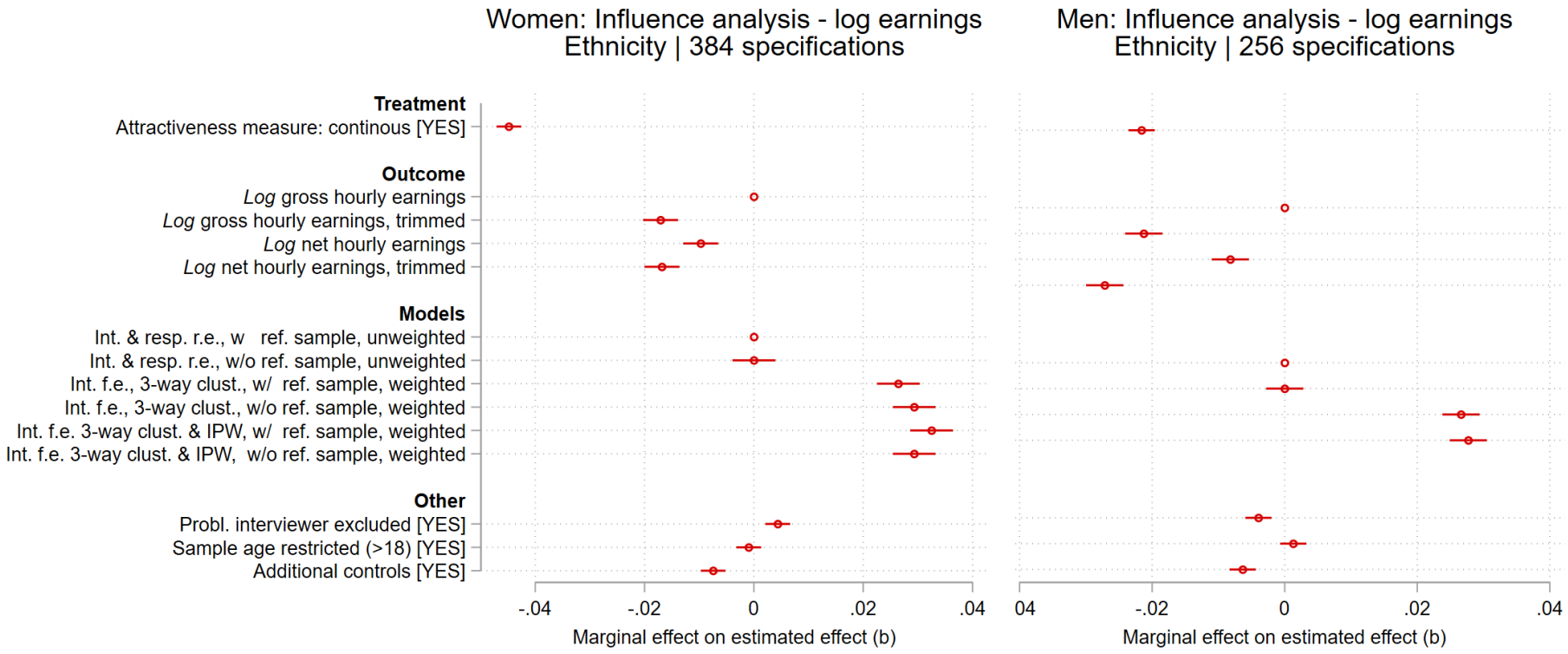
4. Results



Outcome = b
Covariates: specifications

Fewer specifications (no IPW)

4. Results

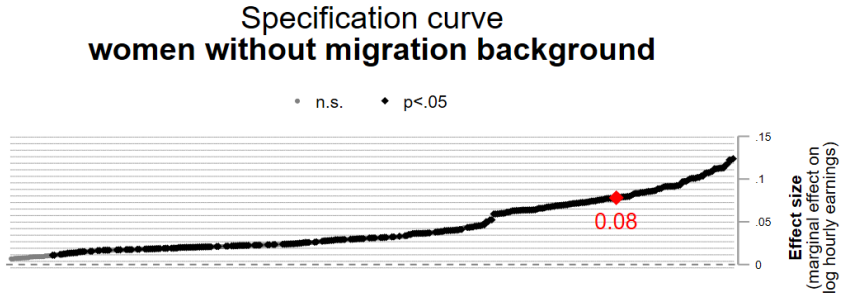

Women: Influence analysis - log earnings
Ethnicity | 384 specifications
Men: Influence analysis - log earnings
Ethnicity | 256 specifications
Treatment
Attractiveness measure: continous [YES]
Outcome
Log gross hourly earnings
Log gross hourly earnings, trimmed
Log net hourly earnings
Log net hourly earnings, trimmed
Models
Int. & resp. r.e., w ref. sample, unweighted
Int. & resp. r.e., w/o ref. sample, unweighted
Int. f.e., 3-way clust., w/ ref. sample, weighted
Int. f.e., 3-way clust., w/o ref. sample, weighted
Int. f.e. 3-way clust. & IPW, w/ ref. sample, weighted
Int. f.e. 3-way clust. & IPW, w/o ref. sample, weighted
Other
Probl. interviewer excluded [YES]
Sample age restricted (>18) [YES]
Additional controls [YES]
-.04
-.02
0
.02
.04
Marginal effect on estimated effect (b)
04
-.02
0
.02
.04
Marginal effect on estimated effect (b)

4. Results

- Presenting the results
  - Summary statistics (Simonsohn et al. 2020; Young & Holsteen 2017)
  - Influence analysis (Young & Holsteen 2017)
  - Specification curves (Simonsohn et al. 2020)
    - Graphical illustration
      - Top panel depicts estimated effect size
      - Bottom panel depicts specifications
      - Displays how specifications interact

      - … not all figures (only German & Turkish)
      - Analysis uses approach / do-files by Simonsohn et al. 2020

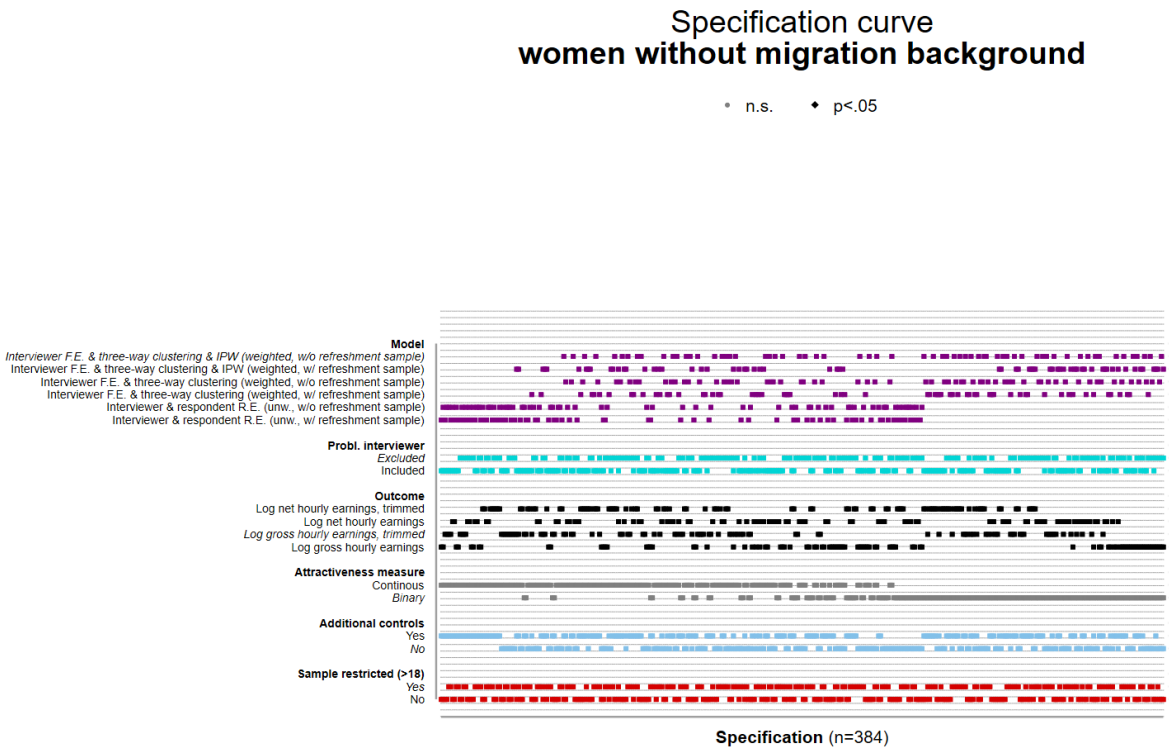
4. Results



Distribution of estimates sorted by size

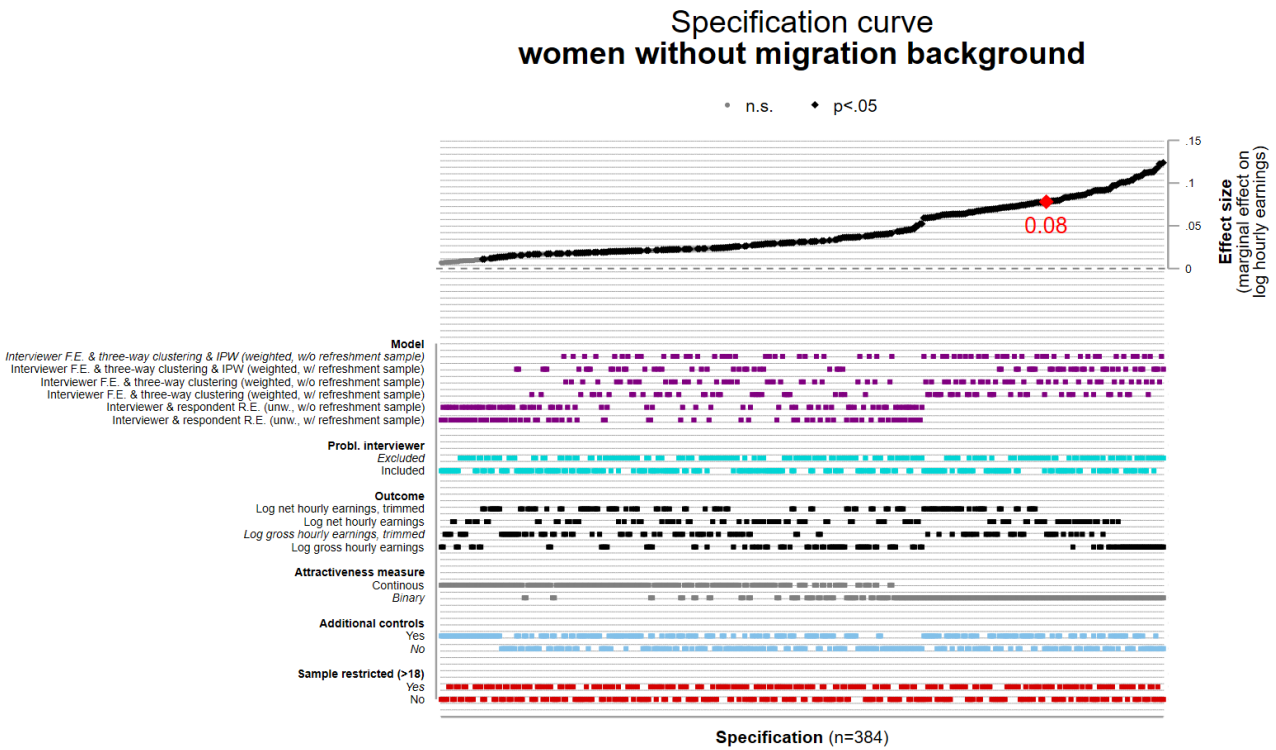
Preferred estimate highlighted in red

4. Results

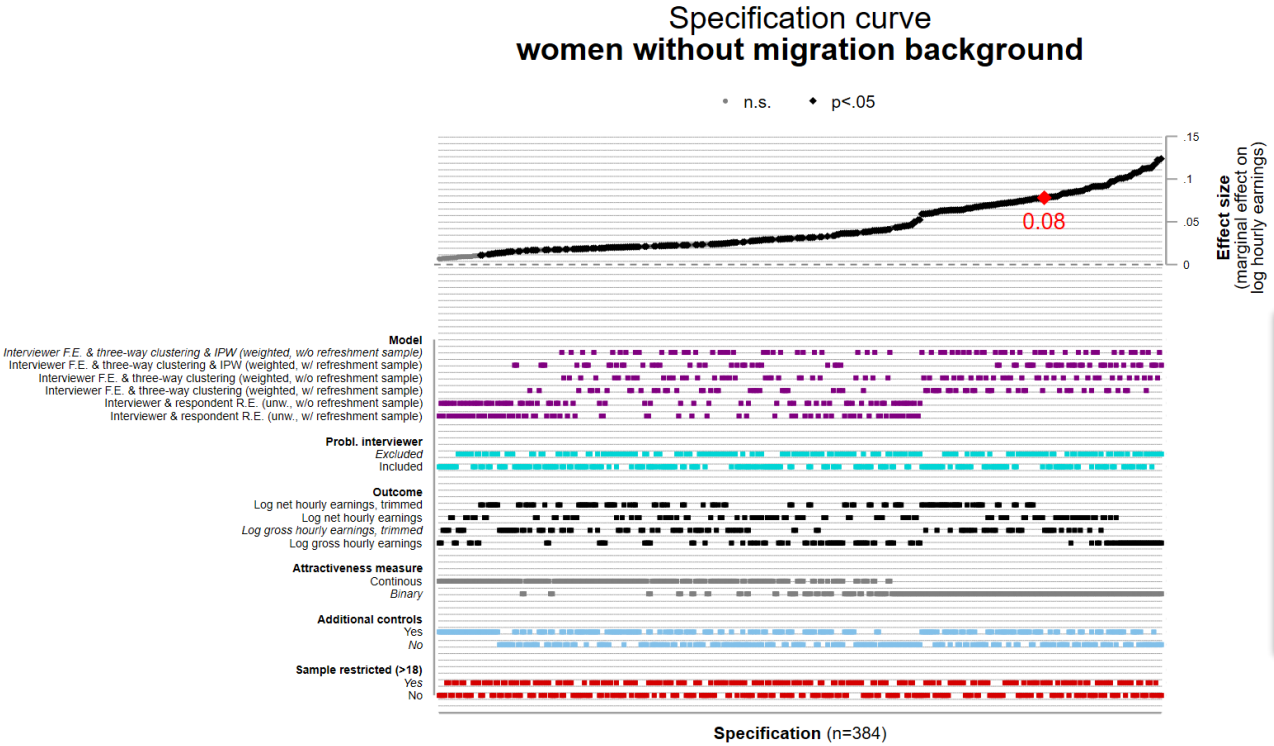


Display of specifications

4. Results


Specification curve
women without migration background
n.s.
p<.05
.15
.1
.05
0
0.08
Effect size
(marginal effect on
log hourly earnings)
Model
Interviewer F.E. & three-way clustering & IPW (weighted, w/o refreshment sample)
Interviewer F.E. & three-way clustering & IPW (weighted, w/ refreshment sample)
Interviewer F.E. & three-way clustering (weighted, w/o refreshment sample)
Interviewer F.E. & three-way clustering (weighted, w/ refreshment sample)
Interviewer & respondent R.E. (unw., w/o refreshment sample)
Interviewer & respondent R.E. (unw., w/ refreshment sample)
Probl. interviewer
Excluded
Included
Outcome
Log net hourly earnings, trimmed
Log net hourly earnings
Log gross hourly earnings, trimmed
Log gross hourly earnings
Attractiveness measure
Continous
Binary
Additional controls
Yes
No
Sample restricted (>18)
Yes
No
Specification (n=384)

4. Results

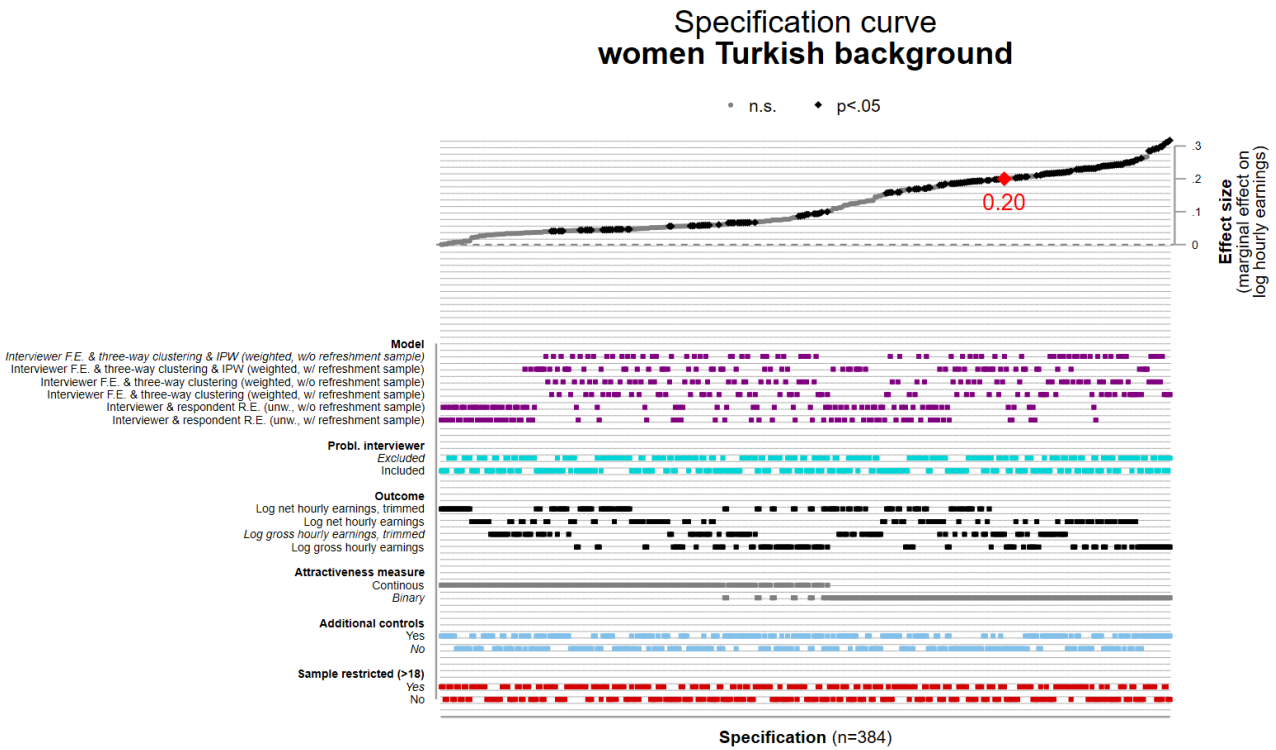


For example:

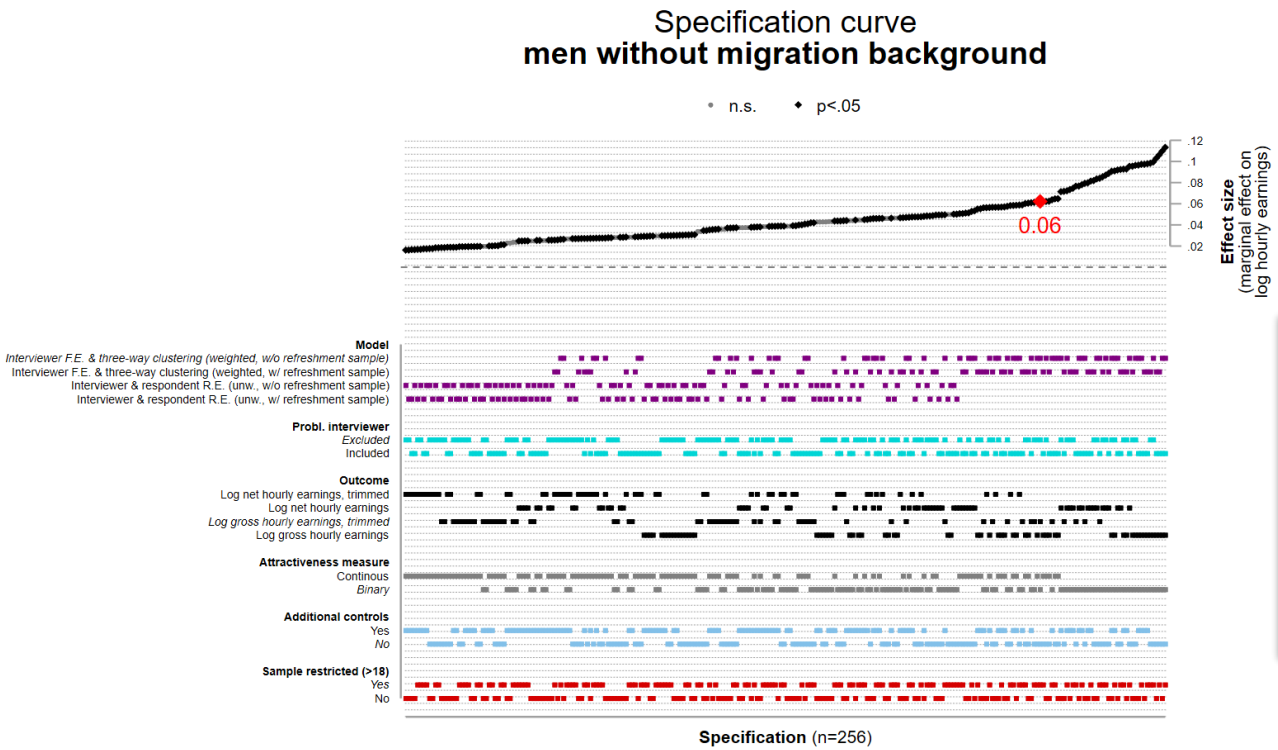
Obtaining a statistically insignificant point estimate requires:

(1) random effect model (unweighted)
(2) continuous measure of attractiveness
(3) inclusion of additional controls
(4) not using trimmed log net earnings

4. Results


Specification curve
women Turkish background
n.s.
p<.05
0.20
.3
.2
.1
0
Effect size
(marginal effect on
log hourly earnings)
Model
Interviewer F.E. & three-way clustering & IPW (weighted, w/o refreshment sample)
Interviewer F.E. & three-way clustering & IPW (weighted, w/ refreshment sample)
Interviewer F.E. & three-way clustering (weighted, w/o refreshment sample)
Interviewer F.E. & three-way clustering (weighted, w/ refreshment sample)
Interviewer & respondent R.E. (unw., w/o refreshment sample)
Interviewer & respondent R.E. (unw., w/ refreshment sample)
Probl. interviewer
Excluded
Included
Outcome
Log net hourly earnings, trimmed
Log net hourly earnings
Log gross hourly earnings, trimmed
Log gross hourly earnings
Attractiveness measure
Continous
Binary
Additional controls
Yes
No
Sample restricted (>18)
Yes
No
Specification (n=384)

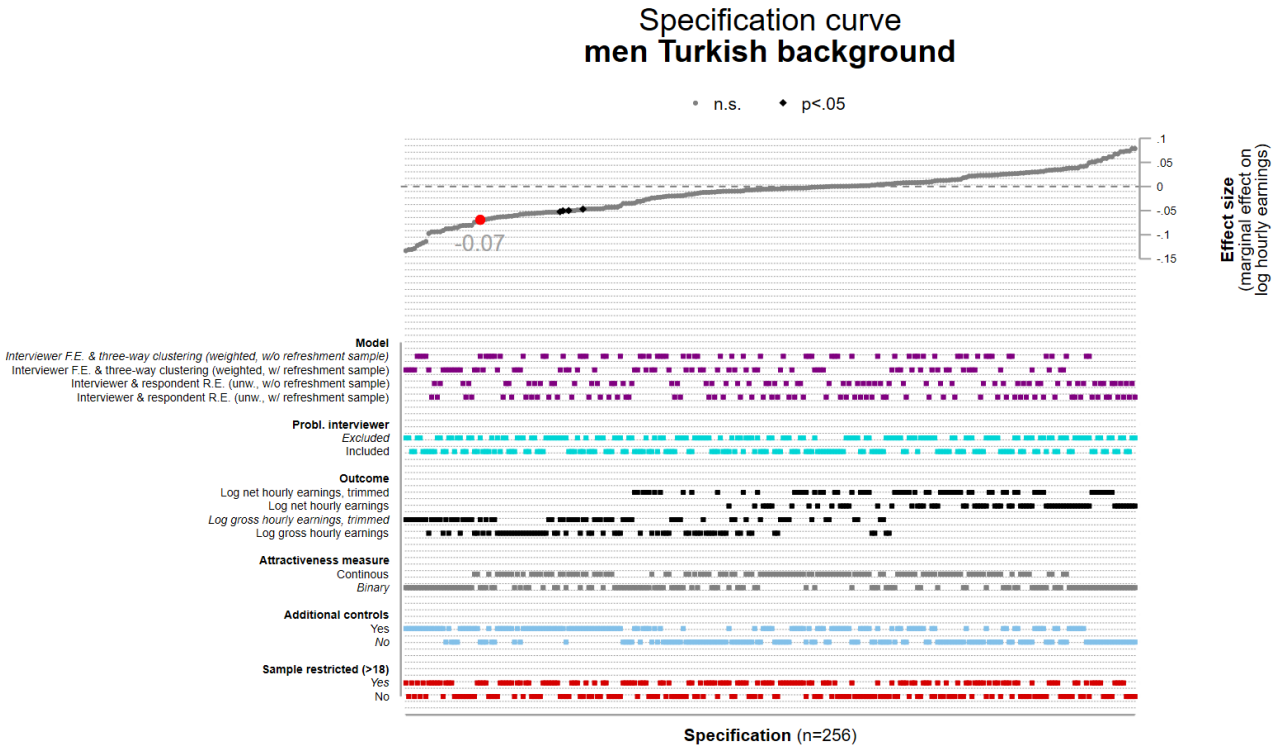
4. Results



For example:

No obvious pattern.

4. Results


Specification curve
men Turkish background
n.s.
p<.05
.1
.05
0
-.05
-.1
-.15
-0.07
Effect size
(marginal effect on
log hourly earnings)
Model
Interviewer F.E. & three-way clustering (weighted, w/o refreshment sample)
Interviewer F.E. & three-way clustering (weighted, w/ refreshment sample)
Interviewer & respondent R.E. (unw., w/o refreshment sample)
Interviewer & respondent R.E. (unw., w/ refreshment sample)
Probl. interviewer
Excluded
Included
Outcome
Log net hourly earnings, trimmed
Log net hourly earnings
Log gross hourly earnings, trimmed
Log gross hourly earnings
Attractiveness measure
Continous
Binary
Additional controls
Yes
No
Sample restricted (>18)
Yes
No
Specification (n=256)

4. Results

- Presenting the results
  - Summary statistics (Simonsohn et al. 2020; Young & Holsteen 2017)
  - Influence analysis (Young & Holsteen 2017)
  - Specification curves (Simonsohn et al. 2020)
  - Inference … (Simonsohn et al. 2020)

4. Results

- Inference (Simonsohn et al. 2020)
  - Considering the full set of specifications, how inconsistent are the results with the null hypothesis of no effect?
  - Test-statistics, e.g.:
    - Median effect estimate: test if estimated median effect estimate is more extreme than would be expected under-the-null.
    - Share of statistically significant estimates in expected direction: test if estimated share estimate is more extreme than would be expected under-the-null.

4. Results

- Inference (Simonsohn et al. 2020)
  - Obvious problem: impossible to generate test-statistic under-the-null analytically.
  - Specifications are not independent and *not part of single model.*

4. Results

- Inference (Simonsohn et al. 2020)
  - Obvious problem: impossible to generate test-statistic under-the-null analytically.
  - Specifications are not independent and *not part of single model.*

  - Solution: generate distribution empirically through resampling under-the-null.
    - Modify observed data: null hypothesis is true
      (→ no association b/w treatment and outcome).
    - Resample (bootstrap) & estimate statistic of interest
      (e.g., median effect estimate).
    - P value: % of resamples with statistic as or more extreme than observed.

4. Results

- Inference (Simonsohn et al. 2020)
  - Problem: computationally very intensive.
    - Multiverse analysis above: “fast” ~ 30 hours.
    - 500-1000 resamples…

4. Results

- Inference (Simonsohn et al. 2020)
  - Problem: computationally very intensive.
    - Multiverse analysis above: “fast” ~ 30 hours.
    - 500-1000 resamples…

- Missing values
  - Similar problem with multiple imputation.

5. Discussion

- Multiverse analysis promising approach: transparency & robustness.
- Can make *theoretical considerations* more important b/c made explicit.

5. Discussion

- Open questions:

1. Birds-eye view: loosing sight of the data / a specific model.
2. Another method to hide behind: GIGO – 100% of models return statistically significant and substantially relevant estimates *but the estimates are biased* (our case?).
3. Affinity towards *statistical significance*?
4. Relation b/w preferred model and the multiverse?
5. …

References

- Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., ... & Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119.
- Brüderl, J., Garrett, M., Hajek, K., Herzig, M., Lenke, R., Lorenz, R., Lutz, K., Phan, T., Schütze, P., & Schumann, N. (2021). Pairfam data manual, Release 12.0. LMU Munich: Technical Report. GESIS Data Archive, Cologne. ZA5678 Data File Version 12.0.0. https://doi.org/doi.org/10.4232/pairfam.5678.12.0.0
- Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.
- Gelman, Andrew, and Eric Loken. (2014). The Statistical Crisis in Science Data Dependent Analysis—a “Garden of Forking Paths”—Explains Why Many Statistically Significant Comparisons Don’t Hold Up.” *American Scientist* 102(6), 460.
- Hernán, M. A., & Cole, S. R. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, 170(8), 959–962
- Hernán M.A., Robins, J.M. (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199-236.
- Muñoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, *48*(1), 1-33.
- Sala-i-Martin, X. (1997). I just ran four million regressions.
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... & Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337-356.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208-1214.
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.
- Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*(1), 3-40.
- Young, C. (2018). Model uncertainty and the crisis in science. *Socius*, *4*, 2378023117737206.