Meta-Analysis for Policy Economists

Practical Tools for Synthesizing Research

Tomas Havranek

Charles University, Prague
Centre for Economic Policy Research, London
Meta-Research Innovation Center, Stanford

Workshop prepared for Tillväxtanalys, Stockholm

August 28–29, 2025

Workshop Outline

1. Motivation
2. Applied Examples
3. Literature Search
4. Data Collection
5. Conventional Tools
6. Publication Bias
7. P-hacking
8. Heterogeneity
9. Extensions & Limitations
10. Takeaways

Workshop Outline

Meta important for research, policy, practice

Research experience:

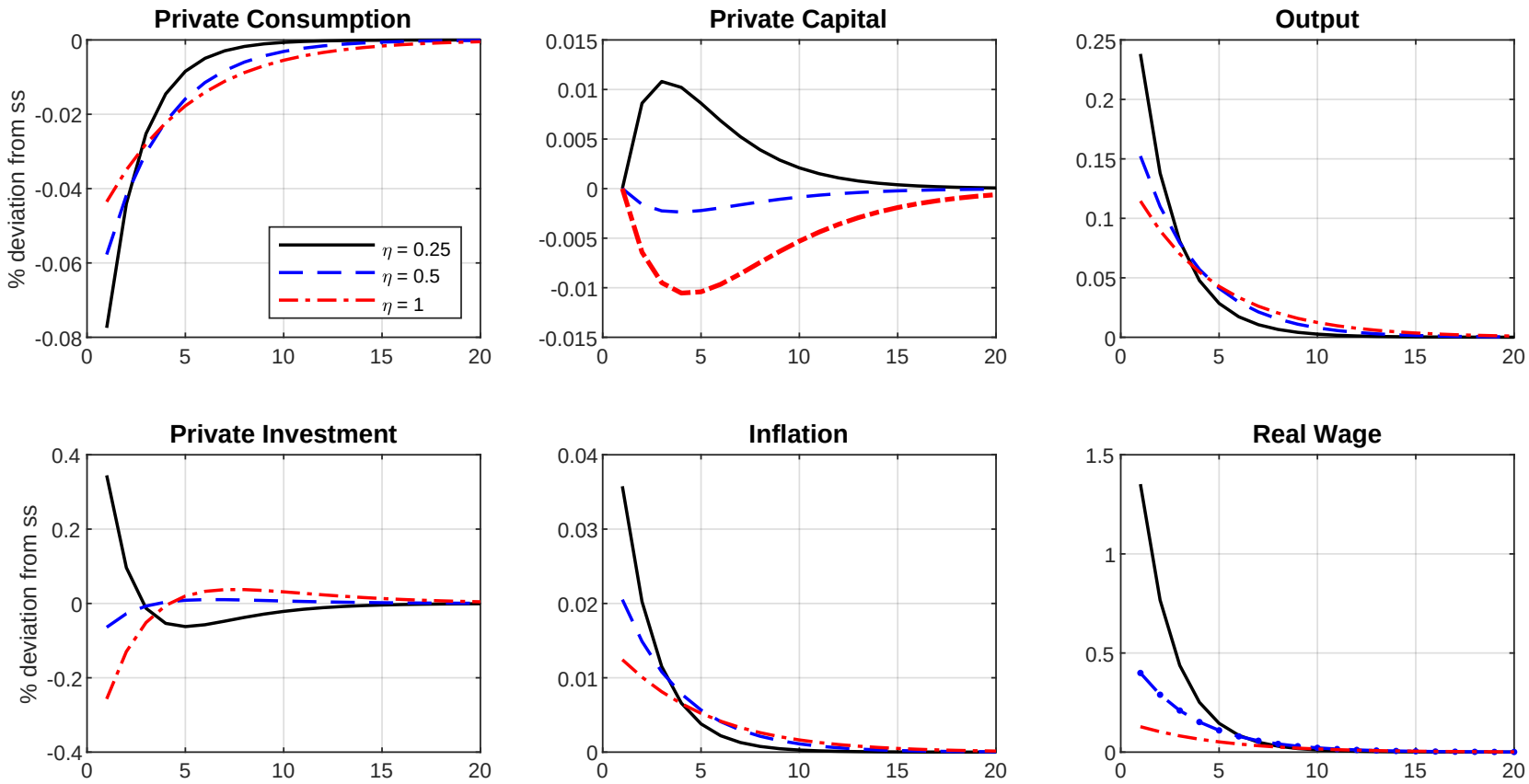
- Affiliate at the Meta-Research Innovation Center, Stanford
- Panel member at ERC Advanced
- Convener of Meta-Analysis in Economics Network
- Meta-Analysis Editor at the Journal of Economic Surveys
- MAIVE method (Nature Communications)
- Over 50 metas published, e.g. REStat, JEEA, JOLE, JIE...

Policy experience:

- Former Advisor to the Board, Czech National Bank
- Former member of the Czech National Economic Council
- Affiliate at the Centre for Economic Policy Research, London

What happens if the government spends more?

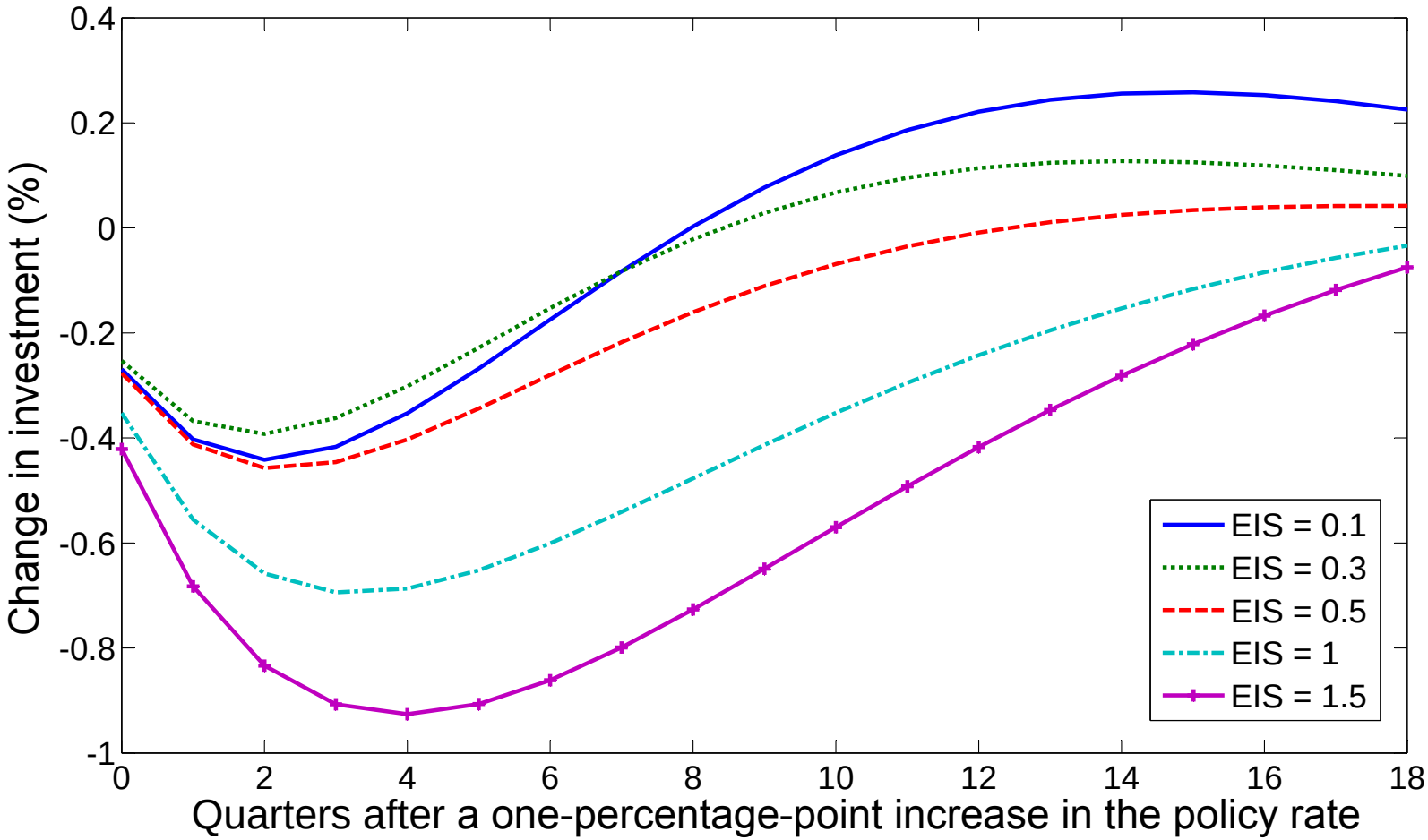
Labor supply elasticity is key (but how to calibrate it?):



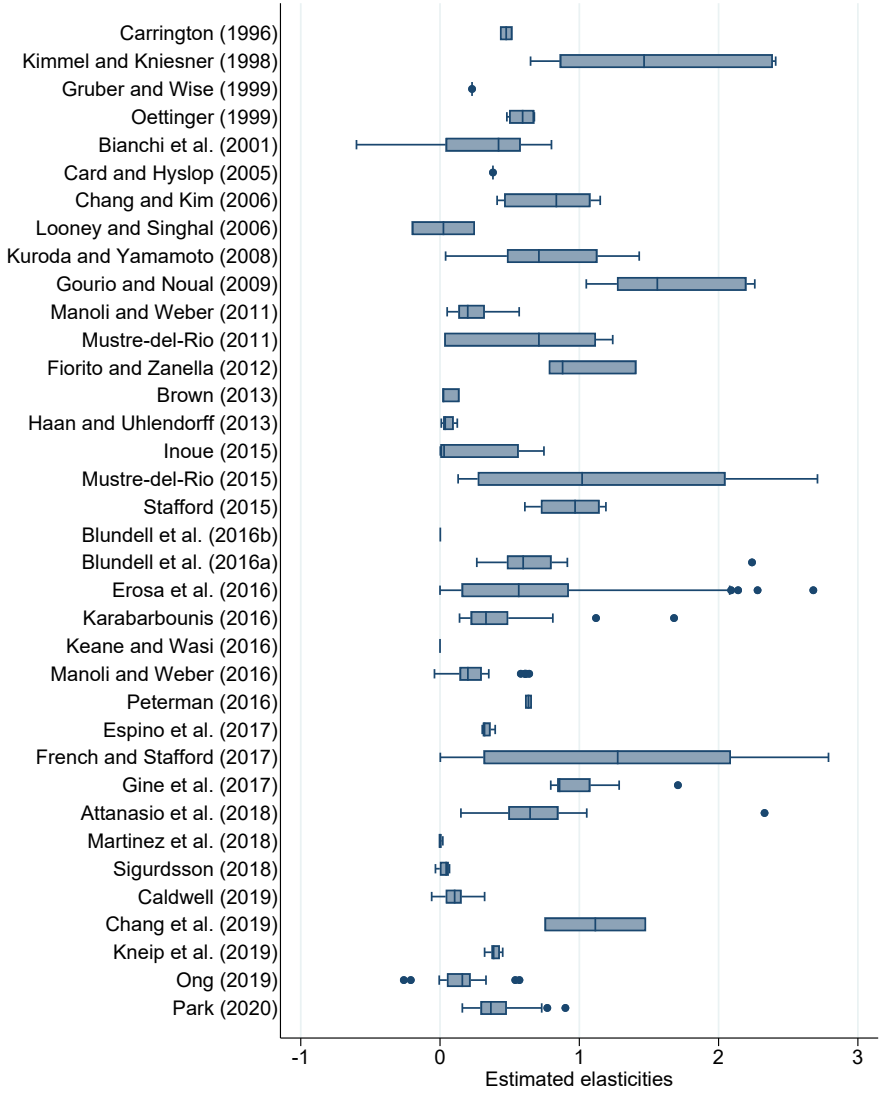
Calibration uncertainty: meta needed!

Parameter		Prior Mean
Preferences		
Habits	h	0.900
Labour supply coef.	ψ	8.832
Frisch elasticity	ϑ	1.250
Wedges		
Euler	κ^{euler}	1.0119
Forex	κ^{forex}	1.0034
Adjustment costs		
Investment	κ	20.000
Capital utilization	$\gamma 2$	0.280
Risk premium	Γ_{b_W}	0.800
Elasticities of substitution		
Domestic goods	ϵ	5.000
Import goods	ϵ_M	9.000
Export goods	ϵ_x	9.400
World goods	ϵ_W	0.860
Consumption goods	ϵ_c	7.600
Investment goods	ϵ_i	7.600
Labour types	η	7.000
Price and wage setting		
Calvo dom. prices	θ_p	0.500
Calvo exp. prices	θ_x	0.080
Calvo imp. prices	θ_M	0.750
Calvo wages	θ_w	0.380
Index. dom. prices	χ_p	0.750
Index. imp. prices	χ_M	0.500
Index. exp. prices	χ_x	0.350
Index. wages	χ_w	0.920
Monetary policy		
Taylor rule (int. rates)	γ_R	0.960
Taylor rule (output gap)	γ_y	0.220
Taylor rule (inflation)	γ_Π	1.150
Fiscal policy		
Public consumption	ρ_g	0.750
Home bias		
Home bias in consump.	n_c	0.280
Home bias in invest.	n_i	0.120
Home bias in export	n_x	0.350
Growth rates		
Invest. spec. tech.	Λ_μ	1.000
General tech.	Λ_A	1.009
Population	Λ_L	1.000
ER appreciation	$\dot{ex}$	0.994
Openness tech.	α_O	1.0035
Export spec. tech.	α_X	1.0058

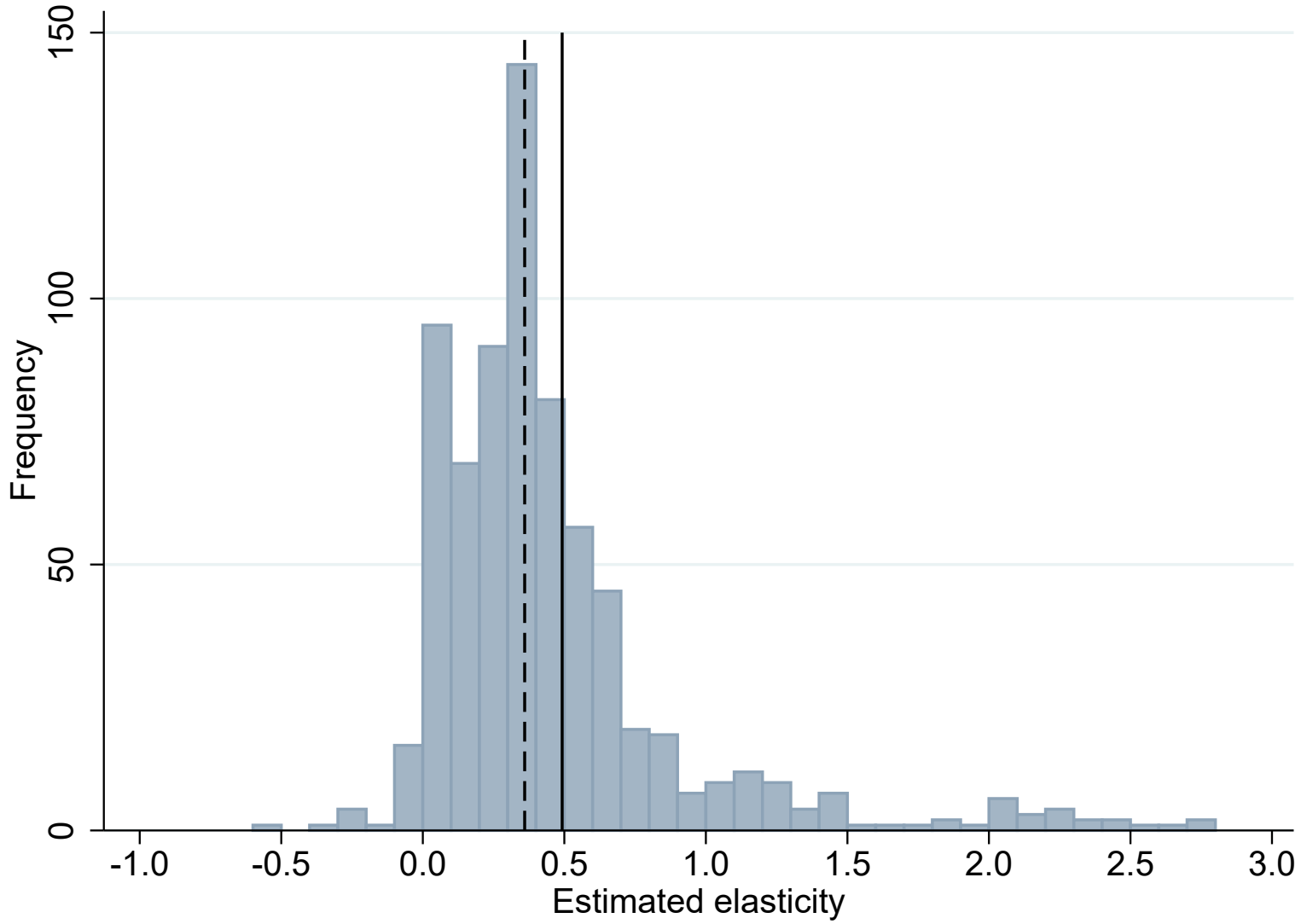
What happens if we change another parameter?



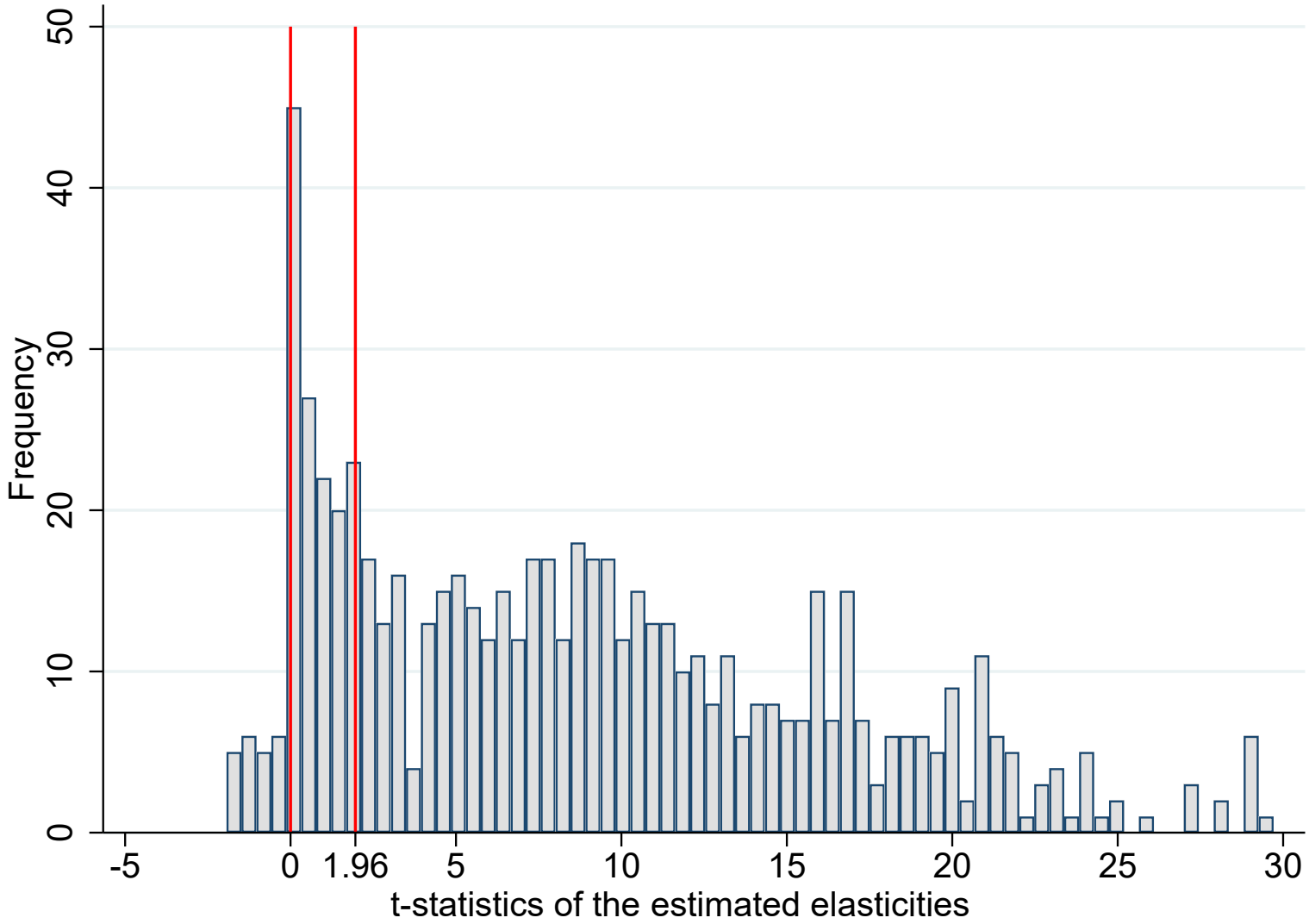
Estimates of the labor supply elasticity vary



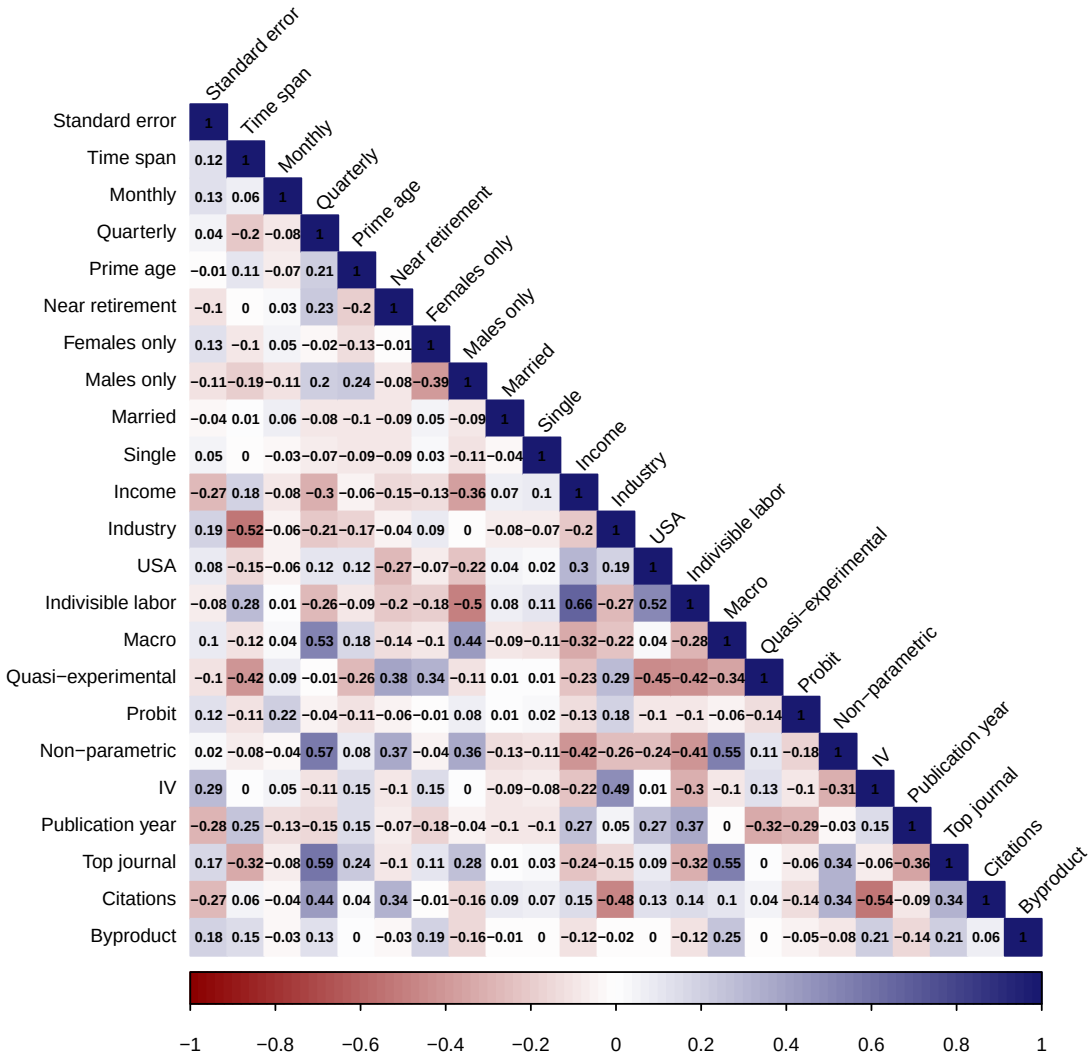
Values around 0.4–0.5 used for calibration (CBO)



Bias: some estimates less likely to be reported



Context: different situations yield different estimates



Implied elasticity: taking bias and context into account

Let's construct a hypothetical study that uses all information reported in the literature but puts more weight on selected aspects (high precision, credible identification, etc.)

	Subjective best practice	Blundell *et al.* (2016a)
Overall	0.02	−0.03
	(−0.22, 0.25)	(−0.25, 0.18)
Near retirement	0.14	0.09
	(−0.14, 0.42)	(−0.14, 0.31)
Women	0.12	0.07
	(−0.08, 0.32)	(−0.10, 0.24)

Guidelines

ARTICLE | Open Access

Meta-analysis of social science research: A practitioner's guide

Zuzana Irsova, Hristos Doucouliagos, Tomas Havranek, T. D. Stanley

First published: 23 November 2023 | https://doi.org/10.1111/joes.12595

SECTIONS PDF TOOLS SHARE

Abstract

This article provides concise, nontechnical, step-by-step guidelines on how to conduct a modern meta-analysis, especially in social sciences. We treat publication bias, *p*-hacking, and systematic heterogeneity as phenomena meta-analysts must always confront. To this end, we provide concrete methodological recommendations. Meta-analysis methods

Guiding example: effect of beauty on success

Hands on

Open Stata and R

Results

Main Findings

1. Mean of 1,159 estimates from 67 studies: good looks $\rightarrow$ earnings $\uparrow$ 4.3%
2. Publication bias. After correction: 2.9%
3. Control for ability? Productivity? Beauty premium $\approx$ 0%

Project Website

www.meta-analysis.cz/beauty

Revisions requested at *Nature Human Behaviour*.

Workshop Outline

Example 1: Daylight saving time

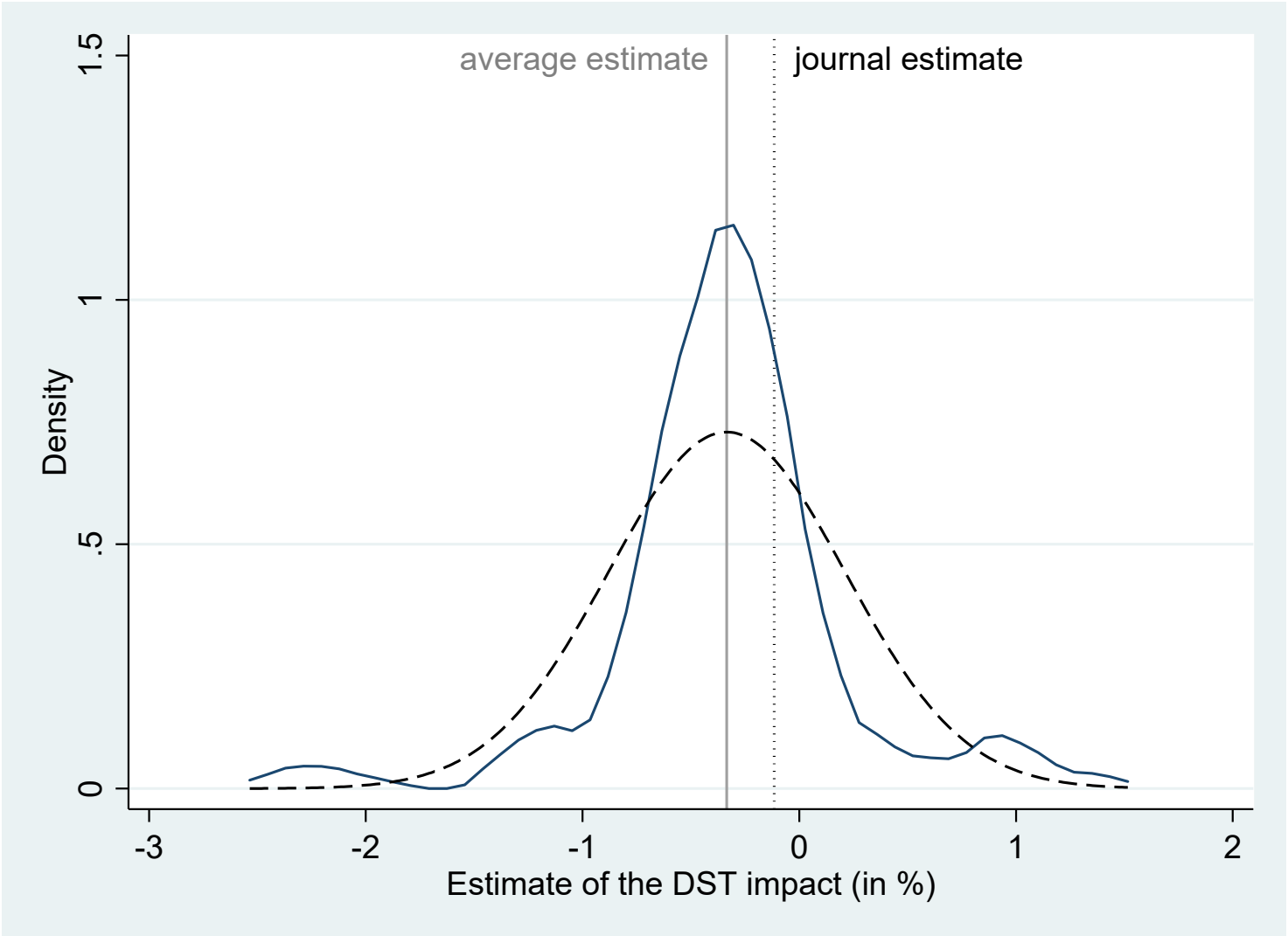
Researchers typically estimate the following model:

$$ln\text{Consumption}_t = \alpha + DST \cdot \text{Treatment effect}_t + \text{Controls}_t + \epsilon.$$

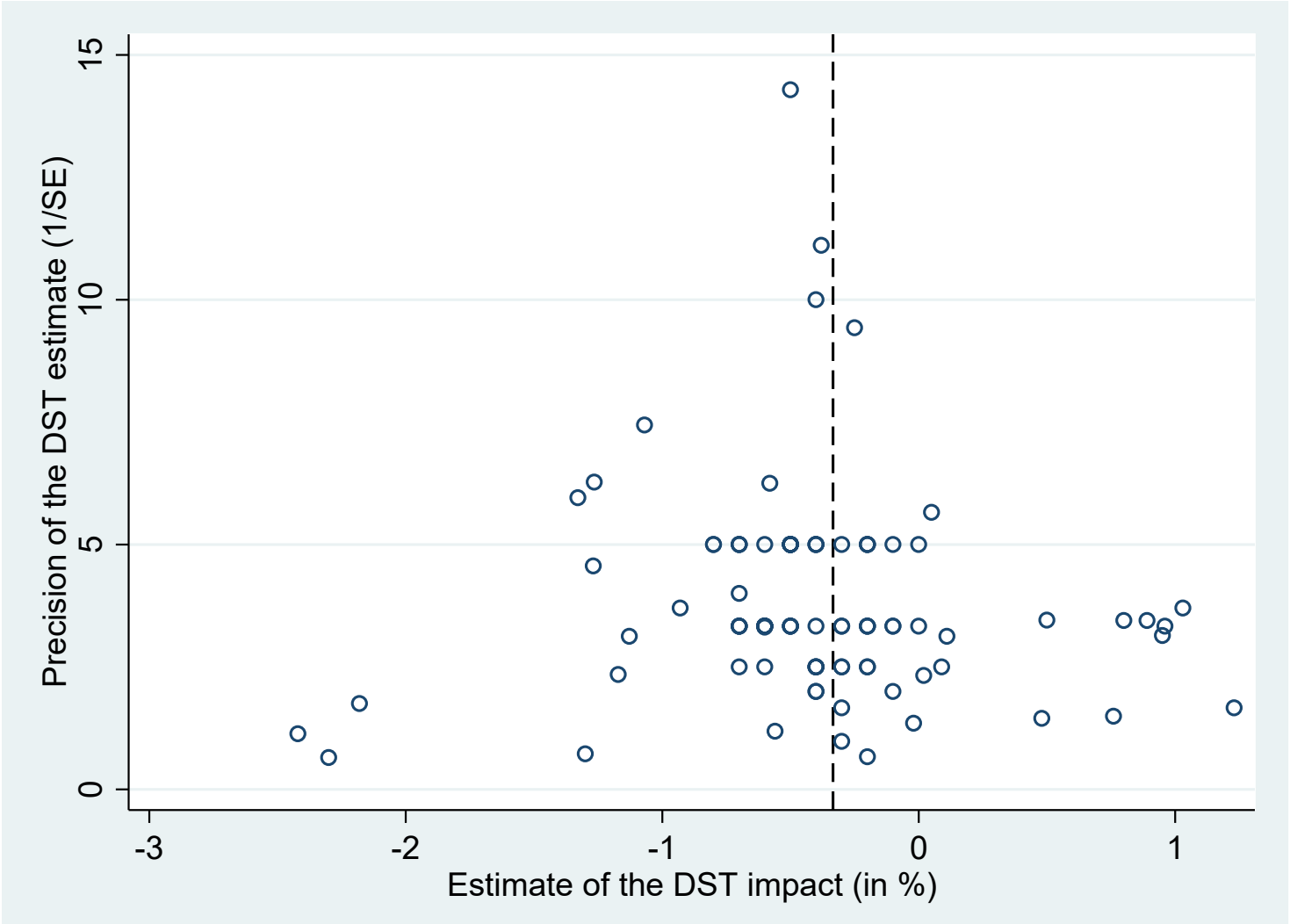
- *Consumption*: the average energy consumption during time *t* for a given hour, day, and year.
- *Treatment effect*: equals 1 for all hours when daylight saving time applies.
- *Controls*: seasonality and holidays, weather, the intensity of sunlight, heterogeneity among consumption units, etc.

No consensus in the literature

Journals report smaller savings



Publication bias? Probably not



Funnel asymmetry tests

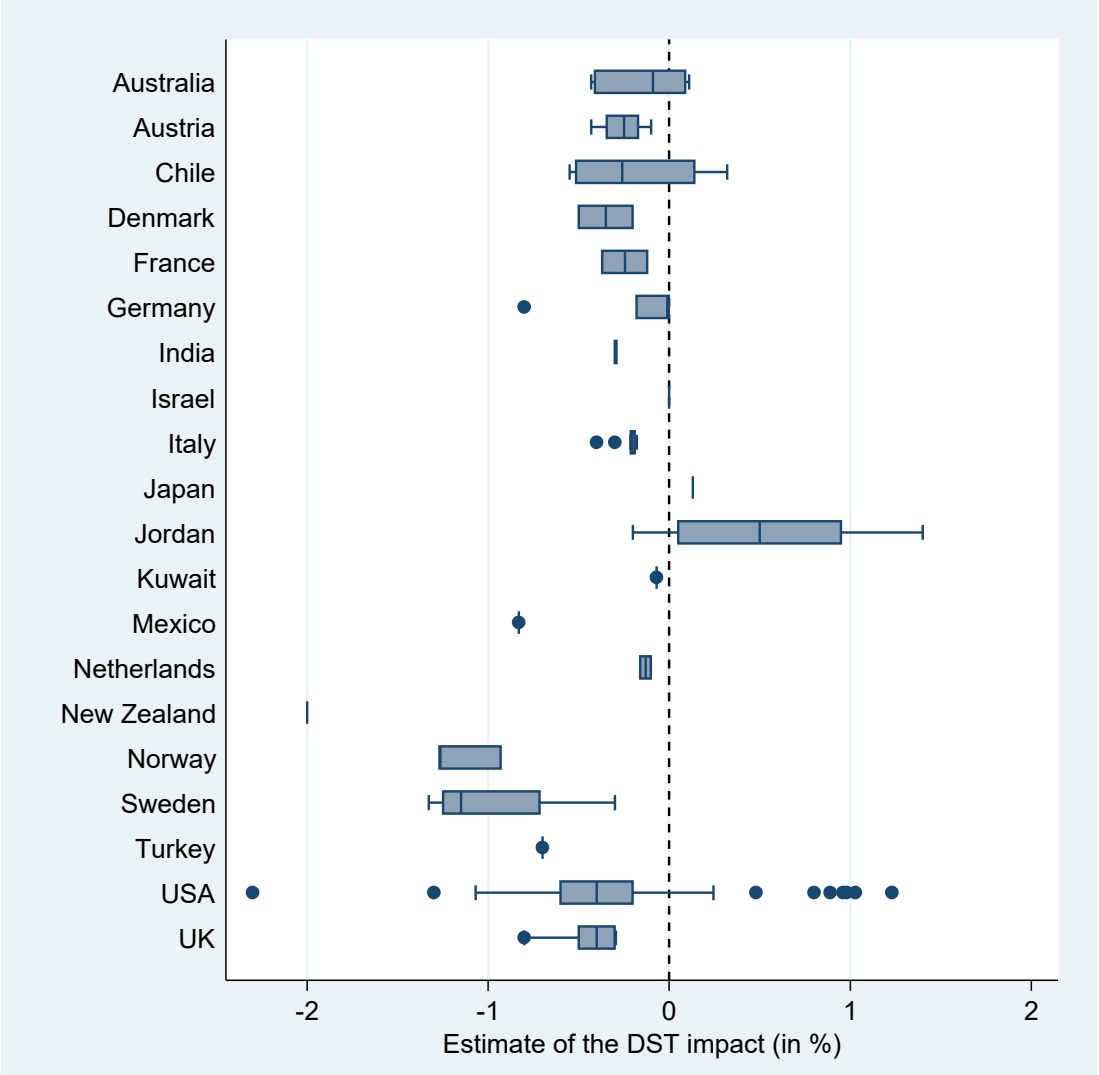
If the ratio of the point estimate to its standard error has a *t*-distribution, then the two quantities should be independent:

$$\text{estimate}_i = \underbrace{\beta}_{\text{true effect}} + \underbrace{\beta_0 SE_i}_{\text{publication bias}} + \mu_i.$$

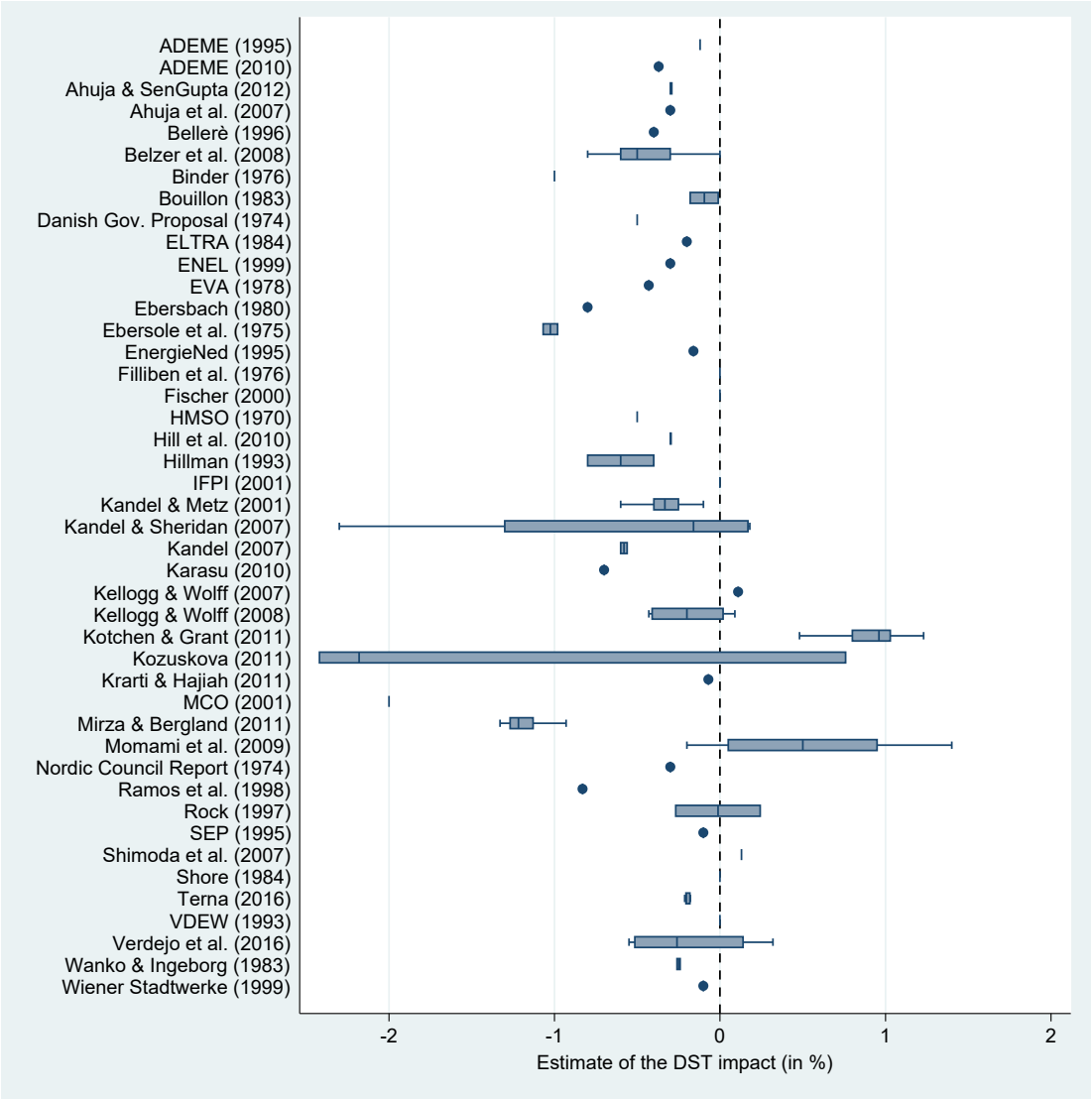
The no. of obs. can be used as an instrument for SE.

	OLS	FE	BE	Country	ME	IV
SE (*publication bias*)	-0.410	-1.217	-0.410	-0.496	-0.449	0.226
	(0.265)	(0.790)	(0.757)	(0.805)	(0.688)	(1.088)
Constant (*true effect*)	-0.293***	-0.222***	-0.294***	-0.278***	-0.291***	-0.445*
	(0.000778)	(0.0700)	(0.00812)	(0.0459)	(0.00731)	(0.243)
Observations	101	101	101	101	101	90

Country heterogeneity



Method heterogeneity (just studies for the US)



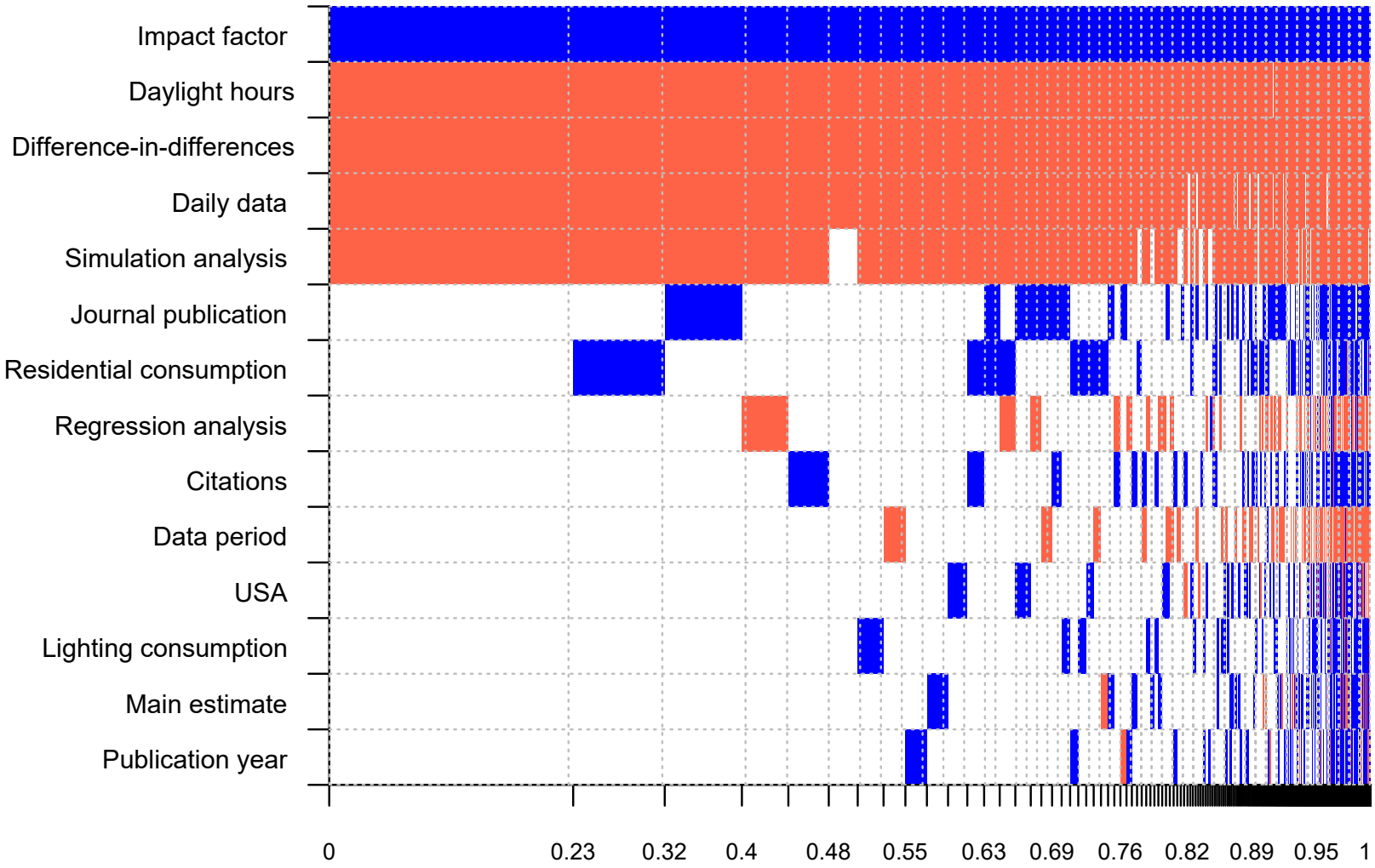
What explains the differences in conclusions? (1)

Data characteristics	
Data period	The number of years used in the estimation.
Main estimate	= 1 if the estimate is preferred by the authors of the study.
Hourly data	= 1 if the data are examined on hourly or higher than hourly granularity.
Daily data	= 1 if the data are examined on a daily basis.
Daylight hours	Average time between sunrise and sunset on the longest day for the country or region under examination.
Europe	= 1 if European countries are examined.
USA	= 1 if US data are examined.
Design of the analysis	
Regression analysis	= 1 if the primary study is based on regression analysis.
Simulation analysis	= 1 if the study is based on simulation.
Difference-in-diff.	= 1 if the difference-in-differences approach is employed.
Residential cons.	= 1 if only residential consumption is examined.
Lighting cons.	= 1 if total energy savings are reported as a result of lighting reduction.

What explains the differences in conclusions? (2)

Publication characteristics	
Publication year	The publication year of the study (base = 1970).
Journal article	= 1 if the study was published in a peer-reviewed journal.
Impact factor	The recursive RePEc impact factor of the outlet.
Citations	The logarithm of the total number of citations of the study in Google Scholar.

Bayesian model averaging



Results

Main Findings

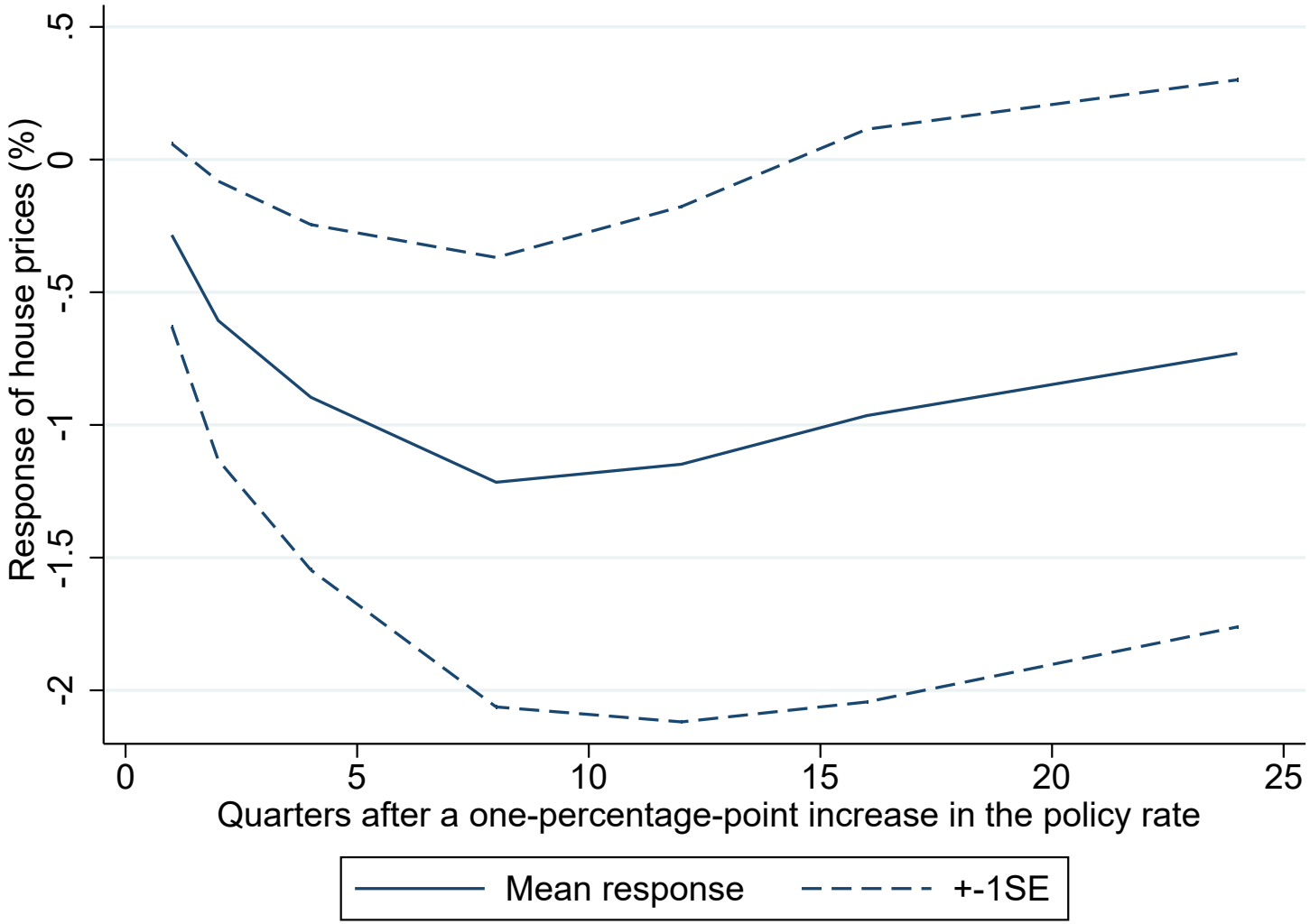
1. The mean reported estimate suggests energy savings of 0.34% due to DST.
2. The results are driven by the method employed (simulation vs. regression).
3. Energy savings are larger for countries farther away from the equator.

Project Website

www.meta-analysis.cz/dst

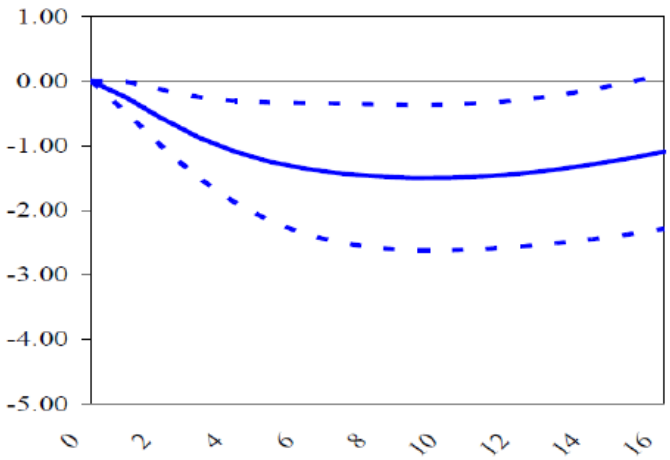
Published in the *Energy Journal*, 2018.

Example 2: Interest rates and house prices



Converting graphs to data

We measure pixel coordinates using WebPlotDigitizer:

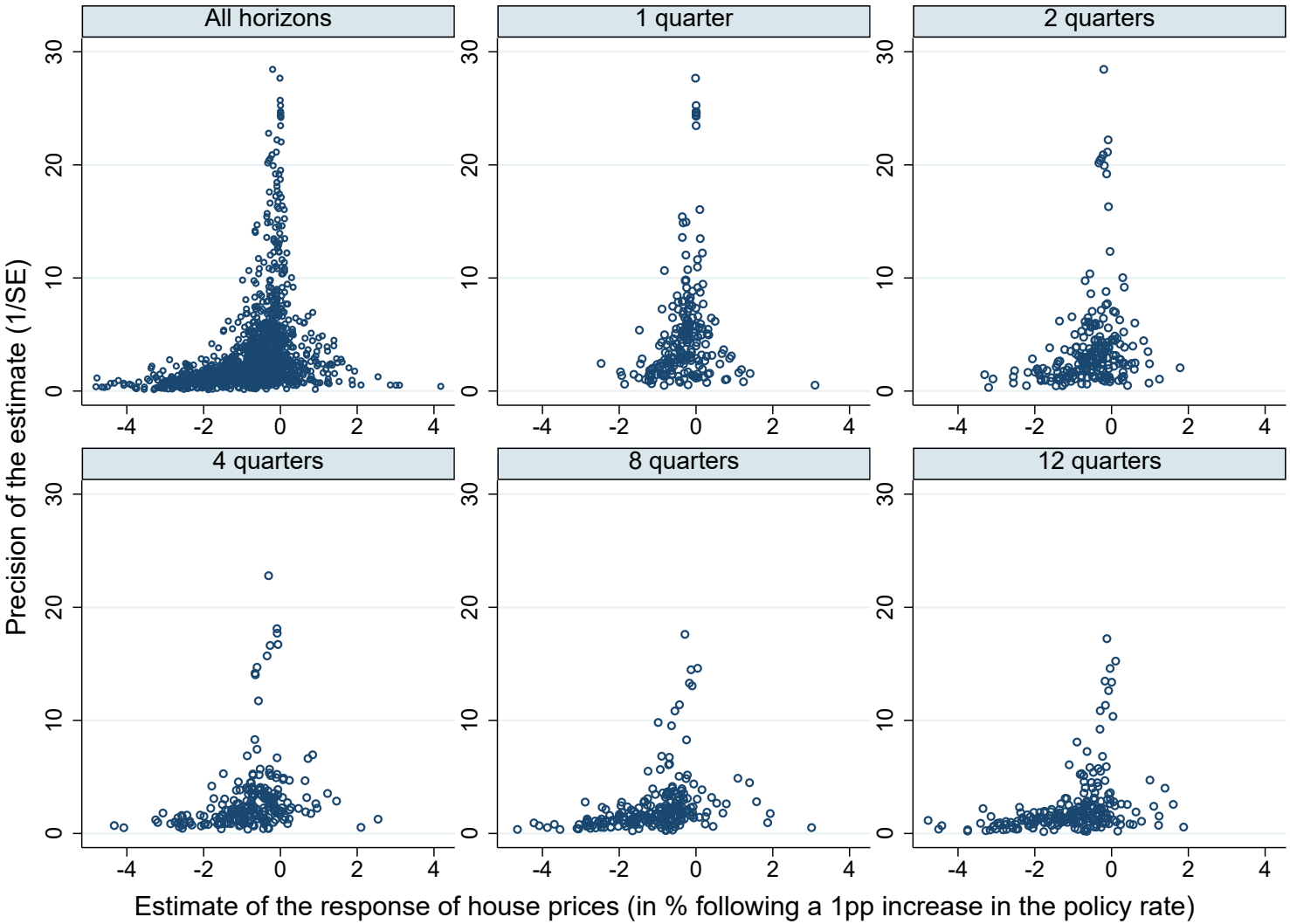


Canada

horizon	response	confidence interval
1	-0.231	(-0.407, -0.016)
2	-0.578	(-1.012, -0.135)
4	-1.076	(-1.853, -0.299)
8	-1.454	(-2.554, -0.359)
12	-1.414	(-2.534, -0.287)
16	-1.084	(-2.259, 0.131)
max	n.a.	n.a.

Obtained from: page 50, Figure 2
Frequency: Quarterly
Note: Responses are in %.

Publication bias for longer horizons?



Corroborated by funnel asymmetry tests

Note that the slope is negative, the intercept is smaller than the reported mean effect:

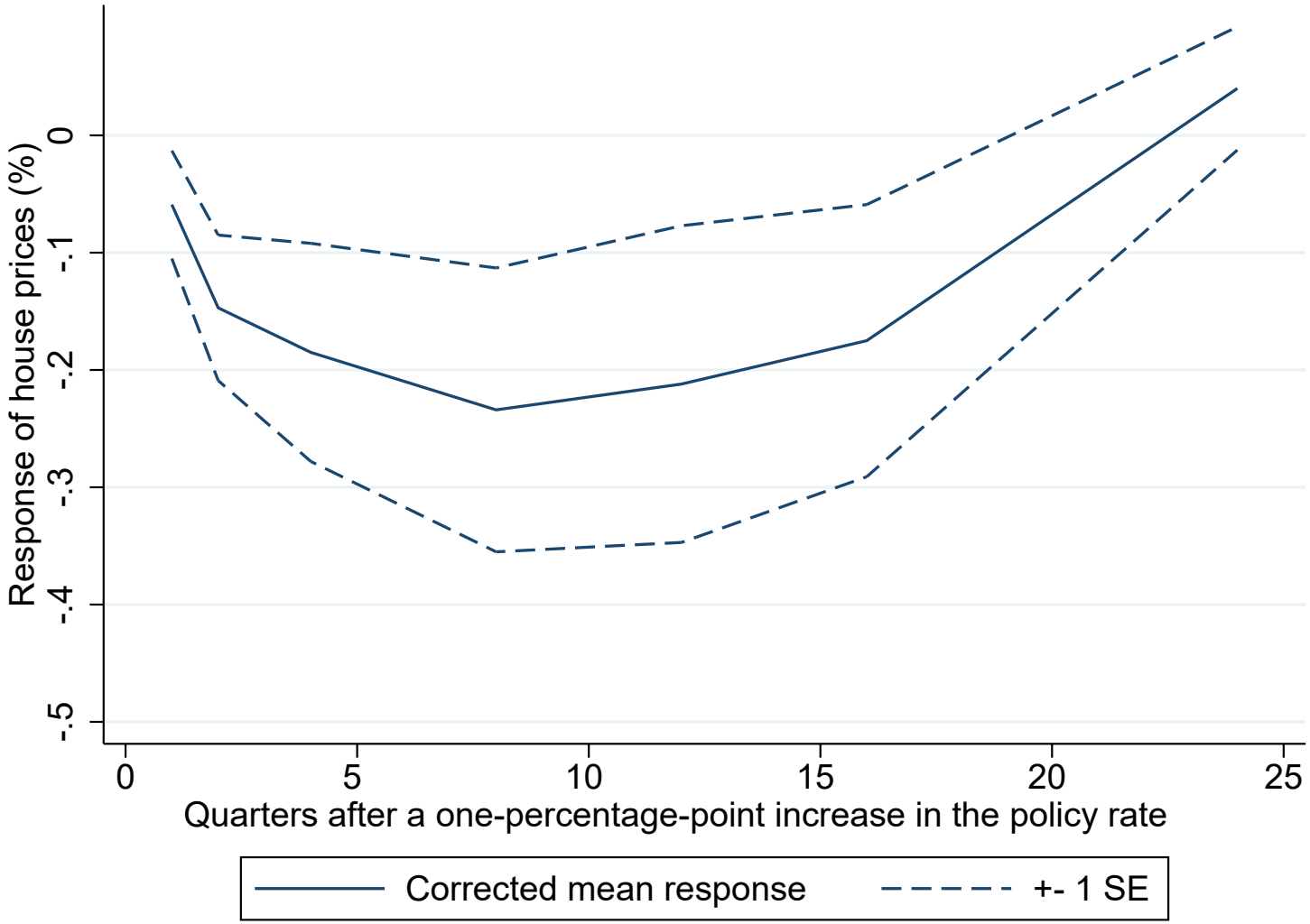
Time after a monetary policy shock:	1 quarter	2 quarters	4 quarters	8 quarters	12 quarters	16 quarters
PANEL A: Linear models						
Regression of reported estimates on their standard errors, ordinary least squares						
Standard error (publication bias)	-0.815*	-1.117***	-1.353***	-1.160***	-0.667**	-0.375*
	(0.463)	(0.367)	(0.427)	(0.308)	(0.317)	(0.215)
	[-1.220, -.015]	[-1.987, -.351]	[-2.373, -.410]	[-2.051, -.150]	[-1.542, .009]	[-1.281, .163]
Constant (corrected mean effect)	-0.014	-0.020	-0.015	-0.233	-0.501**	-0.560***
	(0.168)	(0.185)	(0.248)	(0.219)	(0.252)	(0.201)
	[-.213, .145]	[-.417, .479]	[-.576, .739]	[-.790, .444]	[-1.072, .154]	[-.968, -.039]
Regression of reported estimates on their standard errors, weighted by inverse variance						
Standard error (publication bias)	-0.705***	-0.874***	-1.092***	-1.160***	-0.964***	-0.732***
	(0.179)	(0.147)	(0.203)	(0.204)	(0.234)	(0.199)
	[-1.228, -.393]	[-1.207, -.586]	[-1.555, -.730]	[-1.629, -.710]	[-1.534, -.497]	[-1.343, -.337]
Constant (corrected mean effect)	-0.059	-0.147**	-0.185**	-0.234*	-0.212	-0.175
	(0.046)	(0.062)	(0.093)	(0.121)	(0.135)	(0.116)
	[-.058, .014]	[-.303, .081]	[-.360, .130]	[-.496, .044]	[-.521, -.030]	[-.533, -.010]

Nonliner tests also imply bias

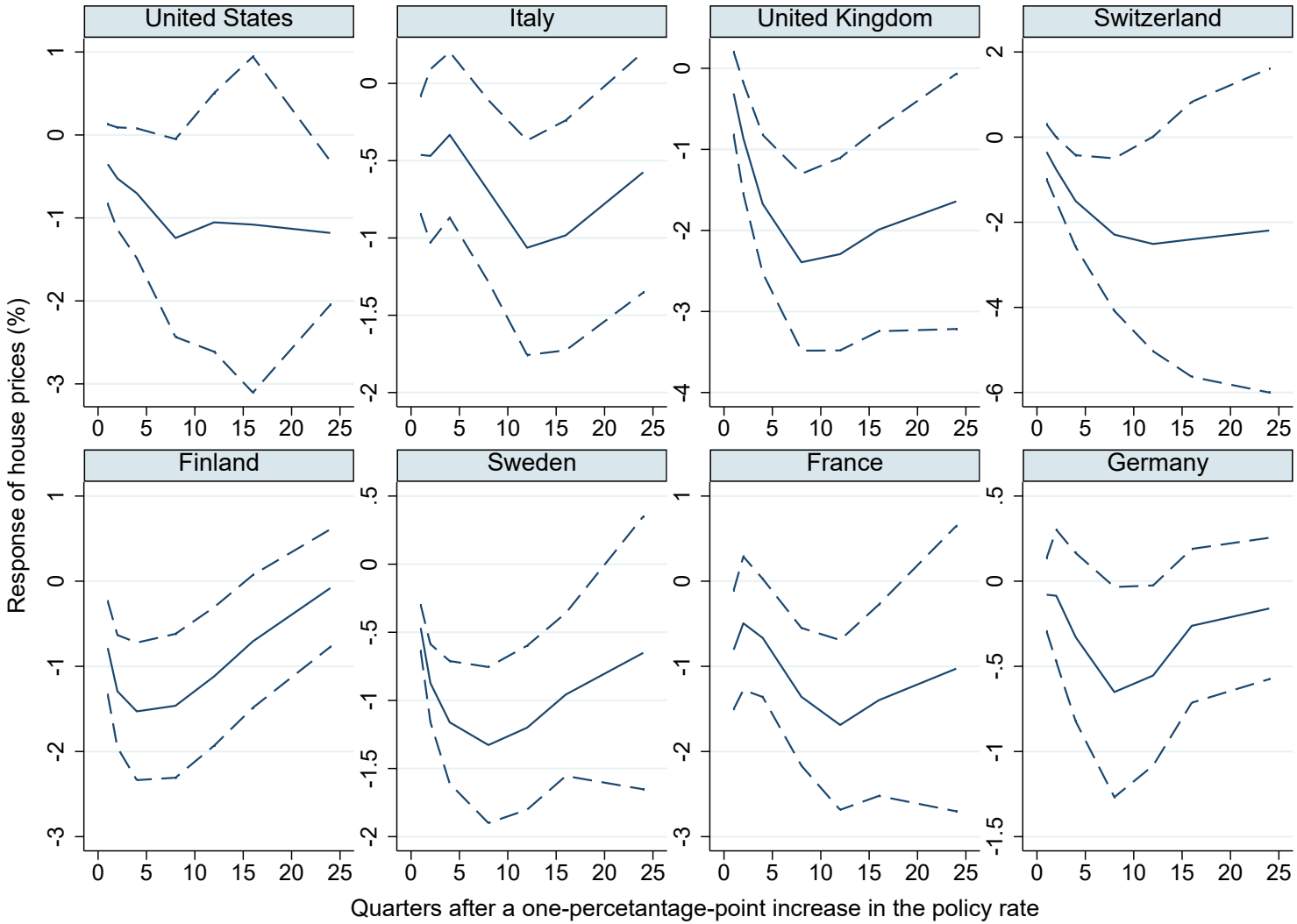
Note that the corrected effect is smaller than the reported mean effect:

Time after a monetary policy shock:	1 quarter	2 quarters	4 quarters	8 quarters	12 quarters	16 quarters
PANEL B: Nonlinear models						
Stem-based method (Furukawa, 2021)						
Corrected mean effect	-0.017	-0.192***	-0.318***	-0.324*	-0.172	-0.146**
	(0.053)	(0.074)	(0.116)	(0.185)	(0.145)	(0.071)
Selection model (Andrews & Kasy, 2019)						
Corrected mean effect, break at $t = 1.645$	-0.004***	-0.155*	-0.361***	-0.404**	-0.205	-0.065
	(0.002)	(0.094)	(0.079)	(0.190)	(0.187)	(0.115)
Corrected mean effect, break at $t = 1$	-0.023	-0.008	-0.213**	-0.149	-0.182**	-0.030
	(0.027)	(0.234)	(0.130)	(0.194)	(0.110)	(0.114)
P-uniform (van Aert & van Assen, 2021)*						
Corrected mean effect, break at $t = 1.645$	-0.133***	-0.121***	-0.140***	-0.134***	-0.118***	-0.095***
Corrected mean effect, break at $t = 1$	-0.072***	-0.091***	-0.103***	-0.098***	-0.093***	-0.075***

Mean impulse response beyond publication bias



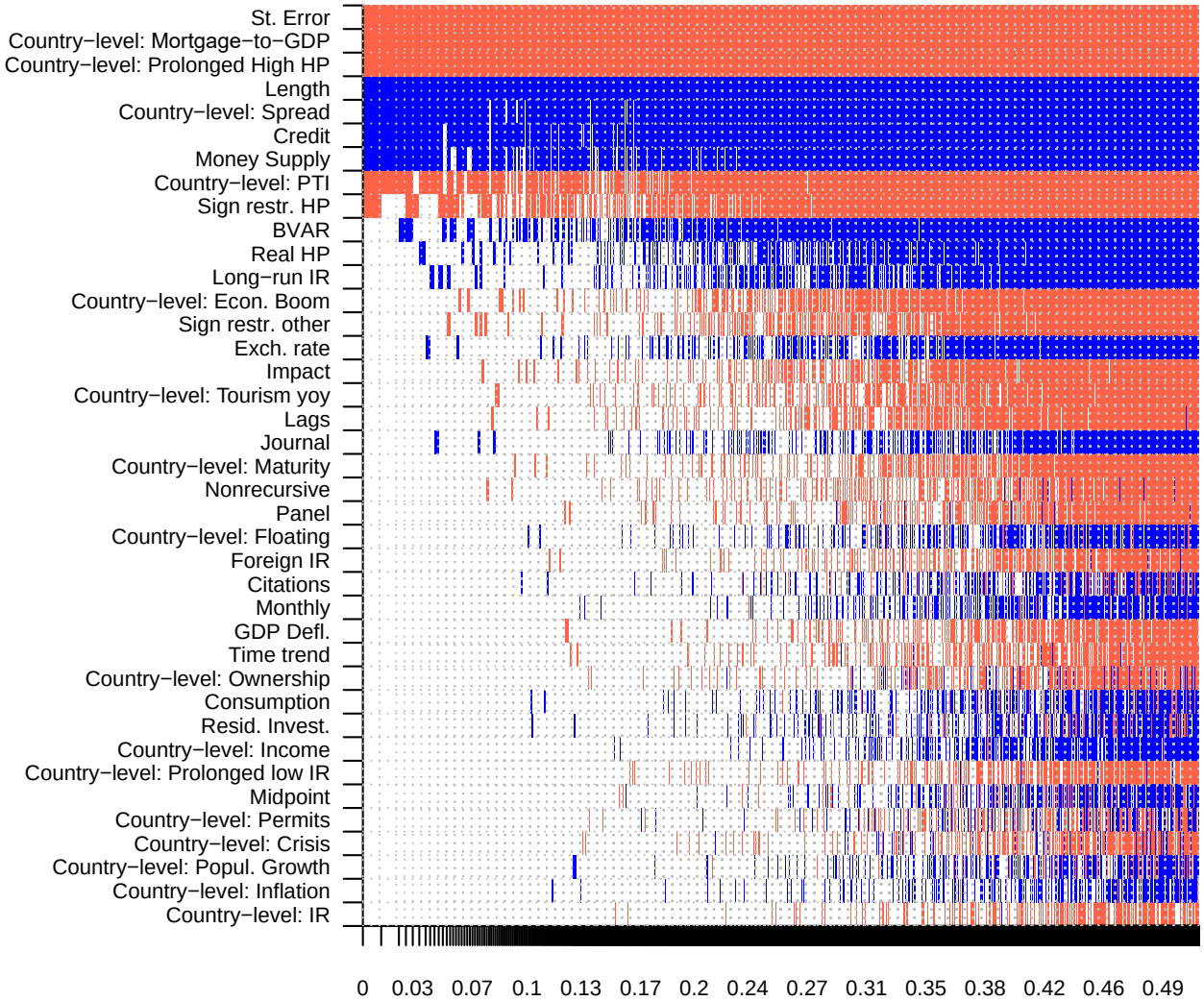
Cross-country heterogeneity



Estimation context

- **Data characteristics**: monthly, panel, length, age
- **Specification characteristics**: foreign IR, credit, consumption, res. invest., money supply, exch. rate, long-run IR, lags
- **Estimation characteristics**: BVAR, sign restr. HP, sign restr. other, nonrecursive
- **Publication characteristics**: citations, impact, published
- **Country characteristics**: mortgage-to-GDP, floating, maturity, IR, spread, prolonged high HP, PTI, crisis, prolonged low IR, tourism, income, inflation, popul. growth, permits, ownership

Bayesian model averaging

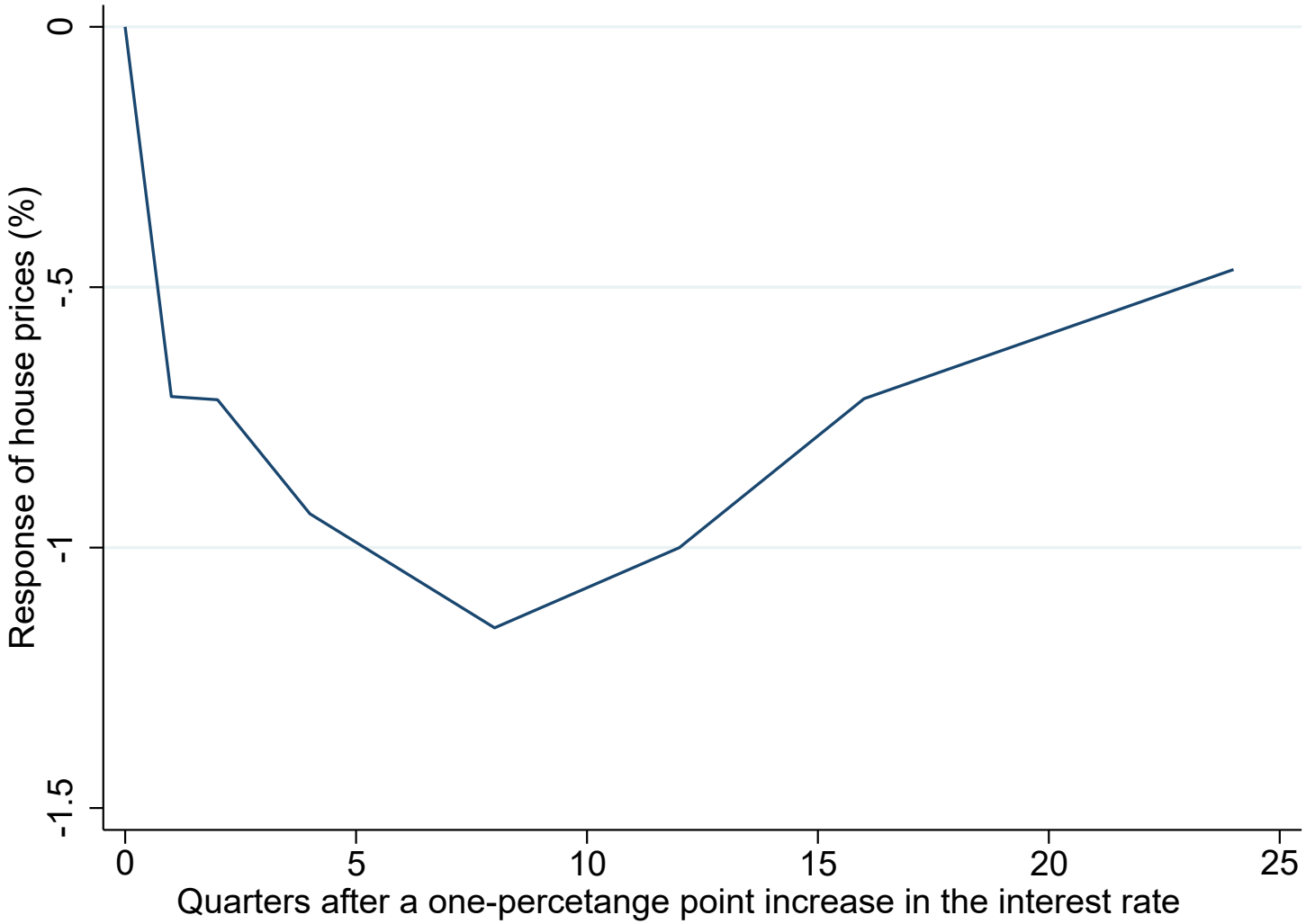


Best practice

We construct a hypothetical ideal study: no publication bias (SE = 0), recent data, long series, nonrecursive identification, published, good journal, highly cited.

Horizon of the corresponding impulse response:	1Q	2Q	4Q	8Q	12Q	16Q
Agnostic on control variables (baseline)	-0.710	-0.716	-0.935	-1.154	-1.000	-0.714
Control for credit, money supply, and long-run IR	-0.016	-0.023	-0.241	-0.461	-0.306	-0.021
Excluding credit, money supply, and long-run IR	-0.920	-0.927	-1.145	-1.365	-1.210	-0.925
Finland	-0.576	-0.583	-0.801	-1.021	-0.866	-0.581
France	-1.477	-1.484	-1.702	-1.922	-1.767	-1.482
Germany	-0.497	-0.503	-0.722	-0.941	-0.787	-0.501
Italy	0.093	0.087	-0.132	-0.352	-0.197	0.088
United Kingdom	-1.344	-1.351	-1.569	-1.789	-1.634	-1.349
United States	-0.910	-0.916	-1.135	-1.354	-1.200	-0.914

Impulse response implied by best practice



Results

Main Findings

1. Among 237 impulse responses from 37 studies, the mean decrease in house prices is 1.2% two years after a 1 pp. increase in the policy rate.
2. The mean is exaggerated because of publication bias.
3. Stronger transmission for i) more developed markets, ii) flatter yield curve, iii) house price boom.

Project Website

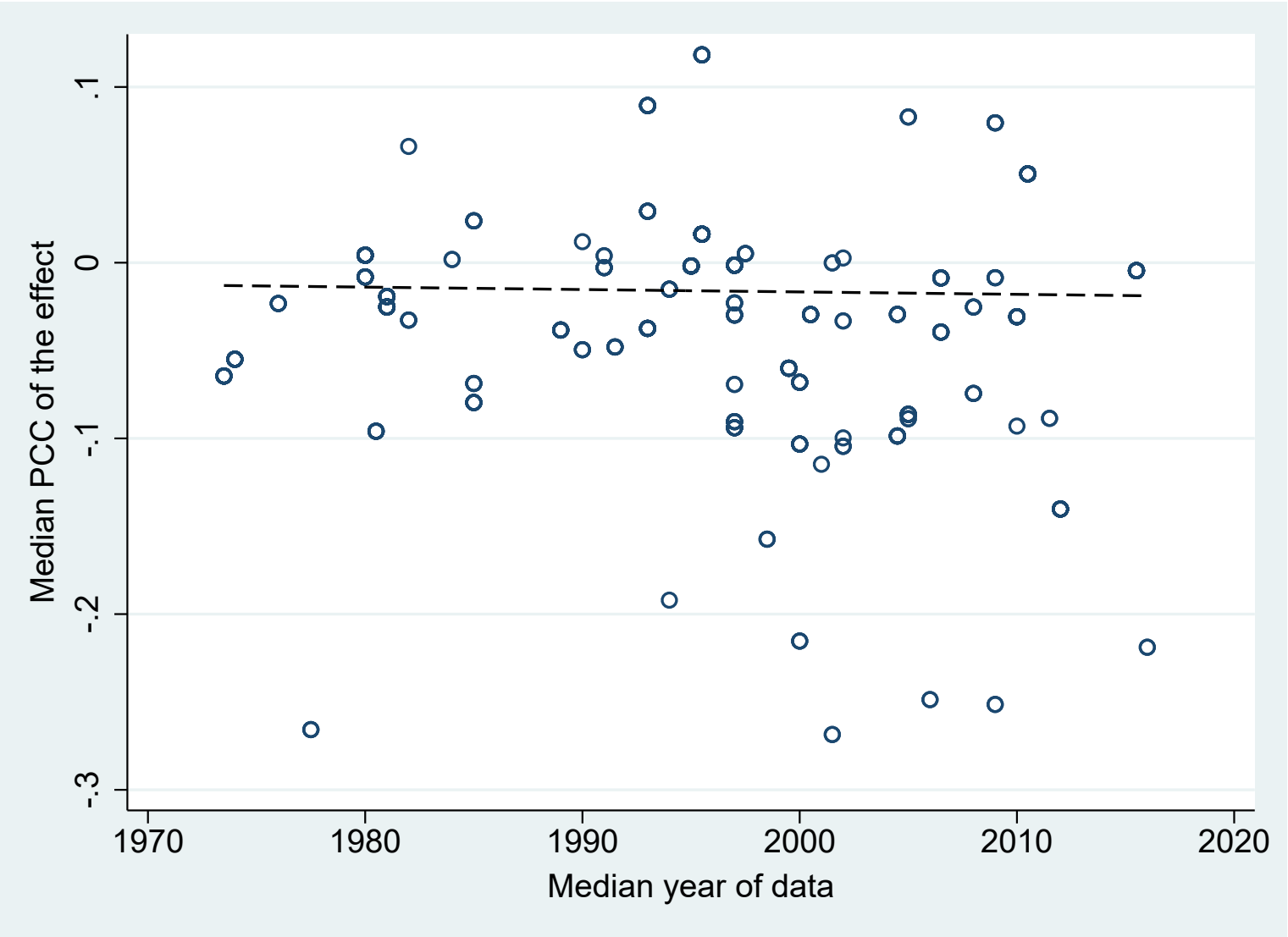
www.meta-analysis.cz/house_prices

Published in the *IMF Economic Review*, 2023.

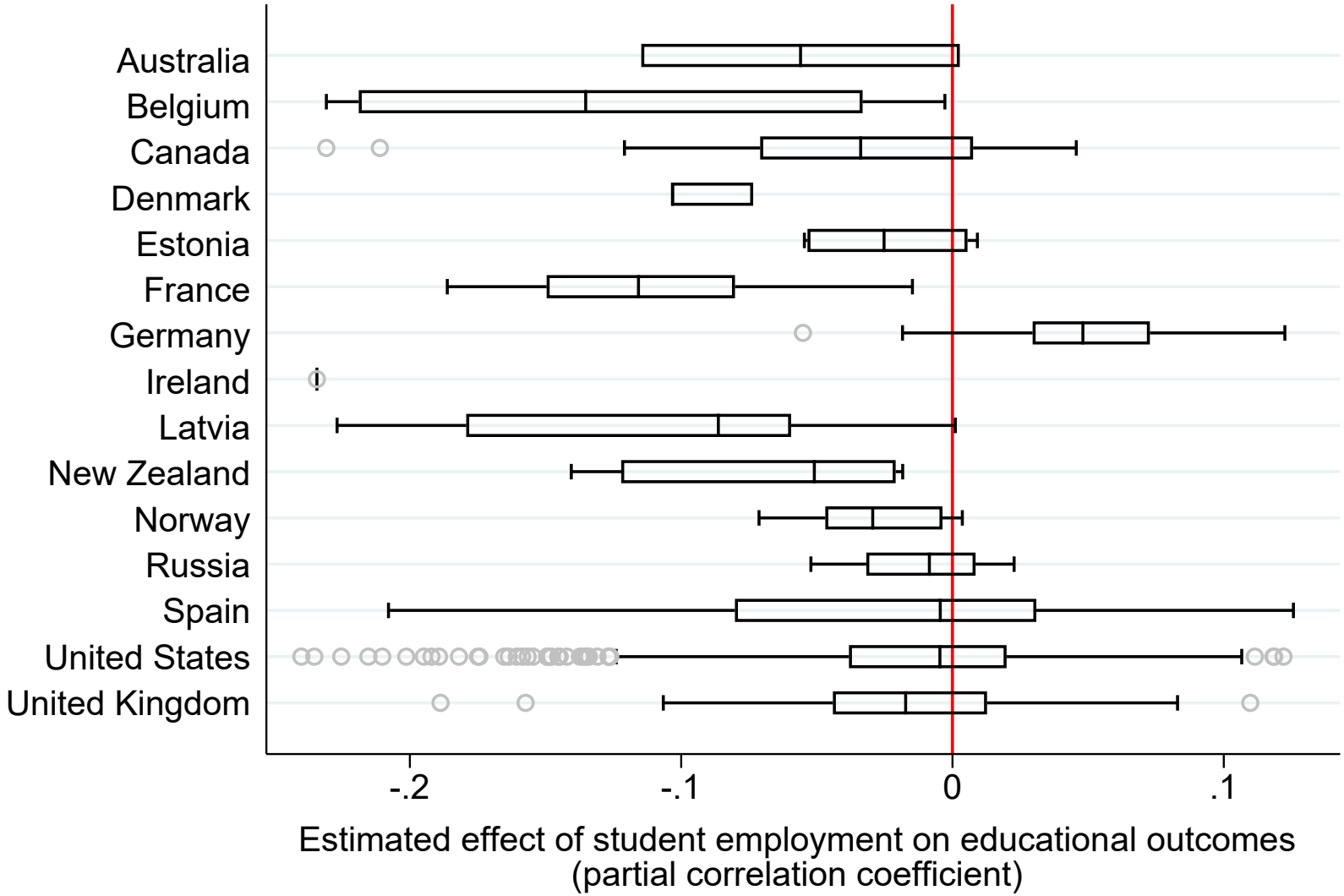
Example 3: Working while in school

- Hard to influence commonly cited determinants of educational outcomes: ability, ethnicity, gender, parental education and affluence
- But most students can decide whether or not to work
- How does that influence educational outcomes?
- Should you motivate your kids to work while in school?
- Mixed empirical results: from large negative to positive estimates
- 861 estimates from 69 studies

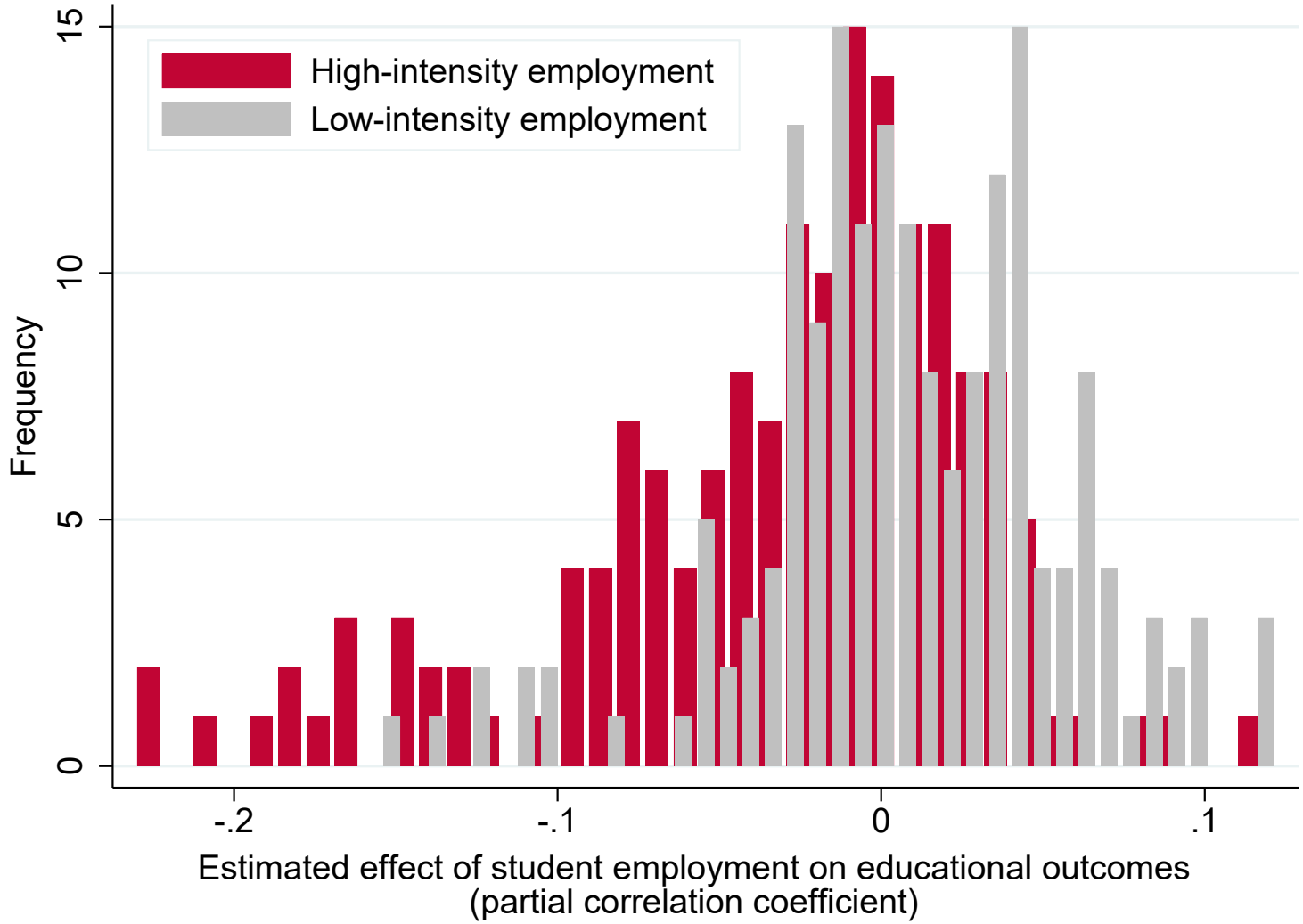
No consensus in the literature



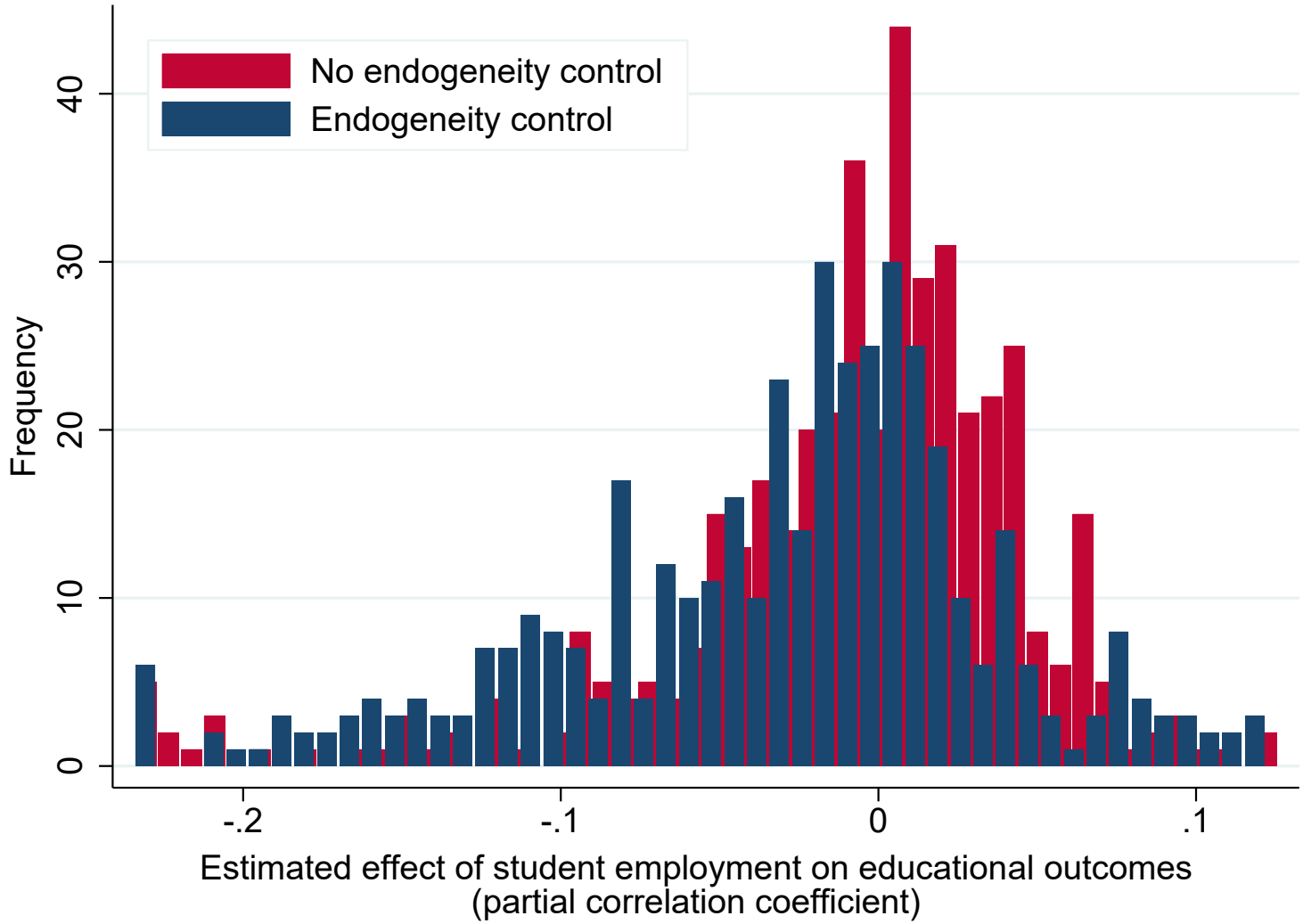
Negative for all countries except Germany



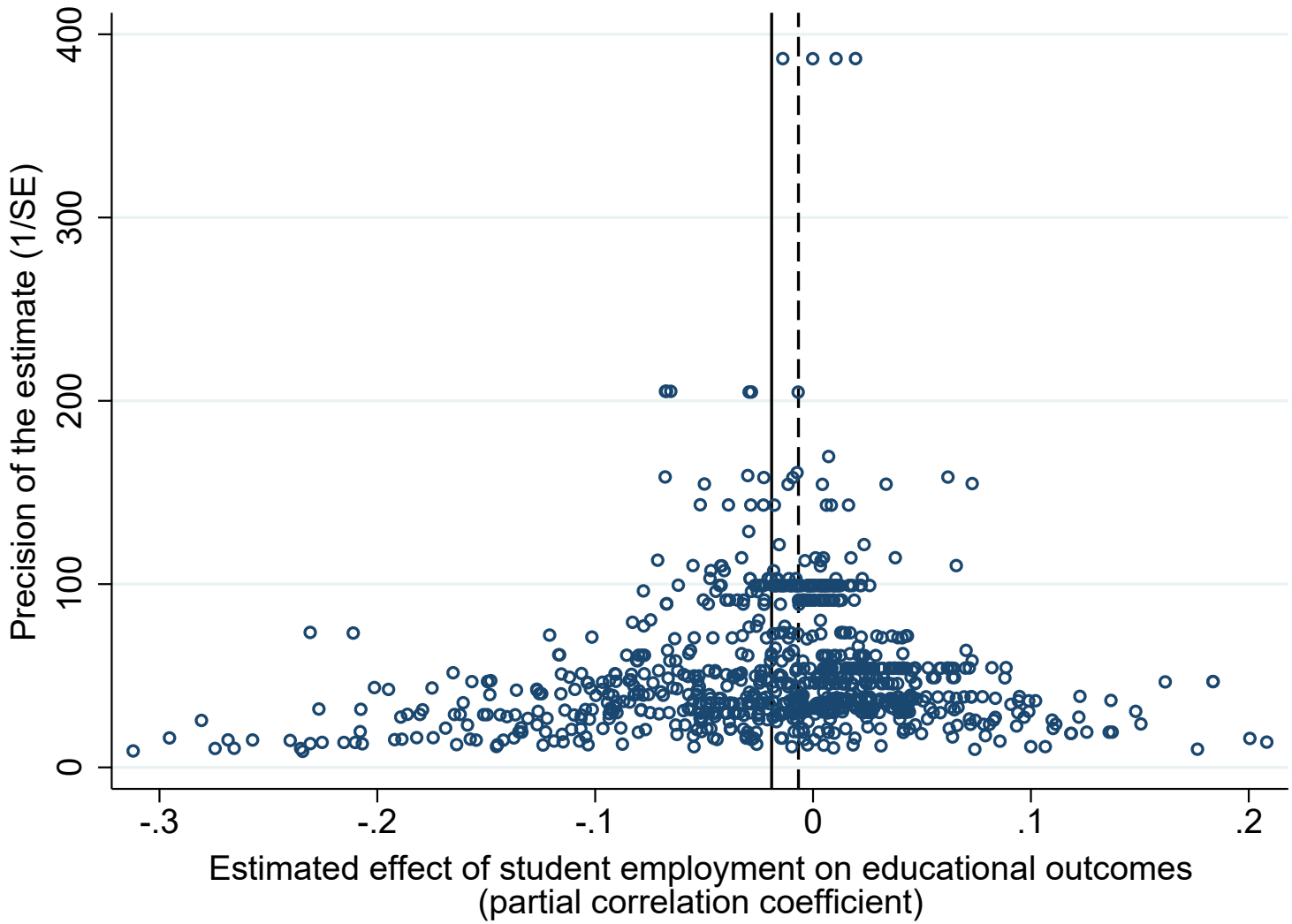
Work intensity clearly matters



Endogeneity control is important



Little publication bias on average



No bias in studies addressing endogeneity

[Block 2] Studies trying to take endogeneity into account

Panel A: Linear techniques

	OLS	IV	Study	Precision
Standard error *(Publication bias)*	-0.311 (0.347) [-1.174, 0.710]	-0.311 (0.392) [-1.306, 0.829]	-0.244 (0.431) [-1.556, 0.987]	-0.449 (0.387) [-1.3, 0.489]
Constant *(Effect beyond bias)*	-0.0189** (0.00932) [-0.040, -0.0001]	-0.0185* (0.00984) [-0.039, -0.0002]	-0.0218 (0.0159) [-0.057, 0.011]	-0.0151** (0.00751) [-0.034, 0.002]
Observations	425	425	425	425

Panel B: Between- and within-study variation

	BE	FE	RE
Standard error *(Publication bias)*	-1.334*** (0.457)	1.041*** (0.373)	0.235 (0.287)
Constant *(Effect beyond bias)*	0.00317 (0.0169)	-0.0556*** (0.0101)	-0.0431*** (0.0126)
Observations	425	425	425

Panel C: Nonlinear techniques

	WAAP	Stem method	Kinked model	Selection model	p-uniform*
Effect beyond bias	-0.0200*** (0.00687)	0.00996 (0.0294)	-0.0138*** (0.00369)	-0.0260*** (0.008)	-0.0319*** (0.0106)
Observations	425	425	425	425	425

Strong bias in studies ignoring endogeneity

[Block 1] Studies ignoring endogeneity

Panel A: Linear techniques

	OLS	IV	Study	Precision
Standard error	-1.858***	-1.945***	-2.420***	-0.959**
(Publication bias)	(0.457)	(0.483)	(0.545)	(0.397)
	[-3.189, -0.866]	[-3.23, -0.921]	[-3.83, -1.17]	[-2.058, -0.24]
Constant	0.0405**	0.0425**	0.0591***	0.0171
(Effect beyond bias)	(0.0172)	(0.0176)	(0.0189)	(0.0112)
	[0.0001, 0.084]	[0.002, 0.086]	[0.0003, 0.105]	[-0.001, 0.067]
Observations	436	436	436	436

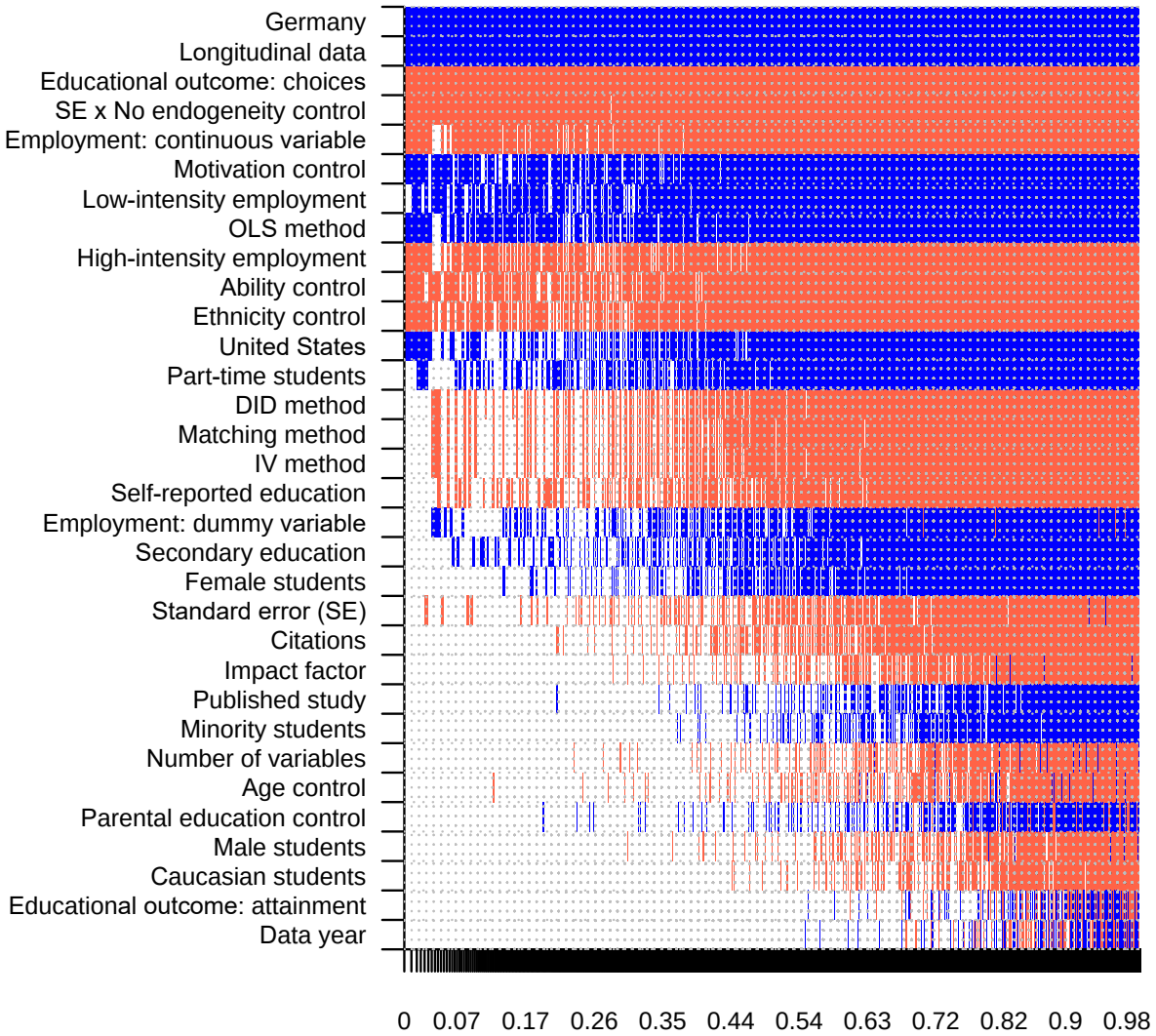
Panel B: Between- and within-study variation

	BE	FE	RE
Standard error	-2.535***	-0.903	-1.310***
(Publication bias)	(0.489)	(0.794)	(0.255)
Constant	0.0331	0.0156	-0.0111
(Effect beyond bias)	(0.0217)	(0.0207)	(0.0146)
Observations	436	436	436

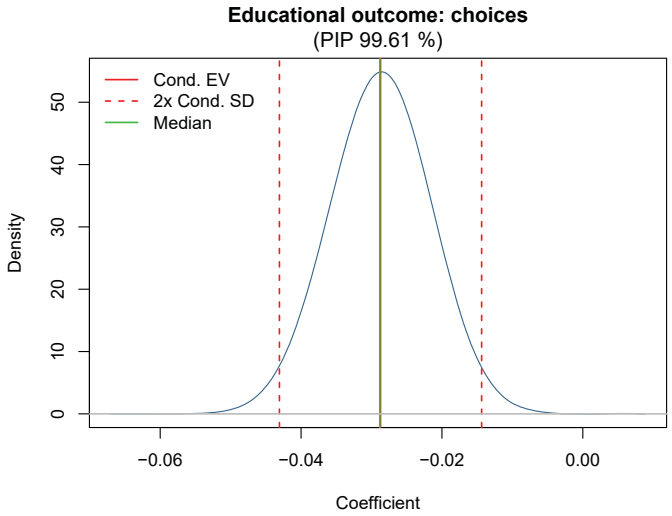
Panel C: Nonlinear techniques

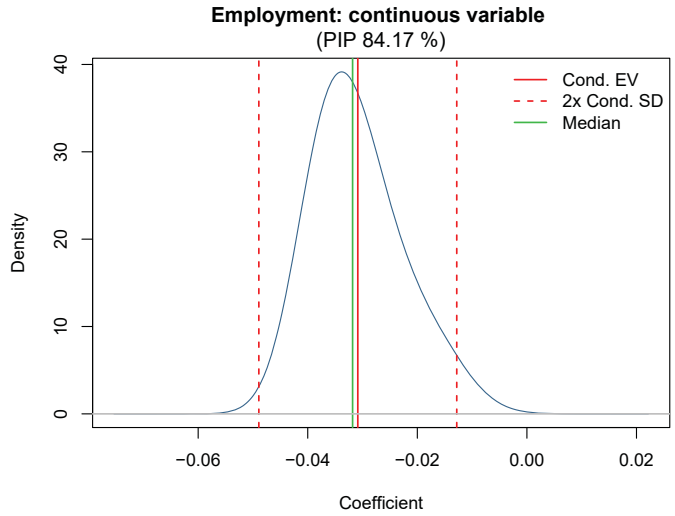
	WAAP	Stem method	Kinked model	Selection model	p-uniform*
Effect beyond bias	·	-0.00959**	0.00187	0.000	·
	(·)	(0.00432)	(0.00386)	(0.00500)	(·)
Observations	436	436	436	436	436

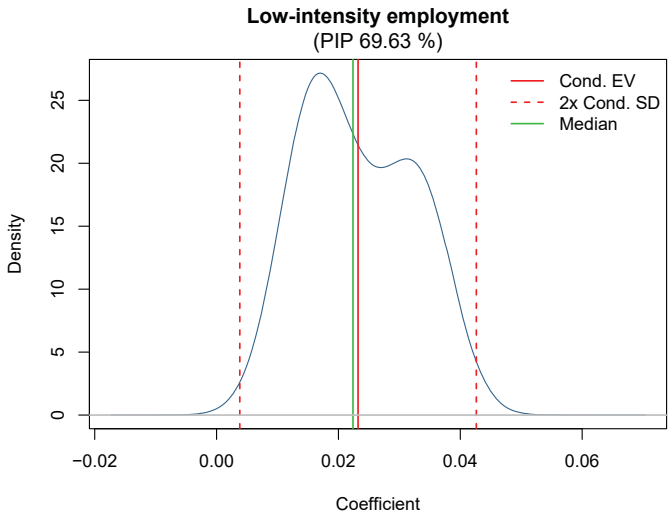
Model uncertainty; endogeneity control is key

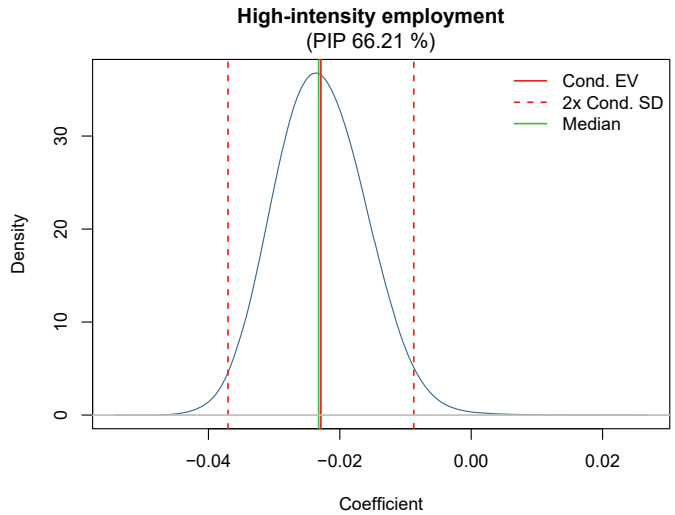


Closer look: work intensity, measurement

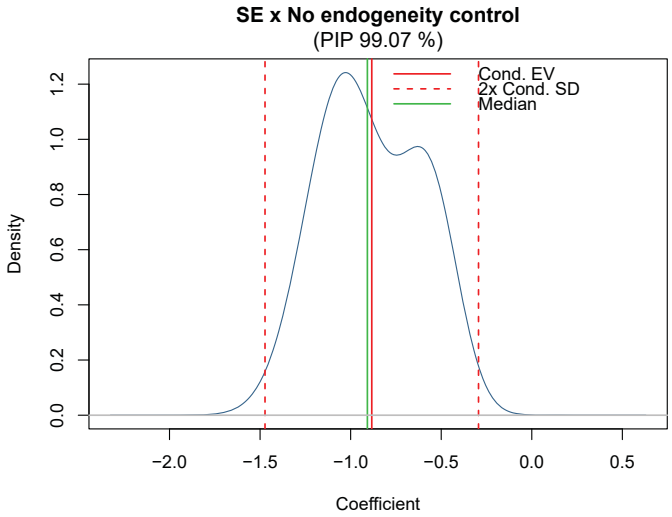


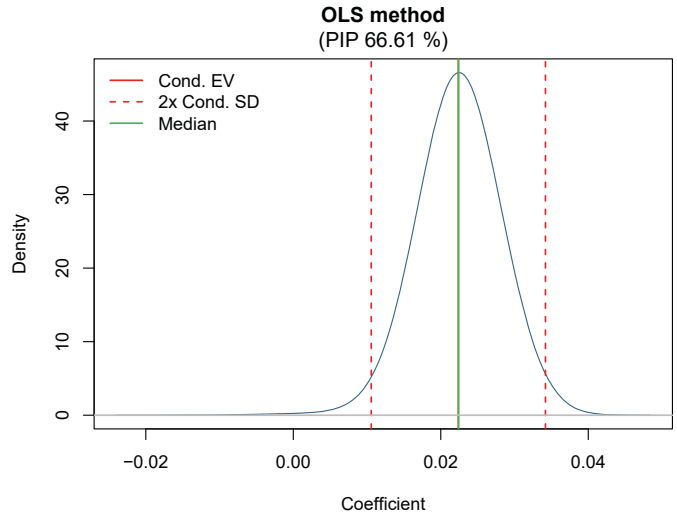


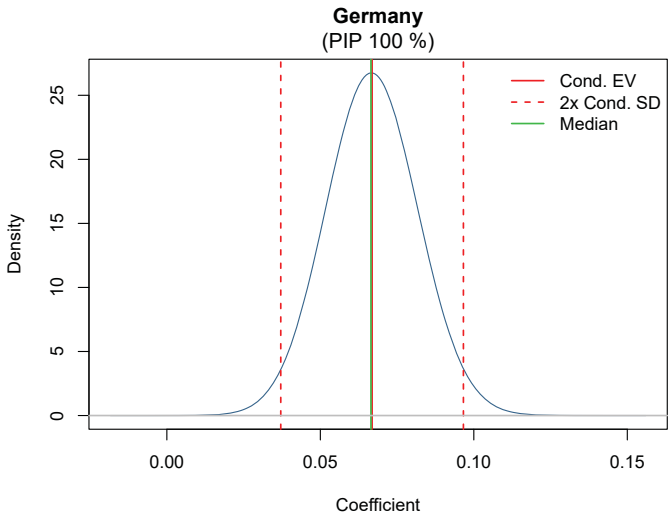


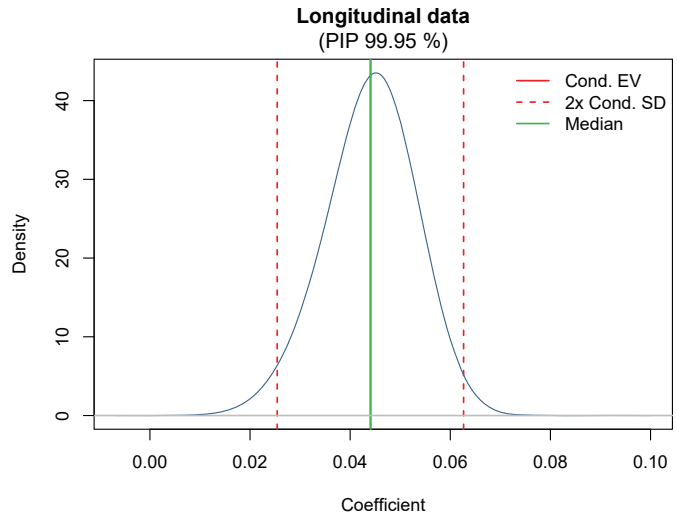


Closer look: endogeneity, panel data, Germany









Best practice estimates

	Mean	95% conf. int.	
USA	-0.039	-0.078	0.001
Germany	0.019	-0.046	0.083
Other countries	-0.048	-0.104	0.008
Male students	-0.039	-0.083	0.005
Female students	-0.037	-0.079	0.005
Part-time students	-0.025	-0.069	0.020
Low-intensity employment	-0.023	-0.068	0.022
High-intensity employment	-0.054	-0.098	-0.010
Educational outcome: choices	-0.062	-0.107	-0.017
Educational outcome: test scores	-0.033	-0.075	0.009
Educational outcome: attainment	-0.033	-0.080	0.014
Overall effect	-0.038	-0.079	0.003

Results

Main Findings

1. Publication and endogeneity biases interact.
2. The corrected effect is negative but economically zero.
3. Effects negative for high-intensity work, positive for Germany.

Project Website

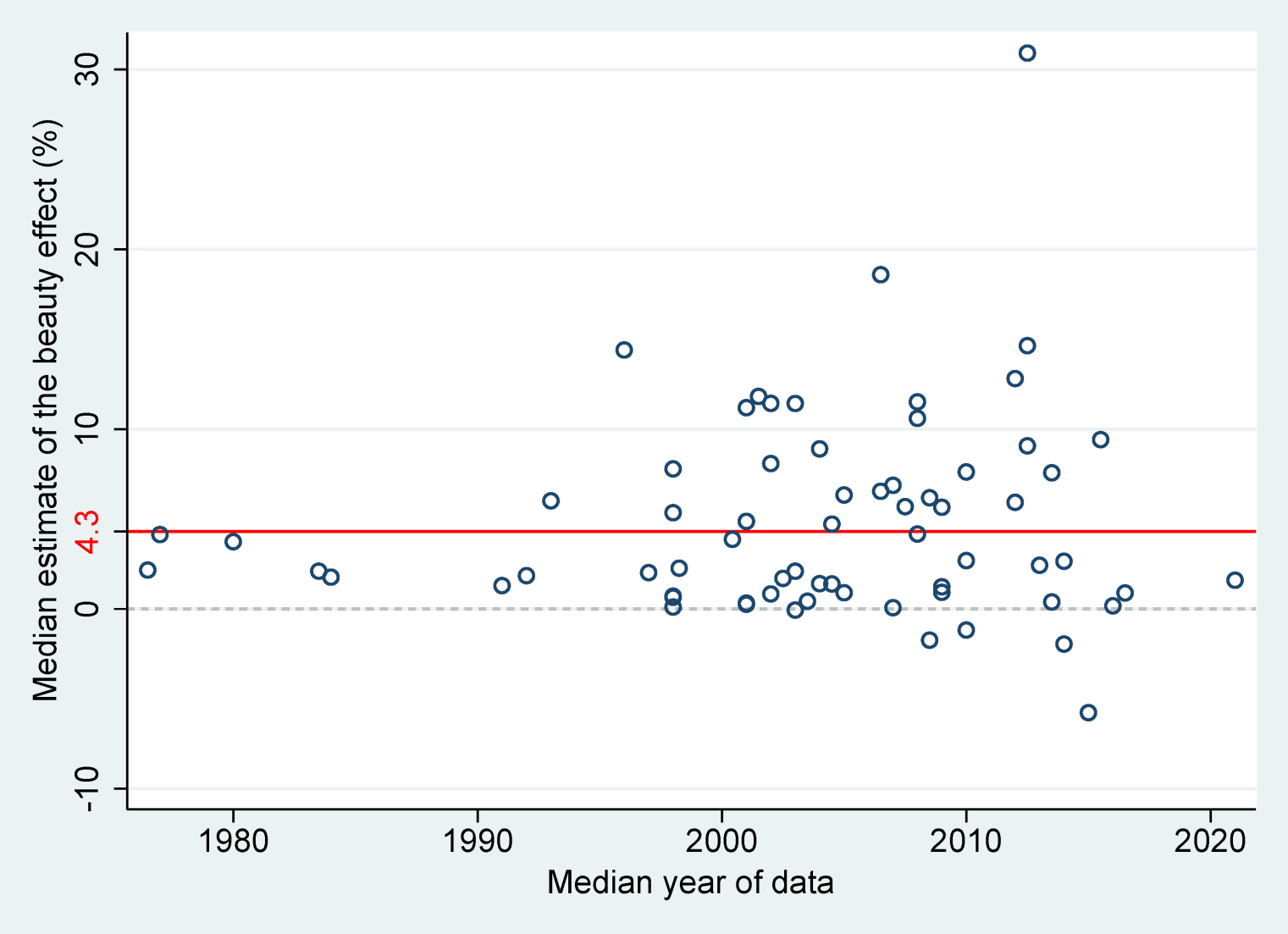
www.meta-analysis.cz/students

Published in the *Economics of Education Review*, 2024.

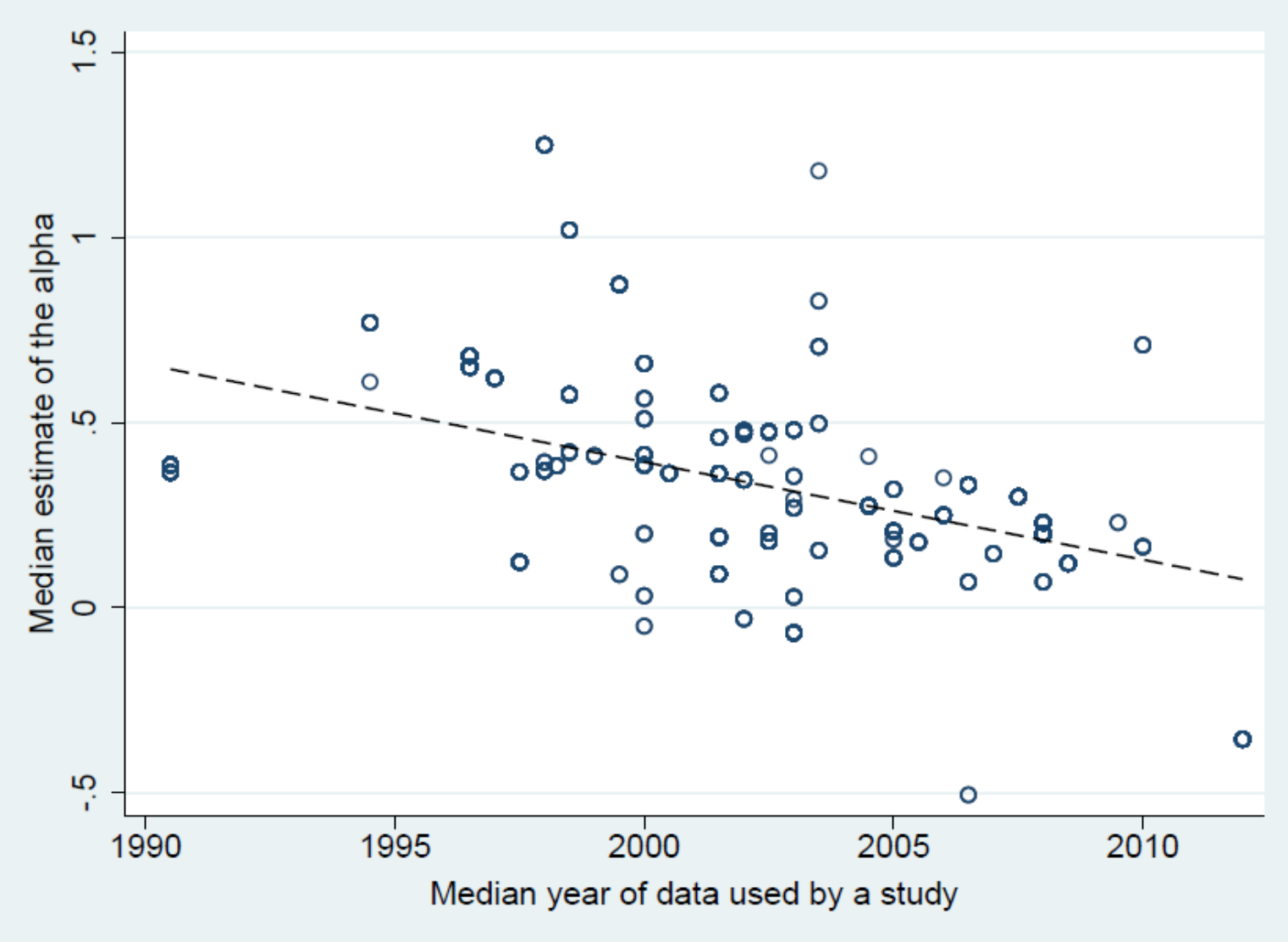
Workshop Outline

Inspiration

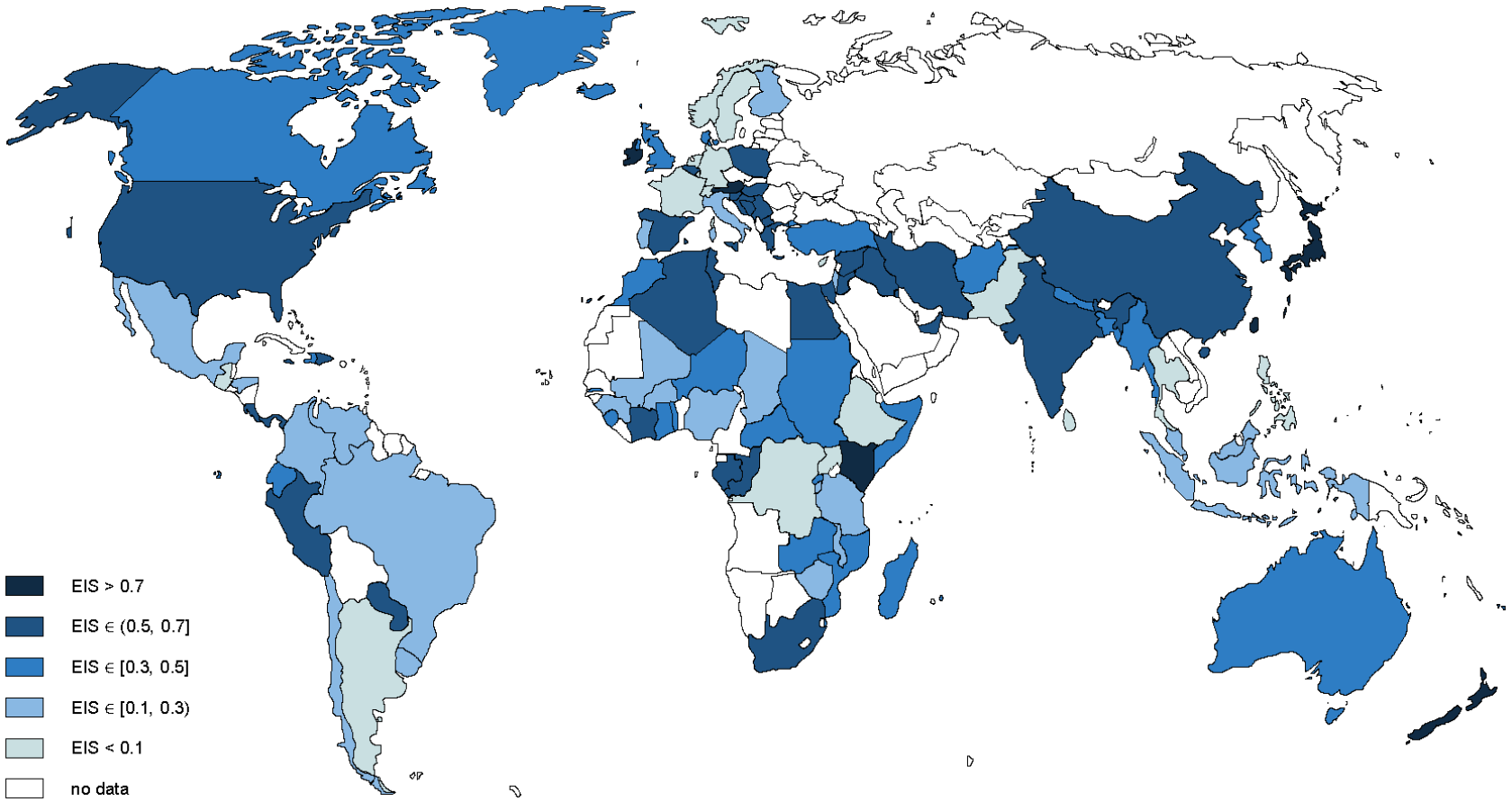
Motivation example: increasing uncertainty



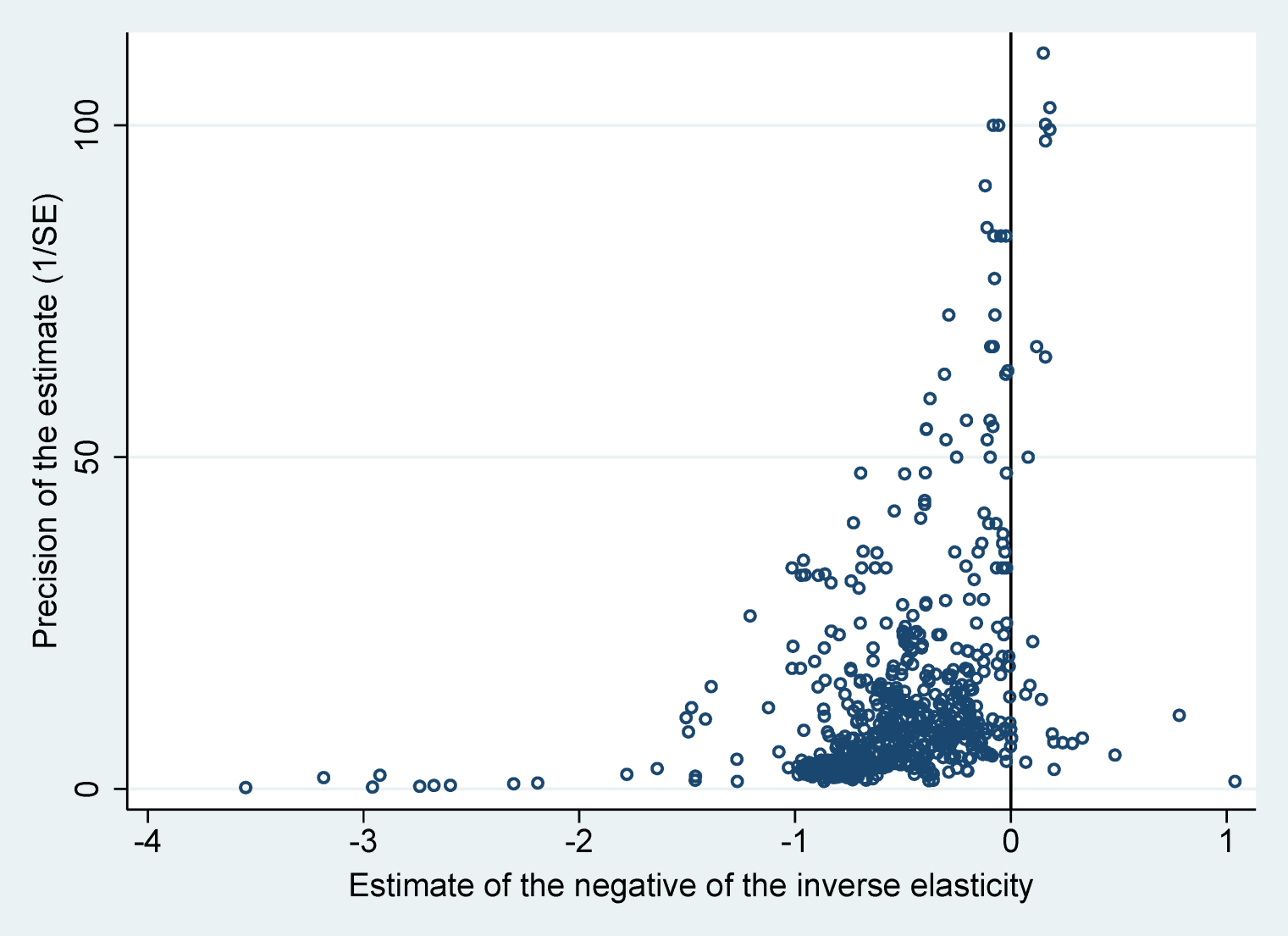
Motivation example: clear trend



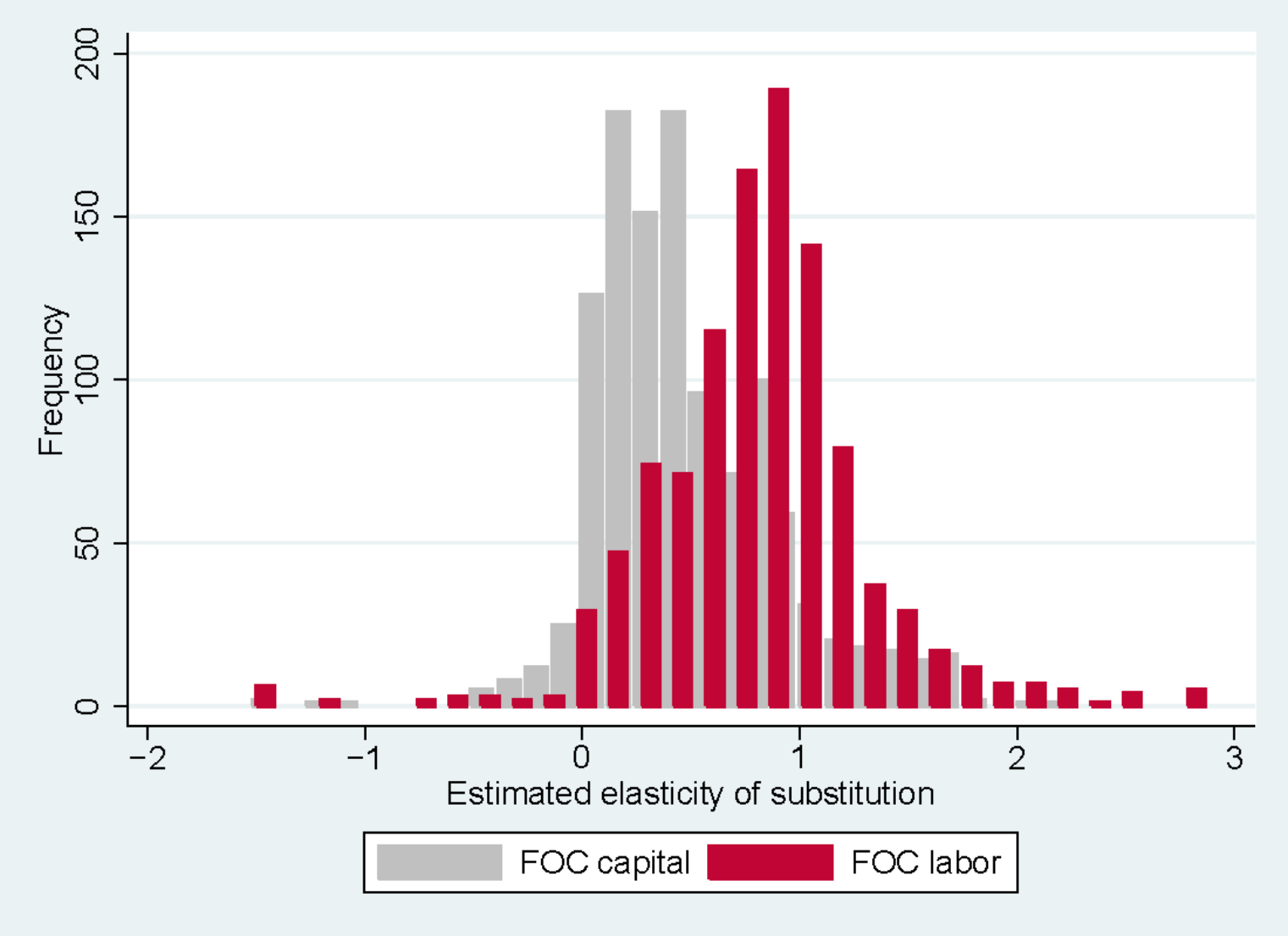
Motivation example: country-level differences



Motivation example: publication bias



Motivation example: method heterogeneity



We use Google Scholar

Advantages:

- Full text, not just the abstract, title, keywords
- Inclusive
- Up-to-date

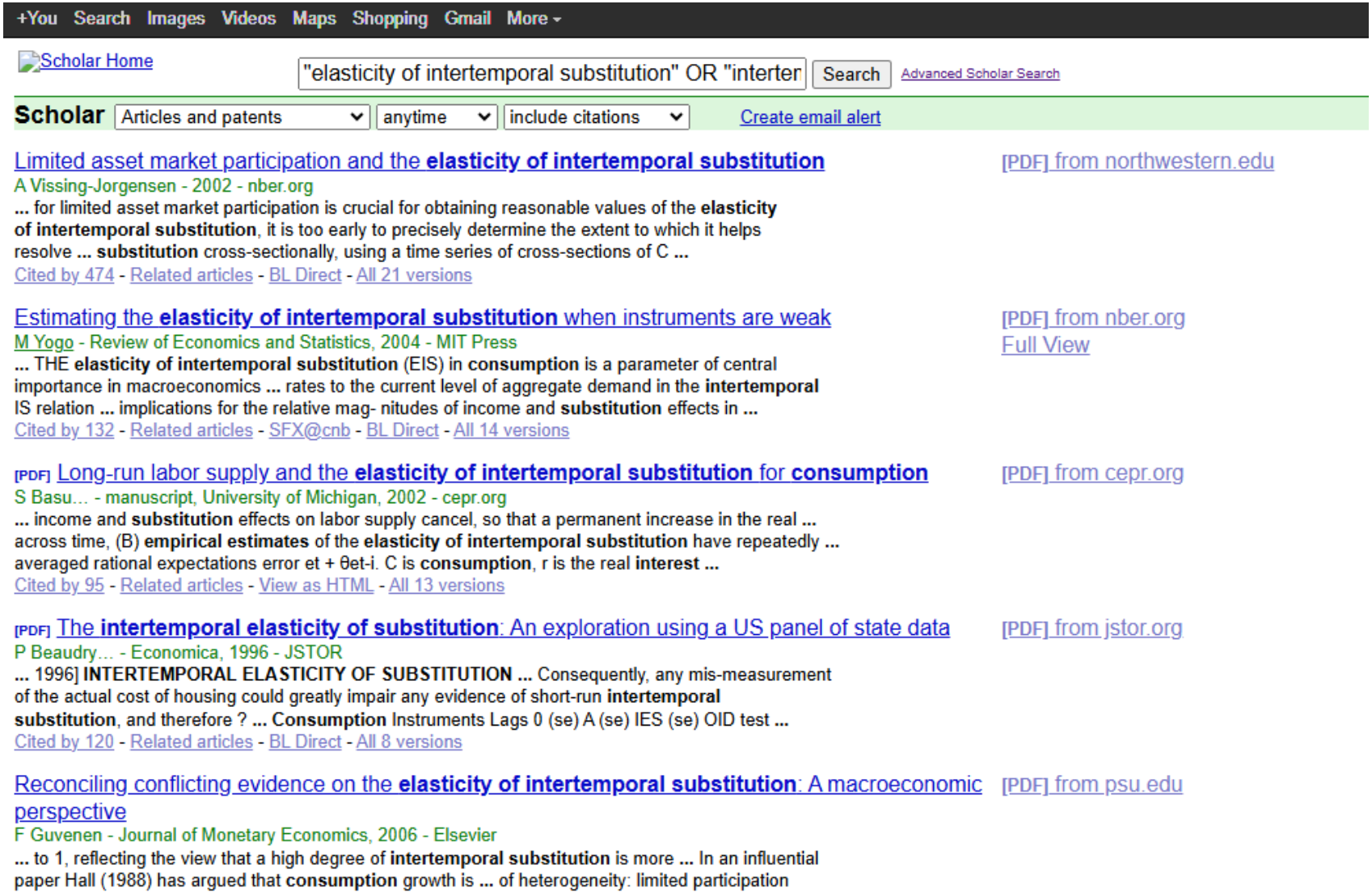
Disadvantages:

- Algorithm changes in time (replicability issues)
- Too many results
- Sometimes relevance unclear

Search steps (without AI for now)

1. Find 5 most relevant studies
2. Design a query that shows the 5 studies near the top
3. Go through first 500 hits
4. Read abstracts
5. Download potentially useful studies
6. Skim them
7. Repeat the procedure for the most recent studies (last 3 years)

Search query 1



+You Search Images Videos Maps Shopping Gmail More

Scholar Home "elasticity of intertemporal substitution" OR "interter Search Advanced Scholar Search

Scholar Articles and patents anytime include citations Create email alert

Limited asset market participation and the **elasticity of intertemporal substitution** [PDF] from northwestern.edu
A Vissing-Jorgensen - 2002 - nber.org
... for limited asset market participation is crucial for obtaining reasonable values of the **elasticity of intertemporal substitution**, it is too early to precisely determine the extent to which it helps resolve ... **substitution** cross-sectionally, using a time series of cross-sections of C ...
Cited by 474 - Related articles - BL Direct - All 21 versions

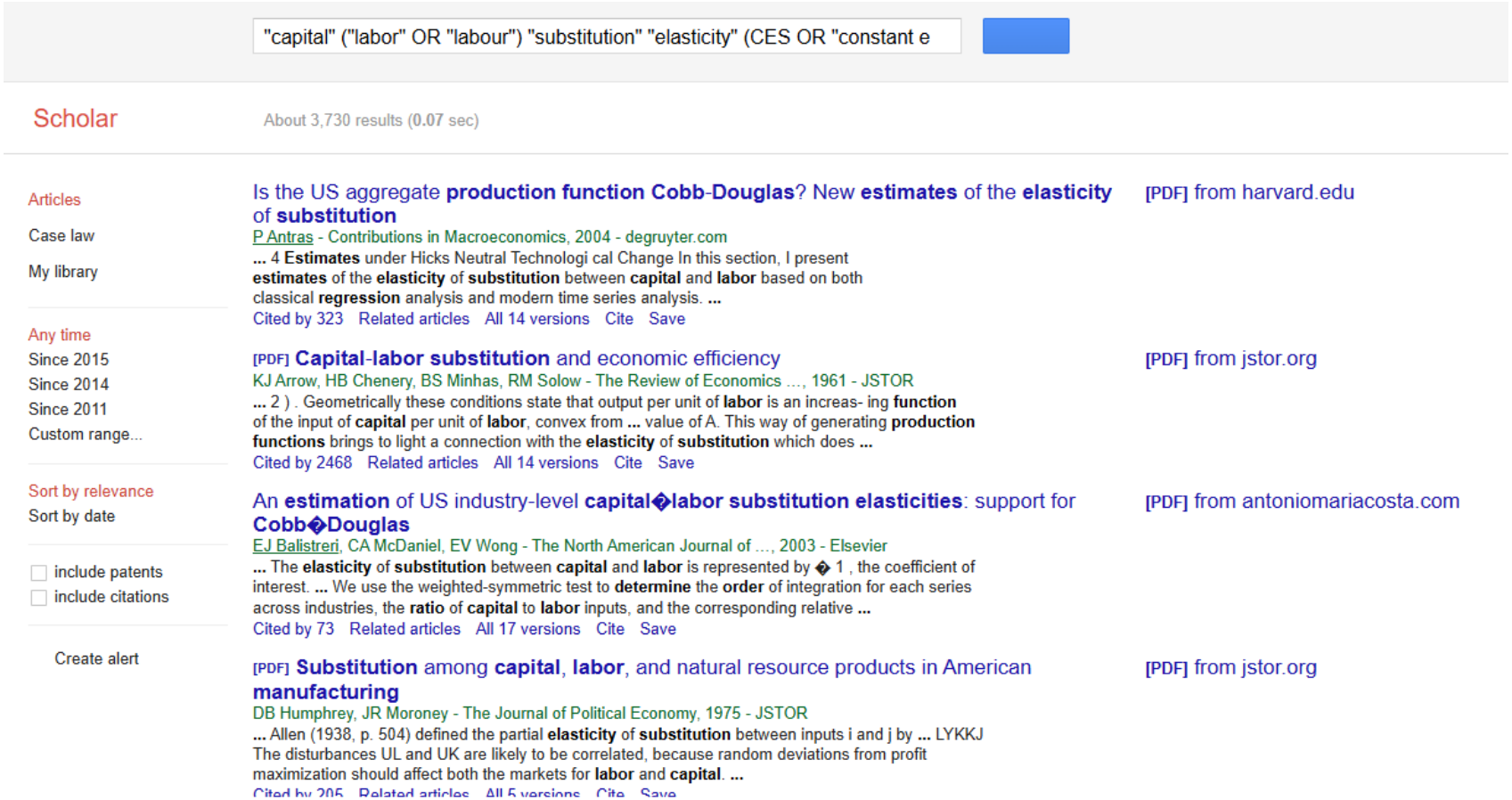
Estimating the **elasticity of intertemporal substitution** when instruments are weak [PDF] from nber.org
M Yogo - Review of Economics and Statistics, 2004 - MIT Press Full View
... THE **elasticity of intertemporal substitution** (EIS) in **consumption** is a parameter of central importance in macroeconomics ... rates to the current level of aggregate demand in the **intertemporal** IS relation ... implications for the relative mag- nitudes of income and **substitution** effects in ...
Cited by 132 - Related articles - SFX@cnb - BL Direct - All 14 versions

[PDF] Long-run labor supply and the **elasticity of intertemporal substitution** for **consumption** [PDF] from cepr.org
S Basu... - manuscript, University of Michigan, 2002 - cepr.org
... income and **substitution** effects on labor supply cancel, so that a permanent increase in the real ... across time, (B) **empirical estimates** of the **elasticity of intertemporal substitution** have repeatedly ... averaged rational expectations error et + θet-i. C is **consumption**, r is the real **interest** ...
Cited by 95 - Related articles - View as HTML - All 13 versions

[PDF] The **intertemporal elasticity of substitution**: An exploration using a US panel of state data [PDF] from jstor.org
P Beaudry... - Economica, 1996 - JSTOR
... 1996] **INTERTEMPORAL ELASTICITY OF SUBSTITUTION** ... Consequently, any mis-measurement of the actual cost of housing could greatly impair any evidence of short-run **intertemporal substitution**, and therefore ? ... **Consumption** Instruments Lags 0 (se) A (se) IES (se) OID test ...
Cited by 120 - Related articles - BL Direct - All 8 versions

Reconciling conflicting evidence on the **elasticity of intertemporal substitution**: A macroeconomic perspective [PDF] from psu.edu
F Guvenen - Journal of Monetary Economics, 2006 - Elsevier
... to 1, reflecting the view that a high degree of **intertemporal substitution** is more ... In an influential paper Hall (1988) has argued that **consumption** growth is ... of heterogeneity: limited participation

Search query 2



"capital" ("labor" OR "labour") "substitution" "elasticity" (CES OR "constant e

Scholar

About 3,730 results (0.07 sec)

Articles
Case law
My library

Any time
Since 2015
Since 2014
Since 2011
Custom range...

Sort by relevance
Sort by date

include patents
include citations

Create alert

Is the US aggregate **production function Cobb-Douglas**? New **estimates** of the **elasticity** of **substitution** [PDF] from harvard.edu
P Antras - Contributions in Macroeconomics, 2004 - degruyter.com
... 4 **Estimates** under Hicks Neutral Technologi cal Change In this section, I present **estimates** of the **elasticity** of **substitution** between **capital** and **labor** based on both classical **regression** analysis and modern time series analysis. ...
Cited by 323 Related articles All 14 versions Cite Save

[PDF] **Capital-labor substitution** and economic efficiency [PDF] from jstor.org
KJ Arrow, HB Chenery, BS Minhas, RM Solow - The Review of Economics ..., 1961 - JSTOR
... 2) . Geometrically these conditions state that output per unit of **labor** is an increas- ing **function** of the input of **capital** per unit of **labor**, convex from ... value of A. This way of generating **production functions** brings to light a connection with the **elasticity** of **substitution** which does ...
Cited by 2468 Related articles All 14 versions Cite Save

An **estimation** of US industry-level **capital**&#65533;**labor substitution elasticities**: support for **Cobb**&#65533;**Douglas** [PDF] from antoniomariacosta.com
EJ Balistreri, CA McDaniel, EV Wong - The North American Journal of ..., 2003 - Elsevier
... The **elasticity** of **substitution** between **capital** and **labor** is represented by &#65533; 1 , the coefficient of interest. ... We use the weighted-symmetric test to **determine** the **order** of integration for each series across industries, the **ratio** of **capital** to **labor** inputs, and the corresponding relative ...
Cited by 73 Related articles All 17 versions Cite Save

[PDF] **Substitution** among **capital**, **labor**, and natural resource products in American **manufacturing** [PDF] from jstor.org
DB Humphrey, JR Moroney - The Journal of Political Economy, 1975 - JSTOR
... Allen (1938, p. 504) defined the partial **elasticity** of **substitution** between inputs i and j by ... LYKKJ The disturbances UL and UK are likely to be correlated, because random deviations from profit maximization should affect both the markets for **labor** and **capital**. ...
Cited by 205 Related articles All 5 versions Cite Save

Primary studies

idstudy	study	pubyear	outlet	title	type
1	Hansen and Singleton (1982)	1982	Econometrica	Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models	article
2	Hansen and Singleton (1983)	1983	Journal of Political Economy	Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns	article
3	Mankiw et al. (1985)	1985	Quarterly Journal of Economics	Intertemporal Substitution in Macroeconomics	article
4	Giovannini (1985)	1985	Journal of Development Economics	Saving and the real interest rate in LDCs	article
5	Miron (1986)	1986	Journal of Political Economy	Seasonal Fluctuations and the Life Cycle-Permanent Income Model of Consumption	article
6	Koskela and Viren (1987)	1987	Economics Letters	Inflation and the Euler equation approach to consumption behaviour: Some empirical evidence	article
7	Ferson and Merrick (1987)	1987	Journal of Financial Economics	Non-stationarity and stage-of-the-business-cycle effects in consumption-based asset pricing relations	article
8	Hall (1988)	1988	Journal of Political Economy	Intertemporal Substitution in Consumption	article
9	Whetley (1988)	1988	Journal of Monetary Economics	Some tests of the consumption-based asset pricing model	article
10	Kugler (1988)	1988	Economics Letters	Intertemporal substitution, taste shocks and cointegration	article
11	Rossi (1988)	1988	Staff Papers - International Monetary	Government Spending, the Real Interest Rate, and the Behavior of Liquidity-Constrained Consumers in De	article
12	Zeldes (1989)	1989	Journal of Political Economy	Consumption and Liquidity Constraints: An Empirical Investigation	article
13	Attanasio and Weber (1989)	1989	Economic Journal	Intertemporal Substitution, Risk Aversion and the Euler Equation for Consumption	article
14	Finn (1989)	1989	Journal of International Money and Fi	An econometric analysis of the intertemporal general-equilibrium approach to exchange rate and current	article
15	Campbell and Mankiw (1989)	1989	NBER Macroeconomics Annual 1989	Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence	chapter
16	Koening (1990)	1990	Quarterly Journal of Economics	Real Money Balances and the Timing of Consumption: An Empirical Investigation	article
17	Epstein and Zin (1991)	1991	Journal of Political Economy	Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical A	article
18	Lawrance (1991)	1991	Journal of Political Economy	Poverty and the Rate of Time Preference: Evidence from Panel Data	article
19	Campbell and Mankiw (1991)	1991	European Economic Review	The response of consumption to income: A cross-country investigation	article
20	Osano and Inoue (1991)	1991	International Economic Review	Testing between competing models of real business cycles	article
21	Runkle (1991)	1991	Journal of Monetary Economics	Liquidity constraints and the permanent-income hypothesis	article
22	Bohn (1991)	1991	Journal of Money, Credit and Banking	On Cash-in-Advance Models of Money Demand and Asset Pricing	article
23	Gleizer (1991)	1991	Brazilian Review of Econometrics	Saving and real interest rates in Brazil	article
24	Fauvel and Samson (1991)	1991	Canadian Journal of Economics	Intertemporal substitution and durable goods: an empirical analysis	article
25	Wirjanto (1991)	1991	Canadian Journal of Economics	Testing the Permanent Income Hypothesis: The Evidence from Canadian Data	article
26	Kuehlwein (1991)	1991	Economics Letters	A test for the presence of precautionary saving	article
27	Auerbach and Hassett (1991)	1991	National Saving and Economic Perform	Corporate Savings and Shareholder Consumption	chapter

Snowballing

- No query is perfect
- Idea: cover studies not identified by the query but frequently mentioned in the literature
- For each primary study, download the list of references
- Inspect the 50 studies most frequently cited by the primary studies
- If you can use some of them, add them to your list

Snowballing example

Meta-analysis of social science research: A practitioner's guide

View PDF ↗ Full text options Export

Abstract
Author keywords
Indexed keywords
SciVal Topics
Metrics
Funding details

Funding details

References (114) View in search results format >

All Export Print E-mail Save to PDF Create bibliography

1 Amini, S.M., Parmeter, C.F.
Comparison of model averaging techniques: Assessing growth determinants
(2012) *Journal of Applied Econometrics*, 27 (5), pp. 870-876. Cited 42 times.
doi: 10.1002/jae.2288
CUNI

2 Andrews, I., Kasy, &.M.
Identification of and correction for publication bias
(2019) *American Economic Review*, 109 (8), pp. 2766-2794. Cited 180 times.

AI-enhanced literature search

1. Use **Deep Research** to explore recent literature and concepts
2. Ask **ChatGPT thinking model** to refine your Scholar query
3. Download **500 abstracts** from Scholar or Publish or Perish
4. Let a **ChatGPT Agent** screen abstracts for relevant estimates
5. Ask Agent to download PDFs (or download manually)
6. Upload PDFs in **batches of 5**; ask which report usable estimates
7. (Optional) Use **ASReview** to double-check or co-train

Using thinking models to design Scholar query

- GPT thinking model helps refine search precision
- Prompt examples:
  - `I want to do a meta-analysis of the beauty premium. Suggest a Google Scholar query to find empirical studies on attractiveness and income.`
  - `Too many irrelevant hits. How can I prioritize regression-based studies only?`
- Output improves over iterations: clarify units, methods, outcomes, and keywords

Prompting ChatGPT Agent for screening

- Batch input: 20–50 abstracts at a time from Scholar export
- Example prompt:
  You are an expert in meta-analysis. For each abstract below and all other information you have about the paper, say whether the paper likely reports quantitative estimates on the effect of beauty on labor market outcomes that can be used in a meta-analysis. Justify briefly.
- Output: `Paper 1: Yes (panel data, Table 2) Paper 2: No (conceptual) Paper 3: Maybe (ambiguous)`

Screening PDFs for quantitative content

- Upload PDFs in small batches (e.g., 5 at a time)
- Example prompt:
  For each of these papers, does it contain new empirical estimates of the effect of beauty on wages that I can use in a meta-analysis? Point to specific tables or pages.
- Output format:
  - Paper A: Yes, Table 4, OLS and IV estimates on beauty premium
  - Paper B: No quantitative analysis
  - Paper C: Yes, appendix regression table

Optional enhancements and caveats

- Combine with **ASReview** for machine-assisted filtering
- Use **Semantic Scholar** for better metadata access
- Avoid over-reliance: AI helps, but expert judgment remains key
- Protect against false positives: Always validate flagged studies manually
- End-to-end AI infrastructure for meta-analysis: **otto-SR**

Minimum requirements for meta-analysis

For modern meta-analysis techniques to work, you need at least:

1. 30 estimates
2. 10 studies
3. effect sizes
4. standard errors
5. sample sizes

What to do if I have, e.g., just 10 estimates from 4 studies? Focus on the median or weighted means.

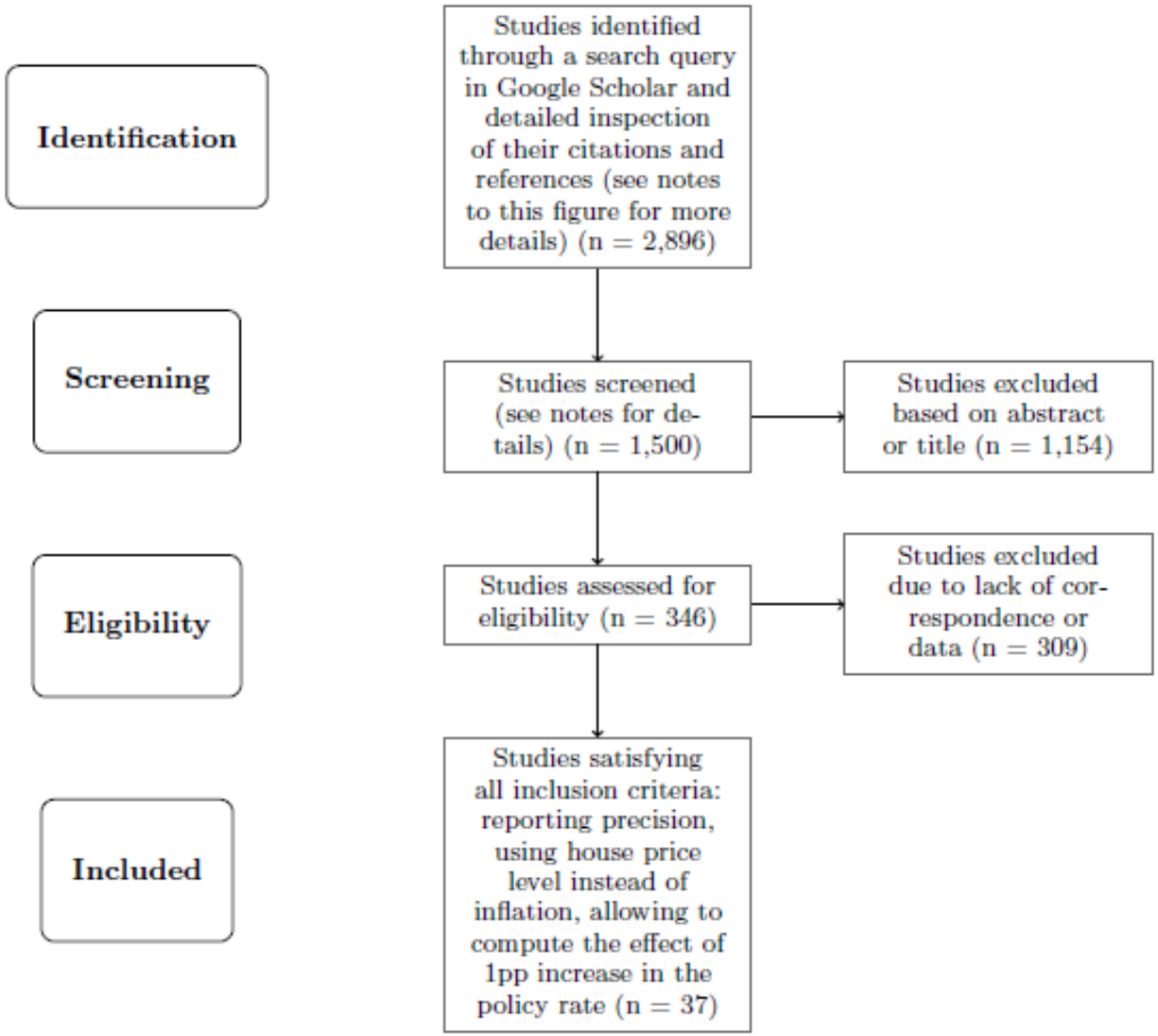
Quality

- Some studies are published in poor journals or use weak identification. What to do?
- **Option 1**: exclude these studies ex ante. Not recommended.
- **Option 2**: include these studies with controls for journal quality and methodology. You can always exclude them as a robustness check.
- Note that some fields have standardized quality benchmarks (risk of bias)

Unpublished papers

- Try to collect all studies, published and unpublished
- If unfeasible, it is possible to focus just on published studies (fewer typos, peer-review)
- Including unpublished papers unlikely to alleviate publication bias
- Still too many papers? Recruit a co-author, AI, or use a random subset (last resort)

PRISMA: document your literature search



Workshop Outline

Effects must be quantitatively comparable

What does it mean?

- When we compare estimates reported in Study 1 and Study 2, it must be clear which one is larger
- It must be clear how much larger it is
- t-statistics are not useful measures of effect size
- Regression coefficients are not comparable in general if studies use different units

Elasticities

$$\text{x-elasticity of } y: \varepsilon = \frac{\frac{\partial y}{y}}{\frac{\partial x}{x}}$$

$$\text{P-elasticity of Q: } \varepsilon = \frac{\frac{\partial Q}{Q}}{\frac{\partial P}{P}} \text{ if continuous}$$

$$\varepsilon = \frac{\frac{Q_2 - Q_1}{Q_1}}{\frac{P_2 - P_1}{P_1}} = \frac{\text{\% change in quantity Q}}{\text{\% change in price P}} \text{ if discrete}$$

Elasticities can be recovered from regressing **log(Y) on log(X)**.

Semi-elasticities, dollar values, std.dev. changes

- Semi-elasticity: effect of euro adoption on trade, effects of foreign investment on domestic productivity, effects of borders on trade
- Dollar values: statistical value of life, social cost of carbon
- Std.dev. changes: effect of class size on student test scores. What to do here?
- Class size is measured the same way in all studies (number of kids).
- Test scores are measured differently.
- Solution: By how many standard deviations do test scores improve if we decrease class size by 10 students?

Standardized mean differences

$$\theta = \frac{\mu_1 - \mu_2}{\sigma},$$

where μ_1 is the mean for one population, μ_2 is the mean for the other population, and σ is a standard deviation based on either or both populations.

- Very popular in psychology, not much used in economics. Why?
- Generally we need an experiment: treatment group, control group.
- In economics we often focus on observational research.
- For many questions the treatment and control are not binary classifications (we look at the intensity of treatment)

Partial correlation coefficients

Example: **tuition fees and university enrollment**. Ideally, we would need price elasticities of demand (PED), but many studies are not log on log.

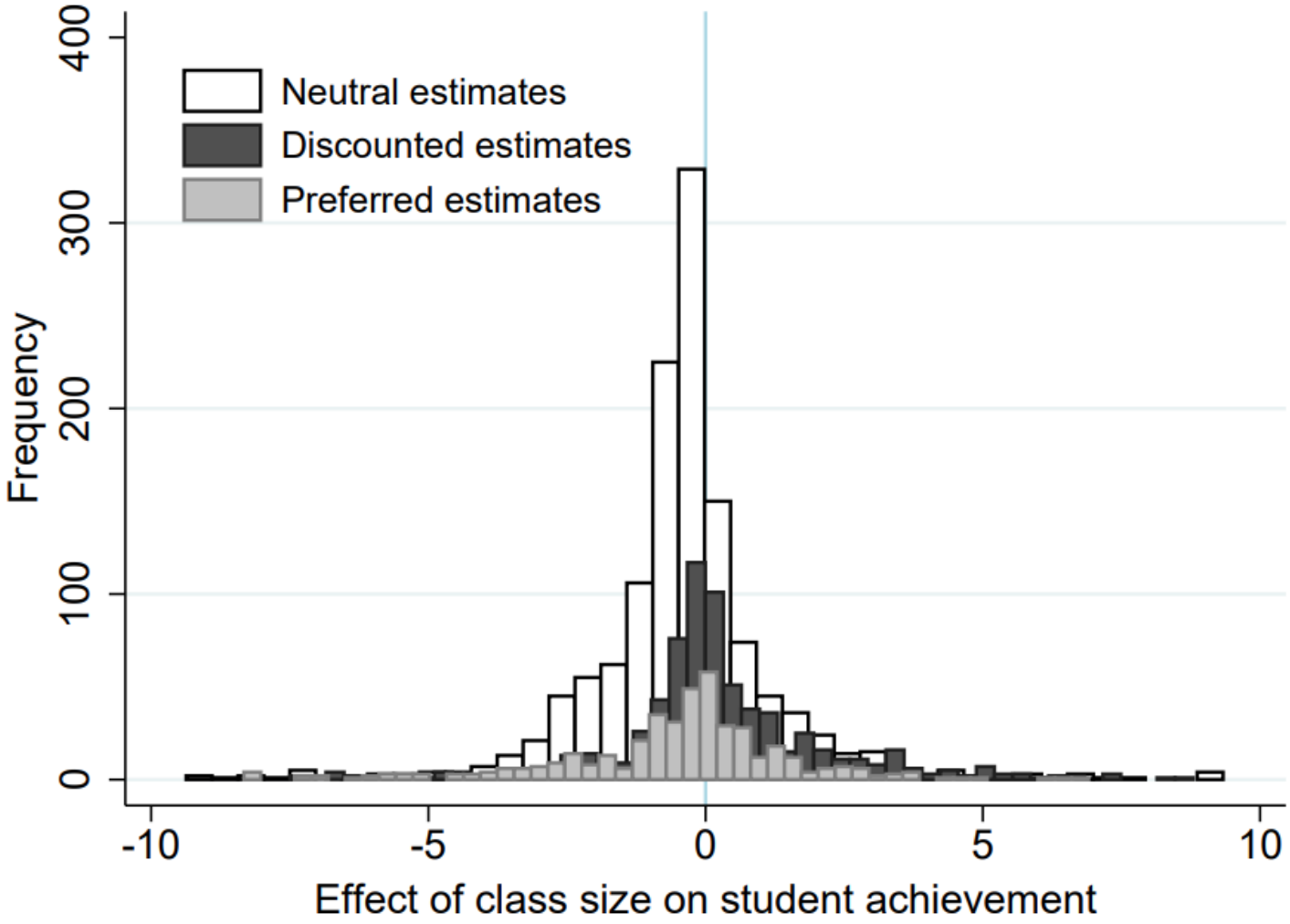
So we transform the estimates to partial correlations using t-statistics and degrees of freedom:

$$\text{PCC(PED)}_{ij} = \frac{\text{T(PED)}_{ij}}{\sqrt{\text{T(PED)}_{ij}^2 + \text{DF(PED)}_{ij}}}.$$

PCC caveats

- Last resort: can be used almost always, but we lose a lot of information
- Difficult to interpret (some guidelines provided by Doucouliagos, 2011)
- Statistical problems for meta-analysis methods (but this is the case also for SMD)
- Include robustness checks: e.g. a subsample of elasticities in the case of tuition fees and university enrollment

Collect all estimates



Standard errors

- Absolutely crucial for meta-analysis
- If not reported directly, can be computed from t-statistics or p-values
- Last resort: approximation based on within-study variation (Matousek et al. 2022)
- Interactions, transformations: delta method

Delta method

Based on Taylor series expansion. Basic formula:

$$\text{Var}(Y) = \text{Var}(f(X)) \approx [f'(\mu_X)]^2 \text{Var}(X).$$

Example: Suppose $Y = X^2$. Then $f(x) = x^2$ and $f'(x) = 2x$, so that:

$$\text{Var}(Y) \approx [2\mu_X]^2 \text{Var}(X) = 4\mu_X^2 \sigma_X^2.$$

Example: Suppose $Y = \frac{1}{X}$. Then $f(x) = \frac{1}{x}$ and $f'(x) = -\frac{1}{x^2}$, so that:

$$\text{Var}(Y) \approx \left[\frac{-1}{\mu_X^2}\right]^2 \text{Var}(X) = \frac{\sigma_X^2}{\mu_X^4}.$$

AI-assisted data collection: Overview

- As of August 2025, **human review still essential**
- AI reasonably good at
  - Identifying key study dimensions (estimation method, outcome type, etc.)
  - Extracting basic data with structured prompts and examples
  - Reducing workload to a single reviewer
- Tools: ChatGPT thinking, Agent Mode, NotebookLM
- End-to-end AI infrastructure for meta-analysis: **otto-SR**

Step-by-step data collection workflow

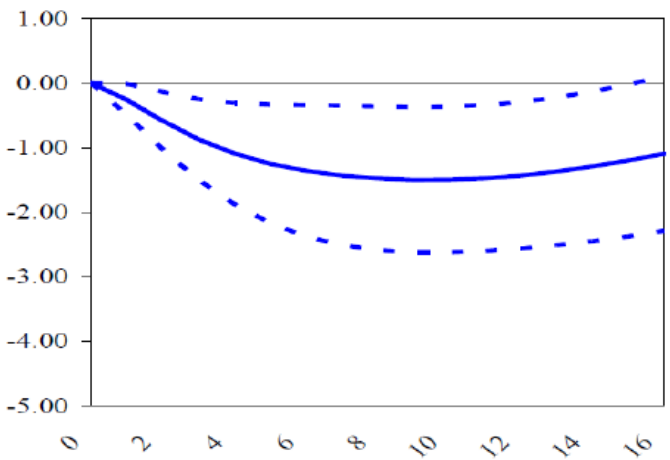
1. Use **Deep Research + thinking model** to find key coding dimensions
2. Examples: OLS vs IV, cross-section vs panel, outcome types, country, controls
3. Upload a few example PDFs + your coded spreadsheet to ChatGPT Agent
4. Prompt Agent: `"Extract these columns from new PDFs using my examples as a guide."`
5. Refine extraction prompts based on feedback
6. Use **NotebookLM** as an independent extraction check
7. Validate discrepancies manually; spot-check the rest

Example prompt for PDF data extraction

Here are 3 PDFs and the corresponding rows from my meta-analysis data sheet. Learn the structure. These data are collected well and I want you to collect data from other papers. I will send 5 new PDFs. Extract estimates of the beauty premium and the corresponding standard error. Collect information on the estimation method used, the definition of the beauty variable, definition of the earnings variable.

- Agent mode allows memory and consistency across batches
- Output can be verified, adjusted, and exported as CSV
- See summary on MAER-Net website.

Graphical estimates, measurement error

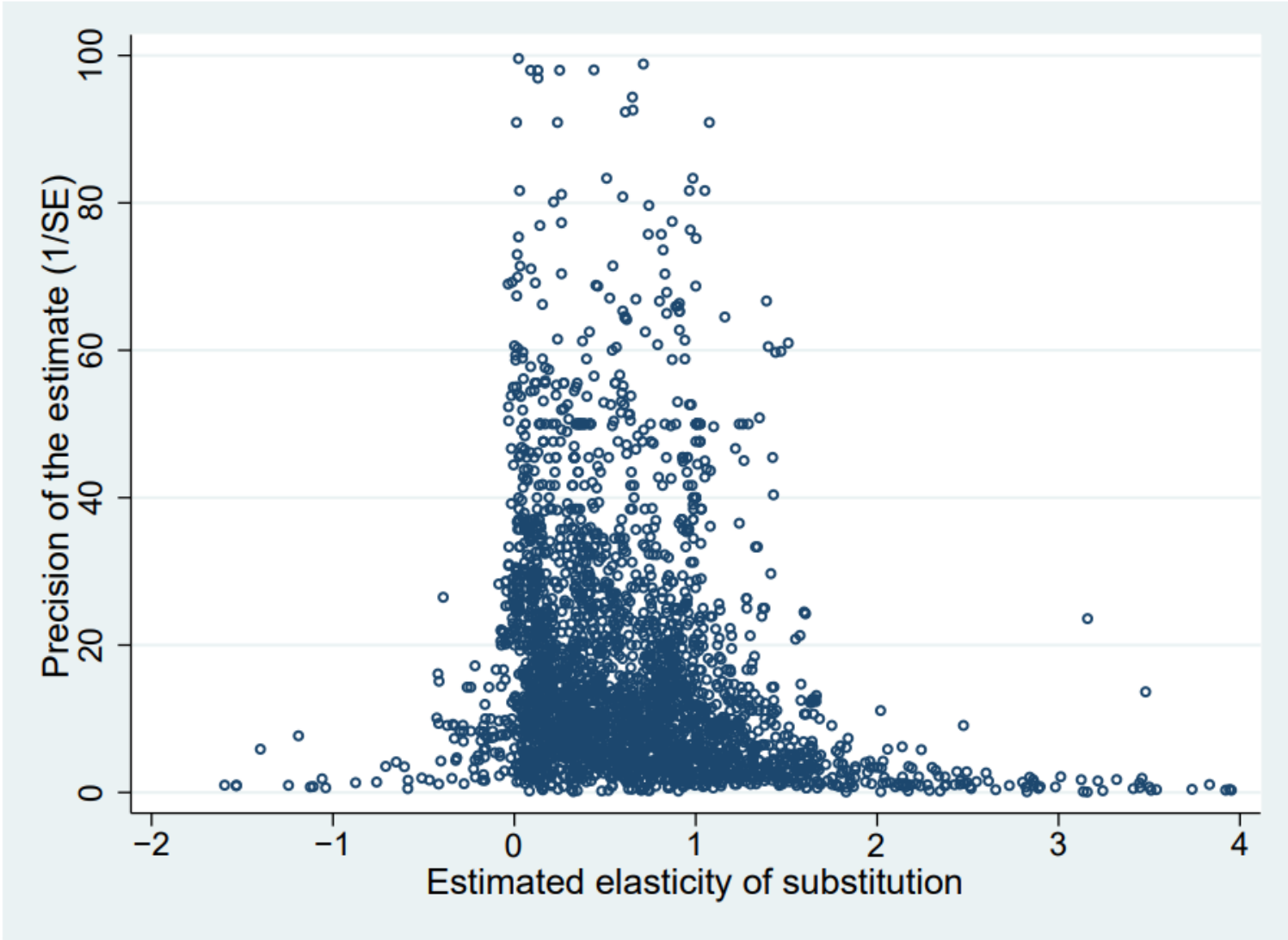


Canada

horizon	response	confidence interval
1	-0.231	(-0.407, -0.016)
2	-0.578	(-1.012, -0.135)
4	-1.076	(-1.853, -0.299)
8	-1.454	(-2.554, -0.359)
12	-1.414	(-2.534, -0.287)
16	-1.084	(-2.259, 0.131)
max	n.a.	n.a.

Obtained from: page 50, Figure 2
Frequency: Quarterly
Note: Responses are in %.

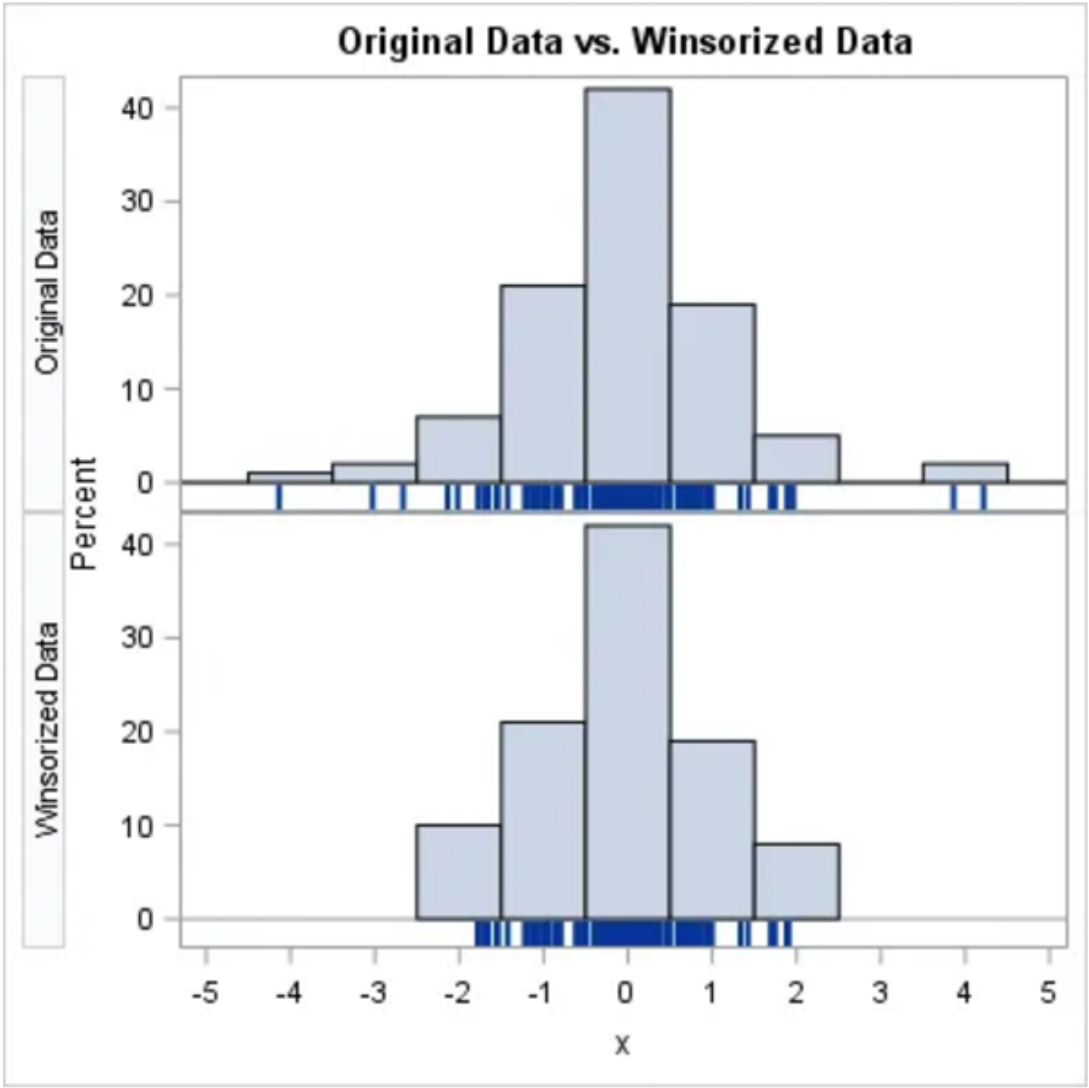
Outliers


Precision of the estimate (1/SE)
0
20
40
60
80
100
−2
−1
0
1
2
3
4
Estimated elasticity of substitution

What to do with outliers?

- No consensus
- Approach 1: omit very large effects (and standard errors), e.g. more than 3 standard deviations away from the mean
- Approach 2: run robust meta-regression or use tools for outlier detection in regression
- Approach 3: use winsorizing (my personal preference)
- Always report robustness checks

Winsorization



Hands on

Open Stata

Reflecting context: the effect of beauty on success

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	study_id	study	country	premium	se_premi	precision	tstat_pre	nobs	instrumer	weight	interview	photo_ra	software_	self_rate	dummy_t	categoric	beauty_p	number_	salary	study_ou	teaching	athletic_	electoral	other_
2	1	Ahmed e	United St	8.91	2.9601	0.3378	3.01	1806	42.497	0.1667	0	0	1	0	0	1	0	4.4427	1	0	0	0	0	
3	1	Ahmed e	United St	8.964	5.2729	0.1896	1.7	1806	42.497	0.1667	0	0	1	0	0	1	0	4.4427	1	0	0	0	0	
4	1	Ahmed e	United St	2.484	3.0293	0.3301	0.82	1806	42.497	0.1667	0	0	1	0	0	1	0	4.4427	1	0	0	0	0	
5	1	Ahmed e	United St	9.18	4.8316	0.207	1.9	1806	42.497	0.1667	0	0	1	0	0	1	0	4.4427	1	0	0	0	0	
6	1	Ahmed e	United St	21.6	6.4671	0.1546	3.34	1806	42.497	0.1667	0	0	1	0	0	1	0	4.4427	1	0	0	0	0	
7	1	Ahmed e	United St	15.012	10.075	0.0993	1.49	1806	42.497	0.1667	0	0	1	0	0	1	0	4.4427	1	0	0	0	0	
8	2	Ahn & Le	World	0.0375	1.0124	0.9878	0.037	774	27.821	0.0714	0	1	0	0	1	0	1	2.7726	0	0	0	0	1	0
9	2	Ahn & Le	World	0.6186	0.2674	3.7391	2.313	774	27.821	0.0714	0	1	0	0	1	0	1	2.7726	0	0	0	0	1	0
10	2	Ahn & Le	World	0.2825	0.1421	7.0396	1.989	774	27.821	0.0714	0	1	0	0	1	0	1	2.7726	0	0	0	0	1	0
11	2	Ahn & Le	World	0.0057	0.5684	1.7593	0.01	774	27.821	0.0714	0	1	0	0	1	0	1	2.7726	1	0	0	0	0	0
12	2	Ahn & Le	World	0.2691	1.4087	0.7099	0.191	774	27.821	0.0714	0	1	0	0	1	0	1	2.7726	1	0	0	0	0	0
13	2	Ahn & Le	World	0.7858	0.8459	1.1822	0.929	774	27.821	0.0714	0	1	0	0	1	0	1	2.7726	1	0	0	0	0	0
14	2	Ahn & Le	World	0.2622	0.4685	2.1346	0.5597	774	27.821	0.0714	0	1	0	0	0	1	0	2.7726	0	0	0	0	1	0
15	2	Ahn & Le	World	0.4506	0.3763	2.6576	1.1974	774	27.821	0.0714	0	1	0	0	0	1	0	2.7726	0	0	0	0	1	0
16	2	Ahn & Le	World	0.1326	2.3261	0.4299	0.057	774	27.821	0.0714	0	1	0	0	0	1	0	2.7726	1	0	0	0	0	0
17	2	Ahn & Le	World	1.2183	2.1218	0.4713	0.5742	774	27.821	0.0714	0	1	0	0	0	1	0	2.7726	1	0	0	0	0	0
18	2	Ahn & Le	World	0.2301	0.1175	8.5093	1.958	399	19.975	0.0714	0	1	0	0	1	0	1	2.7726	0	0	0	0	1	0
19	2	Ahn & Le	World	0.2023	0.2152	4.6471	0.94	375	19.365	0.0714	0	1	0	0	0	1	0	2.7726	0	0	0	0	1	0
20	2	Ahn & Le	World	-0.561	0.8199	1.2196	-0.684	399	19.975	0.0714	0	1	0	0	0	1	0	2.7726	1	0	0	0	0	0
21	2	Ahn & Le	World	0.3107	0.686	1.4578	0.453	375	19.365	0.0714	0	1	0	0	0	1	0	2.7726	1	0	0	0	0	0
22	3	Anyzova	Czech Re	0	7.1	0.1408	0	1031	32.109	0.25	1	1	0	1	1	0	0	0.6931	1	0	0	0	0	0
23	3	Anyzova	Czech Re	9.697	5.1087	0.1957	1.8982	1031	32.109	0.25	1	1	0	1	1	0	1	0.6931	1	0	0	0	0	0
24	3	Anyzova	Czech Re	17.468	6.9226	0.1445	2.5234	1031	32.109	0.25	1	1	0	1	1	0	0	0.6931	1	0	0	0	0	0
25	3	Anyzova	Czech Re	5.4461	5.5562	0.18	0.9802	1031	32.109	0.25	1	1	0	1	1	0	1	0.6931	1	0	0	0	0	0
26	4	Arunach.	Ecuador	12.412	2.6304	0.3802	4.7186	1960	44.272	0.0417	1	0	0	0	0	1	0	0	1	0	0	0	0	0
27	4	Arunach.	Ecuador	10.672	2.5897	0.3861	4.1209	1960	44.272	0.0417	1	0	0	0	0	1	0	0	1	0	0	0	0	0
28	4	Arunach.	Ecuador	6.4388	3.3209	0.3011	1.9389	1960	44.272	0.0417	1	0	0	0	0	1	0	0	1	0	0	0	0	0
29	4	Arunach.	Ecuador	4.7912	3.2695	0.3059	1.4654	1960	44.272	0.0417	1	0	0	0	0	1	0	0	1	0	0	0	0	0
30	4	Arunach.	Mexico	28.018	5.1207	0.1953	5.4715	923	30.381	0.0417	1	0	0	0	0	1	0	0	1	0	0	0	0	0
31	4	Arunach.	Mexico	16.416	5.8208	0.1718	2.8202	923	30.381	0.0417	1	0	0	0	0	1	0	0	1	0	0	0	0	0
32	4	Arunach.	Mexico	16.416	5.8208	0.1718	2.8202	923	30.381	0.0417	1	0	0	0	0	1	0	0	1	0	0	0	0	0
33	4	Arunach.	Mexico	11.016	6.6609	0.1501	1.6538	923	30.381	0.0417	1	0	0	0	0	1	0	0	1	0	0	0	0	0
34	4	Arunach.	Ecuador	10.885	3.0023	0.3331	3.6254	1960	44.272	0.0417	1	0	0	0	1	0	0	0	1	0	0	0	0	0

Measurement

Measurement of beauty

- Interviewer-rated beauty
- Photo-rated beauty
- Software-rated beauty
- Self-rated beauty
- Dummy beauty
- Categorical beauty
- Beauty premium
- Beauty penalty

Measurement of success

- Earnings
- Study outcomes
- Teaching & research outcomes
- Athletic success
- Electoral success
- Other outcomes

Data, estimation, publication

Data characteristics
- Male subjects
- Female subjects
- Mixed gender
- High-skilled workers
- Prostitutes
- Other dressy occupations
- Non-dressy occupations
- Western culture
- Other cultures
- Panel data
- Cross-section

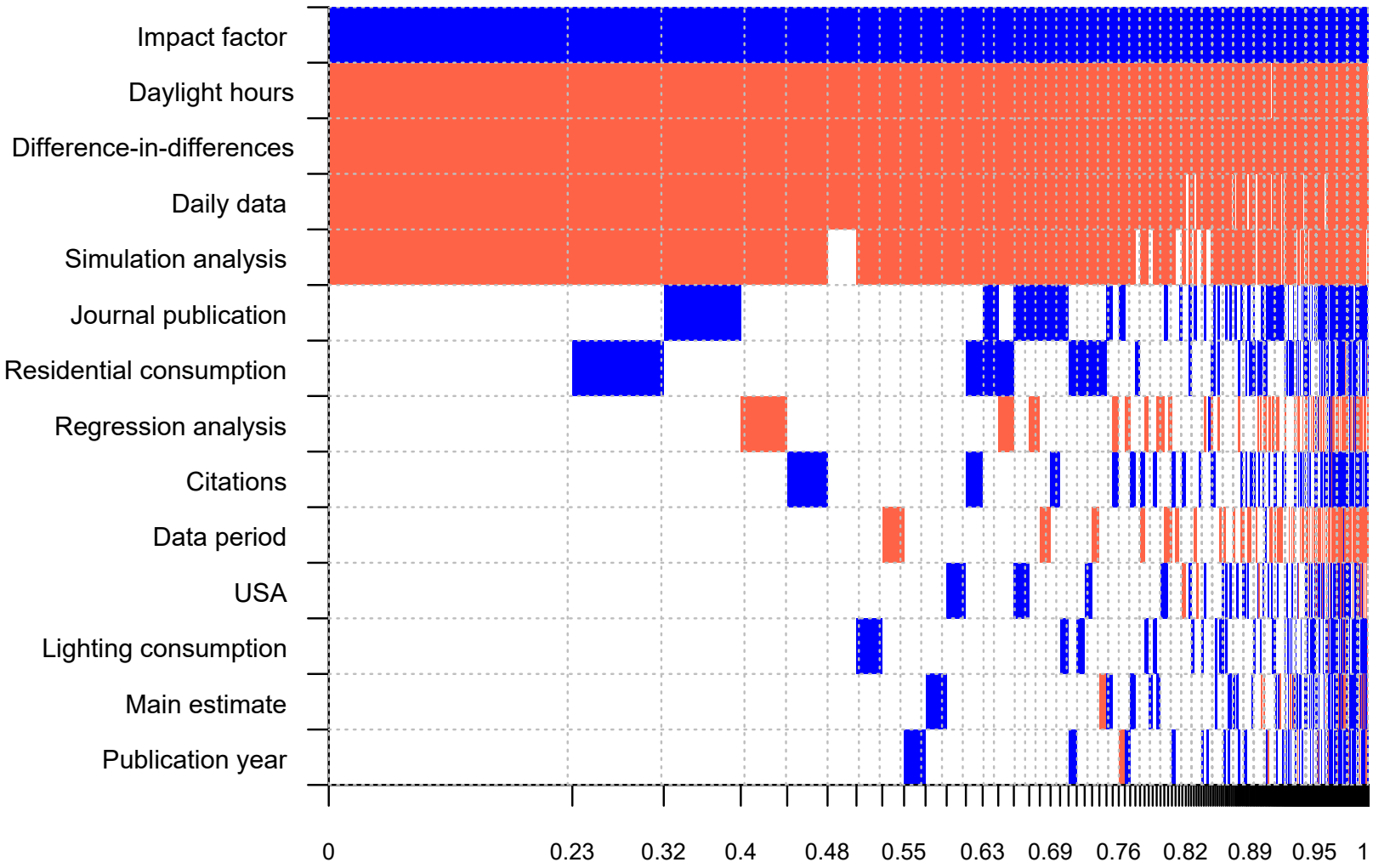
Estimation technique
- Ordinary least squares
- Instrumental variables
- Difference-in-differences
- Other method
- Cognitive skill control
- No cognitive skill control

Publication characteristics
- Published study
- Unpublished study
- High-quality peer review

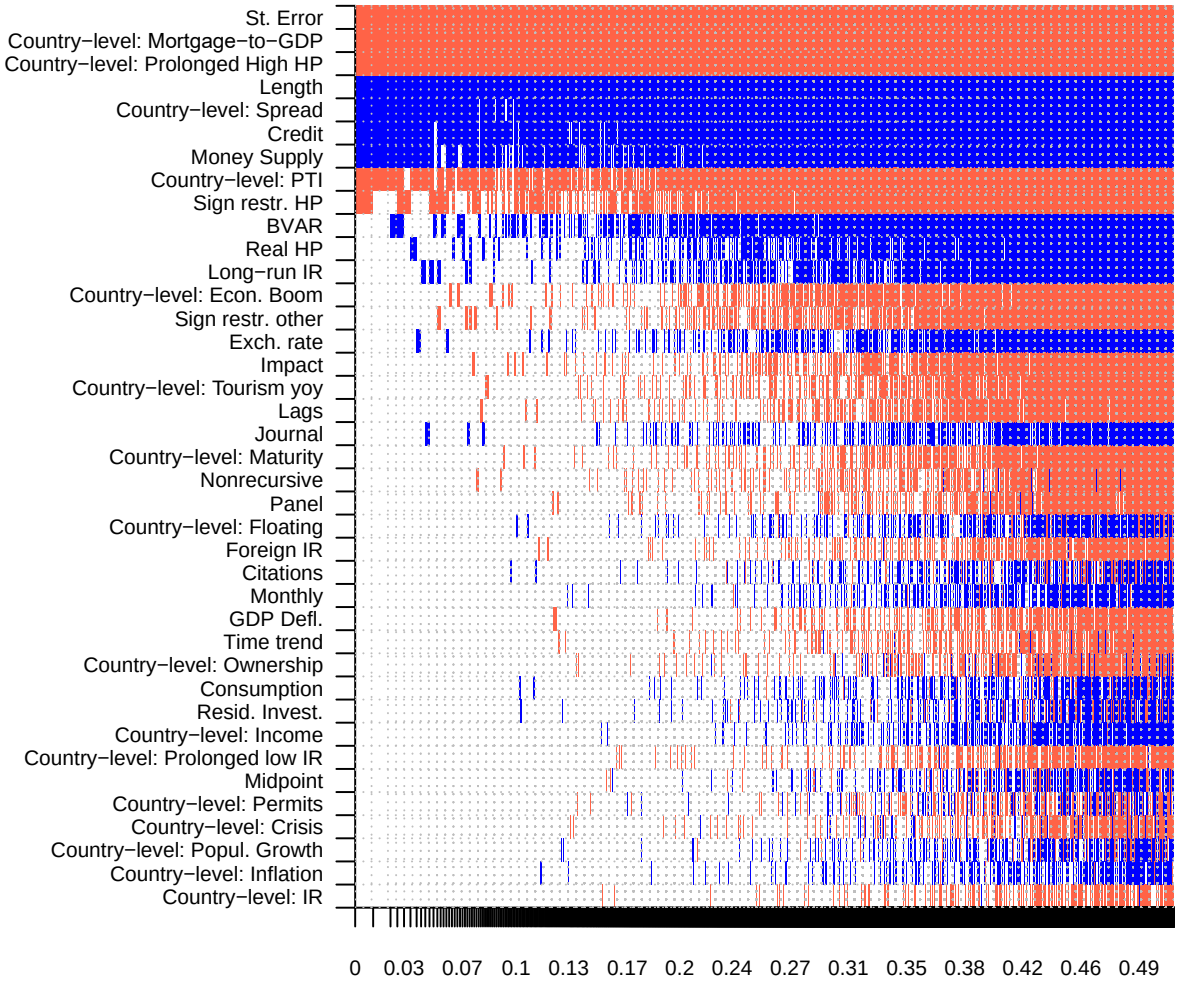
Value added

- Meta-analysis can be used to test hypotheses that primary studies cannot investigate
- Example: the effect of daylight saving time on energy consumption
- Individual studies have data for individual countries (or provinces)
- In meta-analysis, we have data for dozens of countries
- We can investigate the sources of cross-country differences

DST savings and latitude



House prices and monetary policy

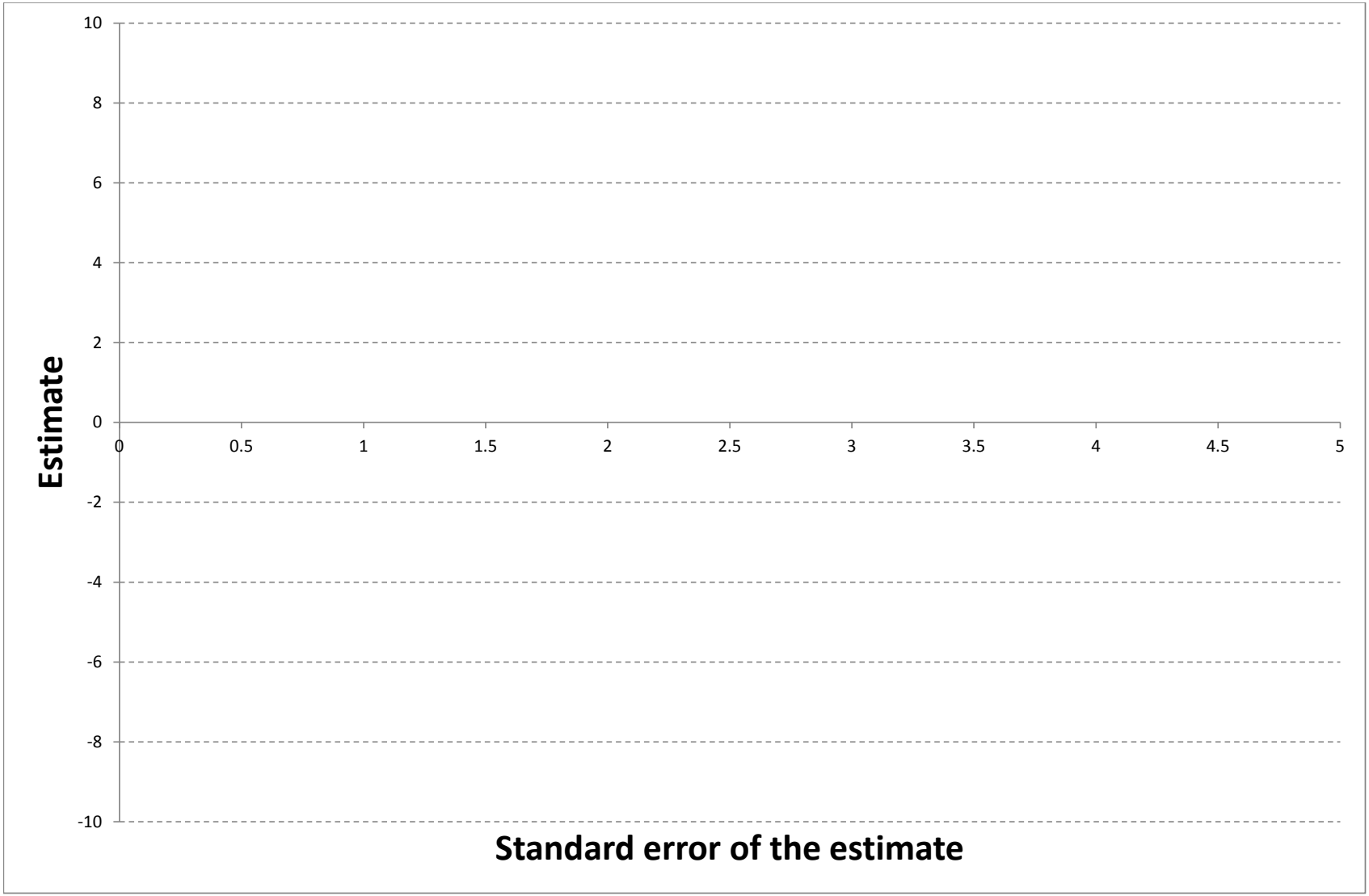


Workshop Outline

Summary statistics for meta-analysis

- Simple mean of reported estimates: natural starting point
- But inefficient and affected by outliers, publication bias, p-hacking
- Median: surprisingly robust to most problems in meta-analysis
- If you have only a few studies or estimates, focus on the (unweighted) median!
- Classical meta-analysis approach: use inverse variance as the weight

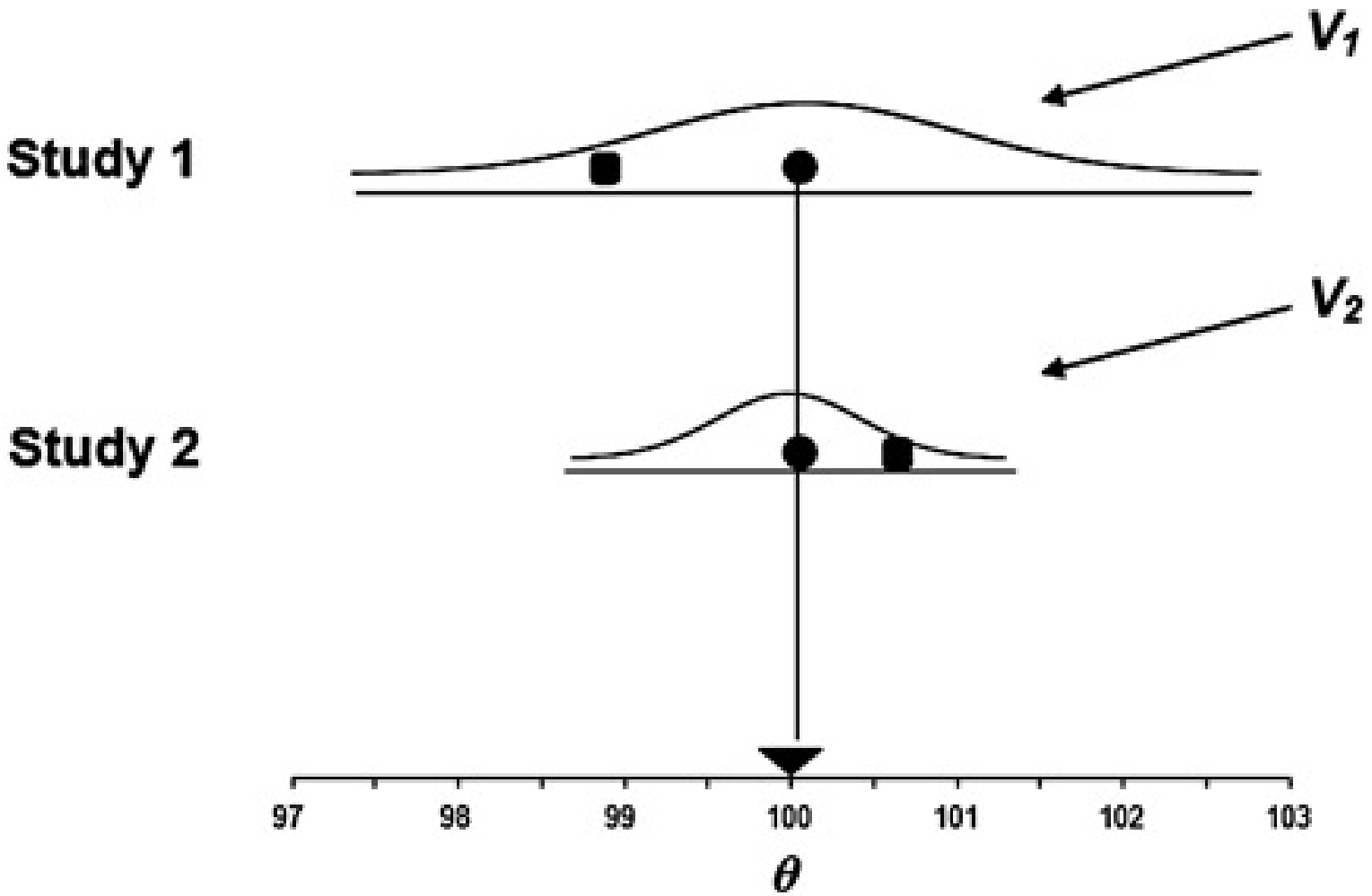
More precision -> more weight



More precision -> more weight

More precision -> more weight

“Fixed effect” estimator in meta-analysis



“Fixed effect” estimator in meta-analysis

Mean weighted by inverse variance of individual estimates.
Very sensitive to outliers in precision (including typos).

$$Y_i = \theta + \varepsilon_i, \tag{1}$$

$$V_i = \frac{\sigma^2}{n}, \tag{2}$$

$$W_i = \frac{1}{V_i}, \tag{3}$$

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}, \tag{4}$$

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i}. \tag{5}$$

Hands on

Open Stata

FE vs. UWLS: point estimate

- Unrestricted weighted least squares yield the same point estimate:

$$\hat{\theta} = \frac{\sum w_i \theta_i}{\sum w_i}, \qquad w_i = \frac{1}{SE_i^2}$$

- Difference: Variance estimation.

Method	**Point Estimate**
Fixed-Effect (FE)	$\hat{\theta} = \frac{\sum w_i \theta_i}{\sum w_i}$
UWLS	$\hat{\theta} = \frac{\sum w_i \theta_i}{\sum w_i}$

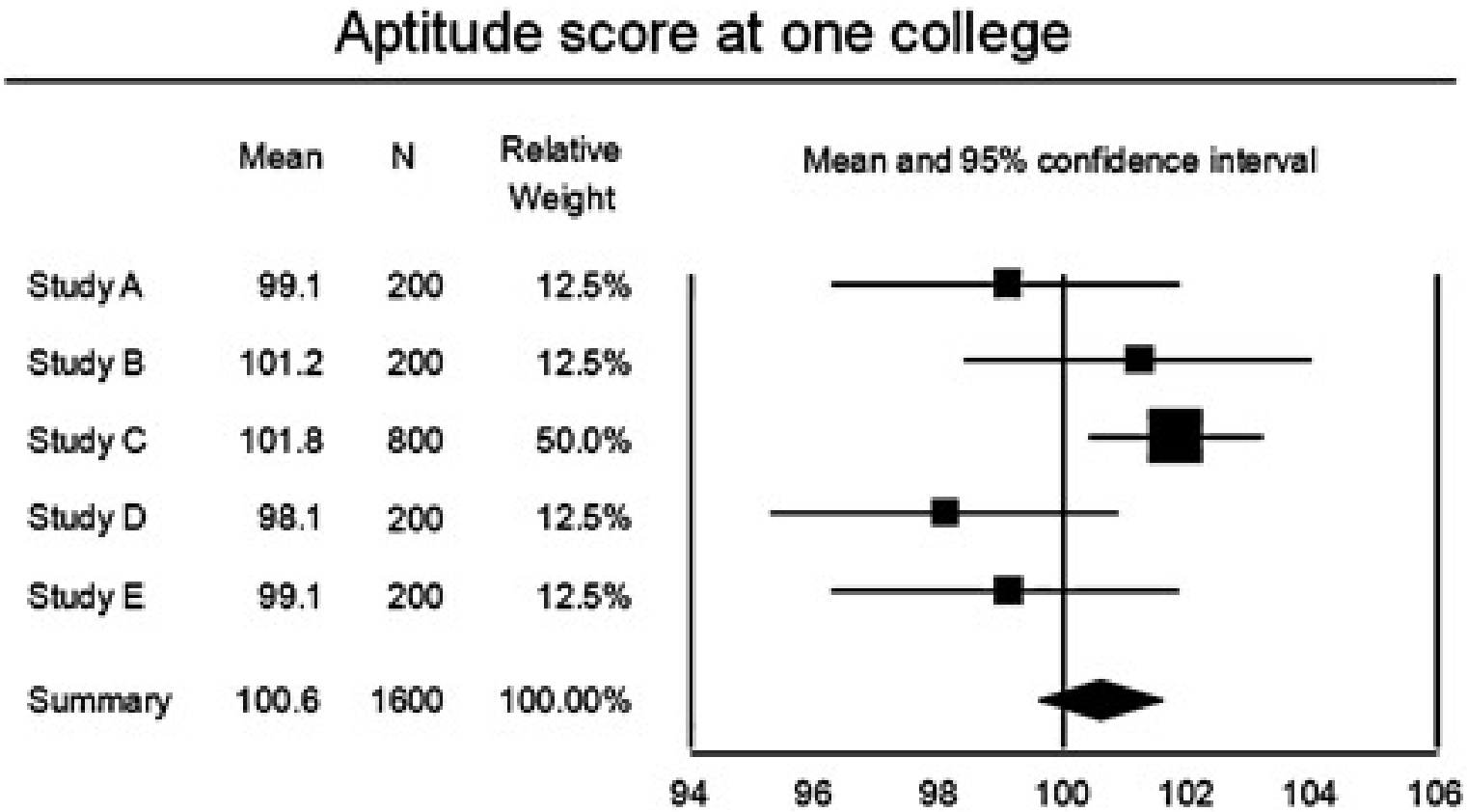
Same estimate, different variance!

FE vs. UWLS: variance

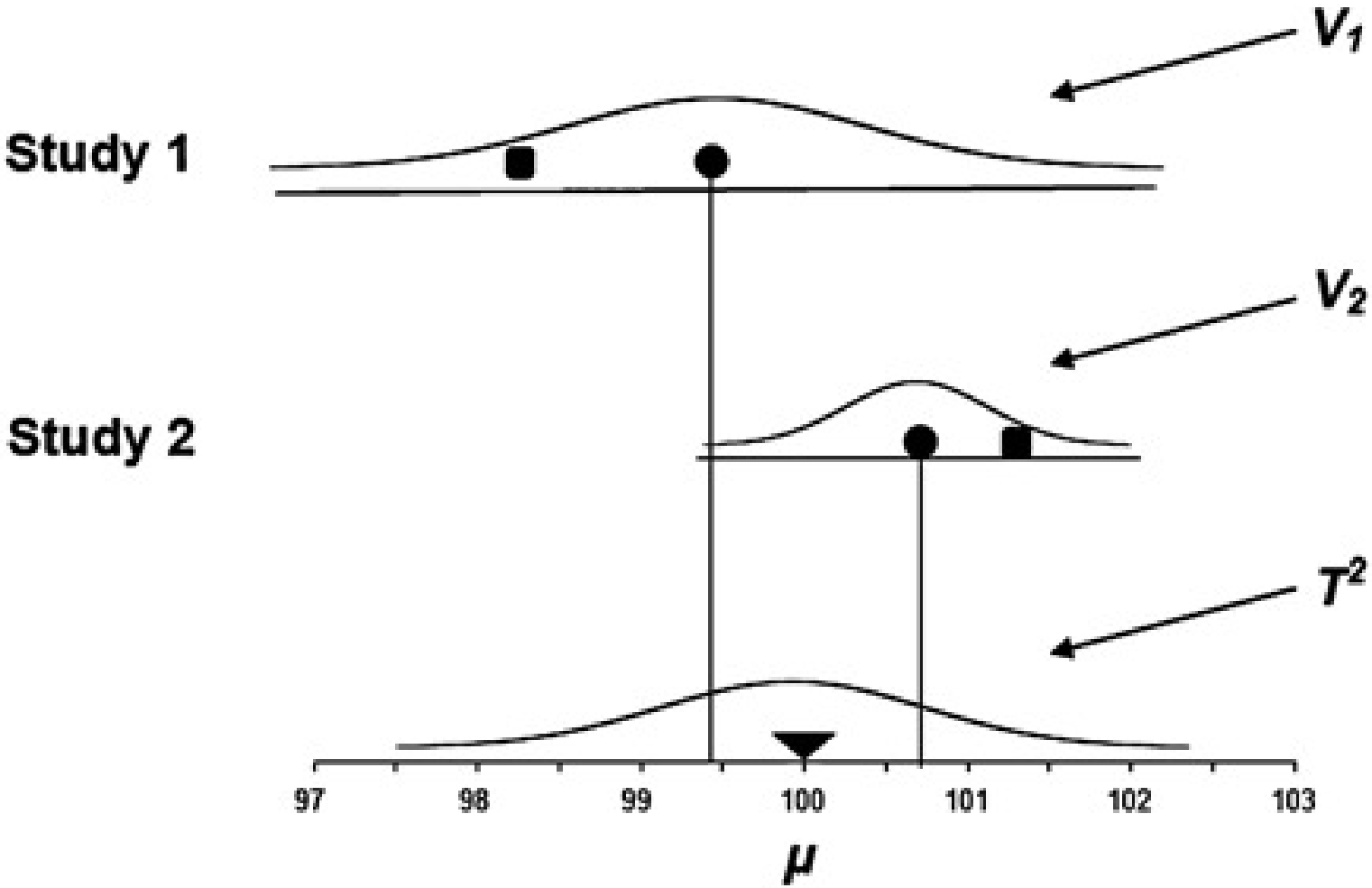
Feature	**Fixed-Effect (FE)**	**UWLS**
Assumption	Identical θ	Allows heterogeneity
Variance formula	$\frac{1}{\sum w_i}$	$\frac{\sum w_i(\theta_i - \hat{\theta})^2}{\sum w_i}$
Variance components	Within-study only	Within + between-study
Robustness	Sensitive to misspecification	More robust

FE typically underestimates variance. Use UWLS!

Fixed effect



Random effects



Random effects

The weight is now diluted by a heterogeneity term. More robust to outliers in precision and heterogeneity, less robust to publication bias.

$$Y_i = \mu + \xi_i + \varepsilon_i, \tag{1}$$

$$V_i = \frac{\sigma^2}{n}, \tag{2}$$

$$W_i = \frac{1}{V_i + \tau^2}, \tag{3}$$

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}, \tag{4}$$

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i}. \tag{5}$$

Estimating τ^2 in random-effects model

Between-study variance τ^2 accounts for heterogeneity across studies. Common estimators:

- **DerSimonian-Laird (DL)**:

$$\tau^2_{DL} = \frac{Q - (k-1)}{C}, \quad C = \sum w_i - \frac{\sum w_i^2}{\sum w_i}$$

  - Simple, but underestimates τ^2 when k is small.

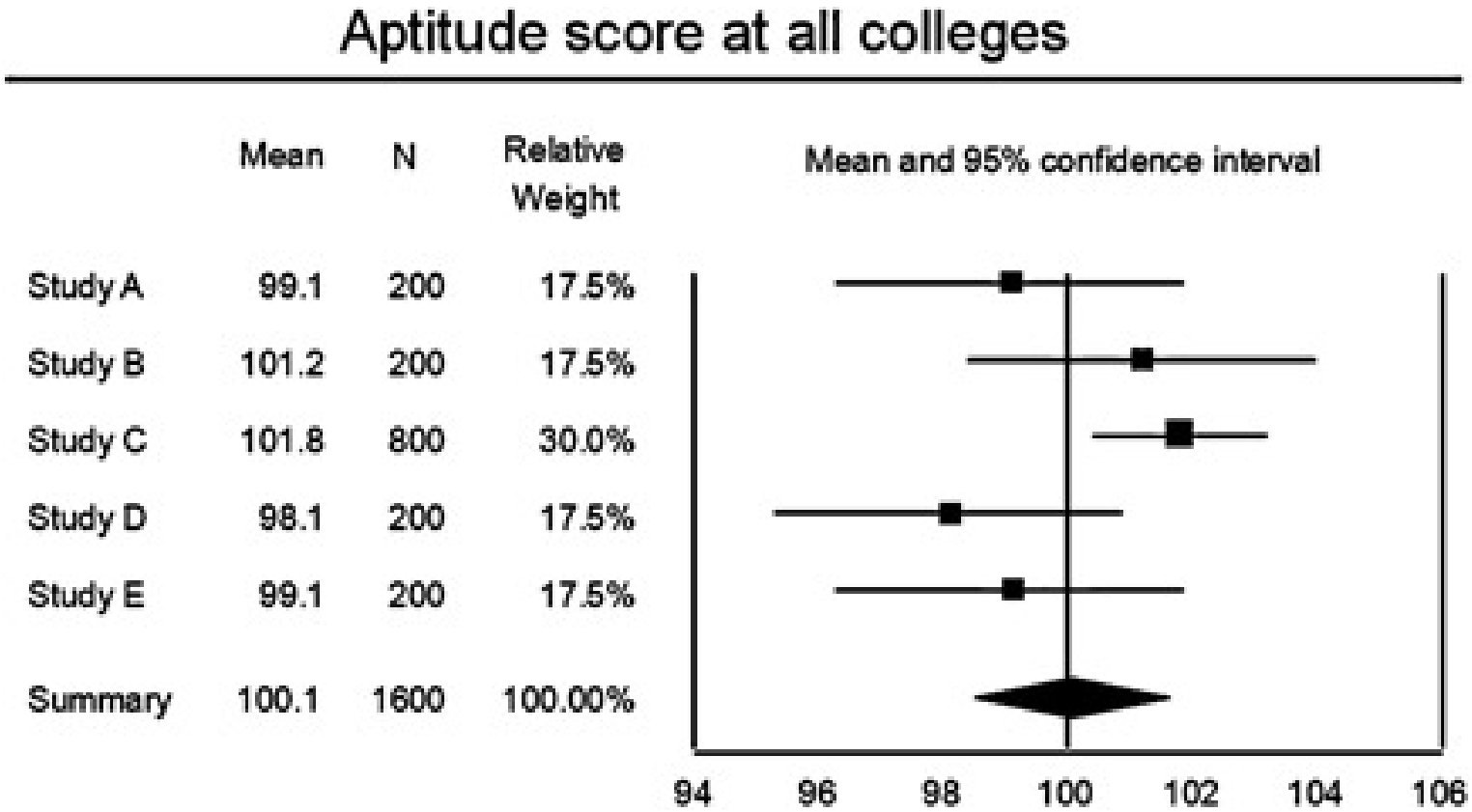
- **Restricted Maximum Likelihood (REML)**:

$$\max_{\tau^2} \sum \log(w_i^* + \tau^2) + \sum \frac{(y_i - \hat{\theta})^2}{w_i^* + \tau^2}$$

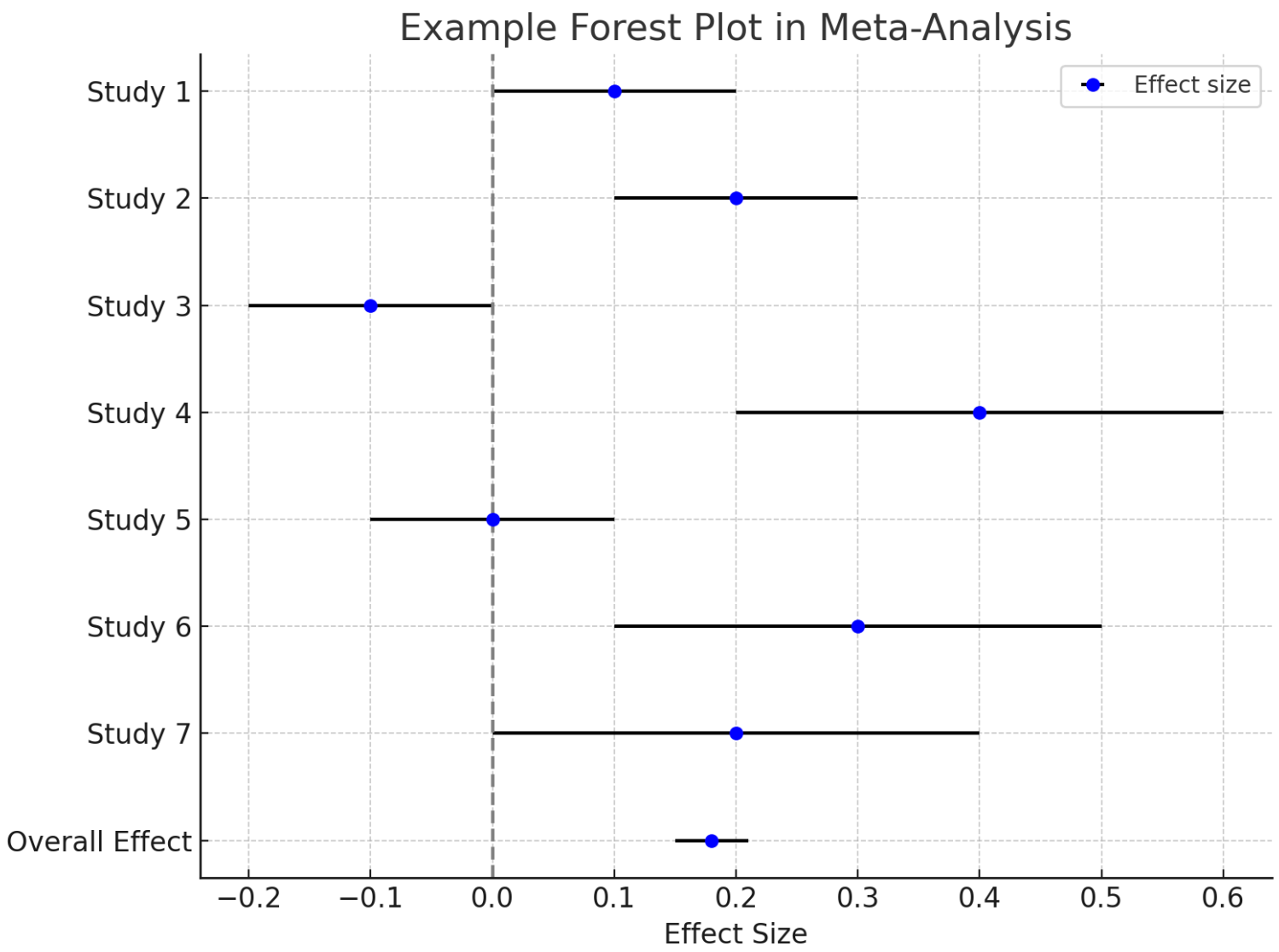
  - More accurate, default in software.

Rule of thumb: Use REML unless simplicity is needed!

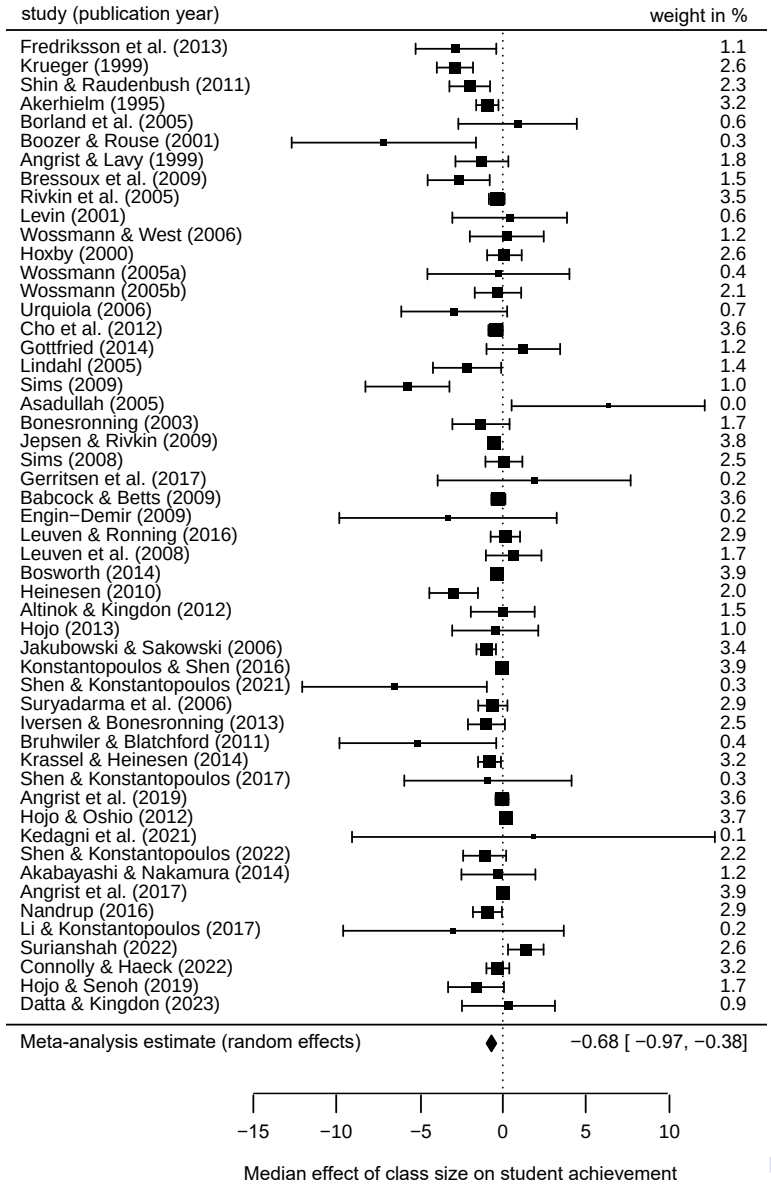
Random effects



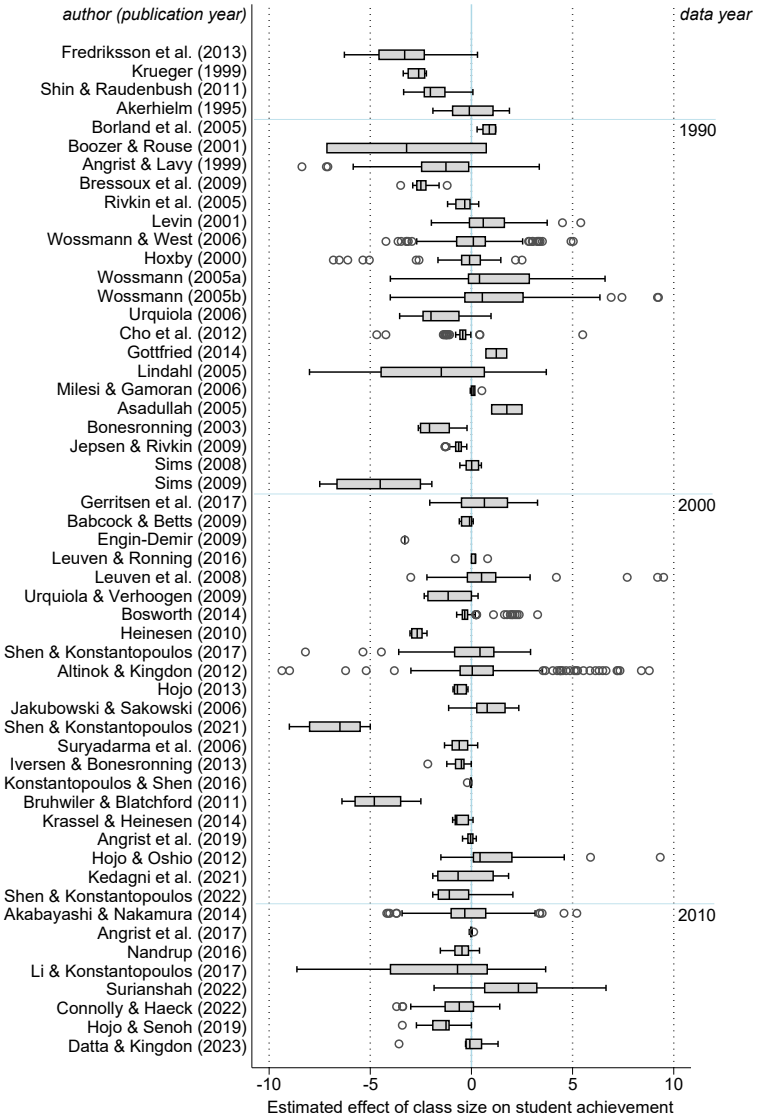
Forest plot


Example Forest Plot in Meta-Analysis
Effect size
Study 1
Study 2
Study 3
Study 4
Study 5
Study 6
Study 7
Overall Effect
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Effect Size

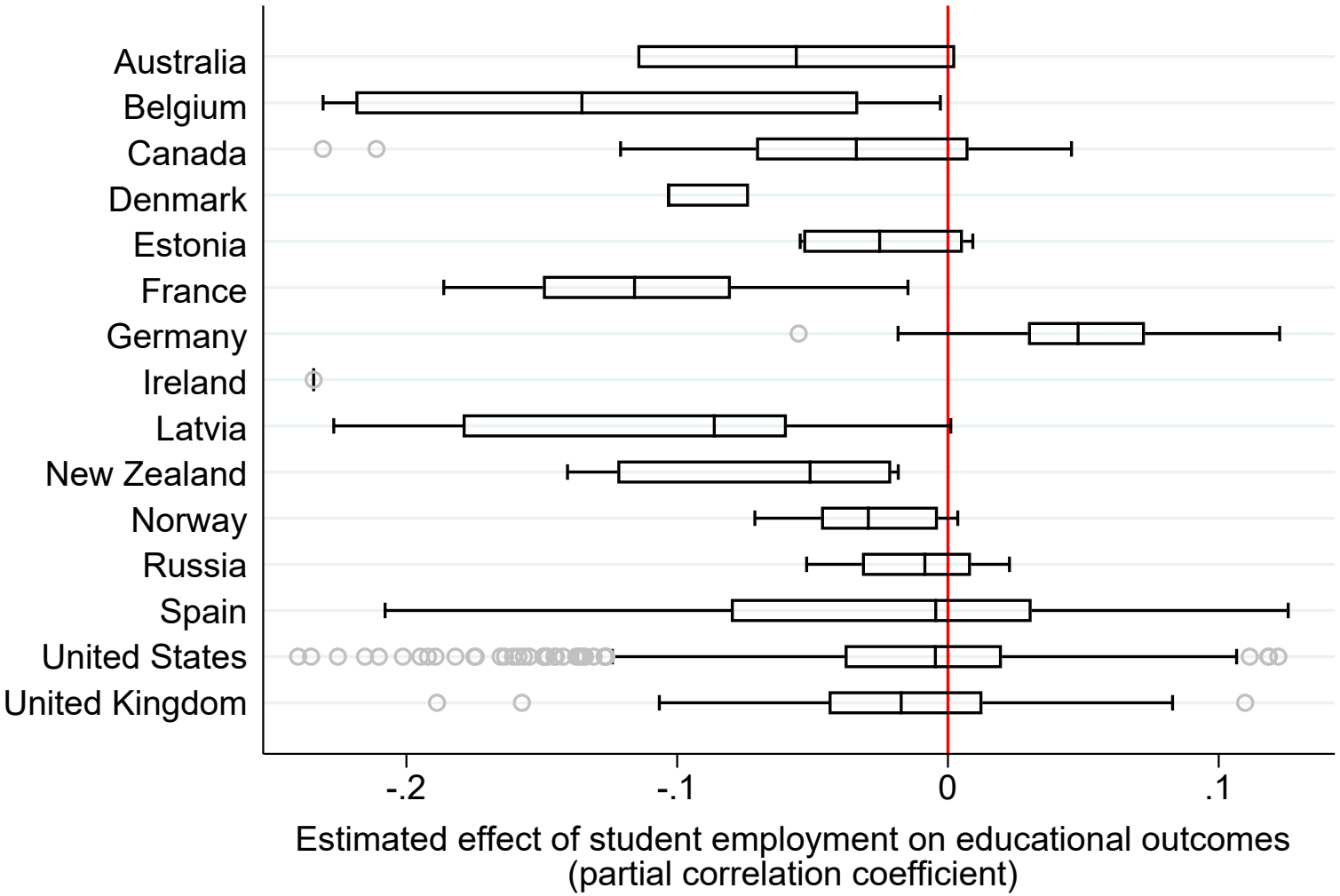
Forest plot


study (publication year)
weight in %
Fredriksson et al. (2013) 1.1
Krueger (1999) 2.6
Shin & Raudenbush (2011) 2.3
Akerhielm (1995) 3.2
Borland et al. (2005) 0.6
Boozer & Rouse (2001) 0.3
Angrist & Lavy (1999) 1.8
Bressoux et al. (2009) 1.5
Rivkin et al. (2005) 3.5
Levin (2001) 0.6
Wossmann & West (2006) 1.2
Hoxby (2000) 2.6
Wossmann (2005a) 0.4
Wossmann (2005b) 2.1
Urquiola (2006) 0.7
Cho et al. (2012) 3.6
Gottfried (2014) 1.2
Lindahl (2005) 1.4
Sims (2009) 1.0
Asadullah (2005) 0.0
Bonesronning (2003) 1.7
Jepsen & Rivkin (2009) 3.8
Sims (2008) 2.5
Gerritsen et al. (2017) 0.2
Babcock & Betts (2009) 3.6
Engin−Demir (2009) 0.2
Leuven & Ronning (2016) 2.9
Leuven et al. (2008) 1.7
Bosworth (2014) 3.9
Heinesen (2010) 2.0
Altinok & Kingdon (2012) 1.5
Hojo (2013) 1.0
Jakubowski & Sakowski (2006) 3.4
Konstantopoulos & Shen (2016) 3.9
Shen & Konstantopoulos (2021) 0.3
Suryadarma et al. (2006) 2.9
Iversen & Bonesronning (2013) 2.5
Bruhwiler & Blatchford (2011) 0.4
Krassel & Heinesen (2014) 3.2
Shen & Konstantopoulos (2017) 0.3
Angrist et al. (2019) 3.6
Hojo & Oshio (2012) 3.7
Kedagni et al. (2021) 0.1
Shen & Konstantopoulos (2022) 2.2
Akabayashi & Nakamura (2014) 1.2
Angrist et al. (2017) 3.9
Nandrup (2016) 2.9
Li & Konstantopoulos (2017) 0.2
Surianshah (2022) 2.6
Connolly & Haeck (2022) 3.2
Hojo & Senoh (2019) 1.7
Datta & Kingdon (2023) 0.9
Meta-analysis estimate (random effects)
−0.68 [−0.97, −0.38]
−15
−10
−5
0
5
10
Median effect of class size on student achievement

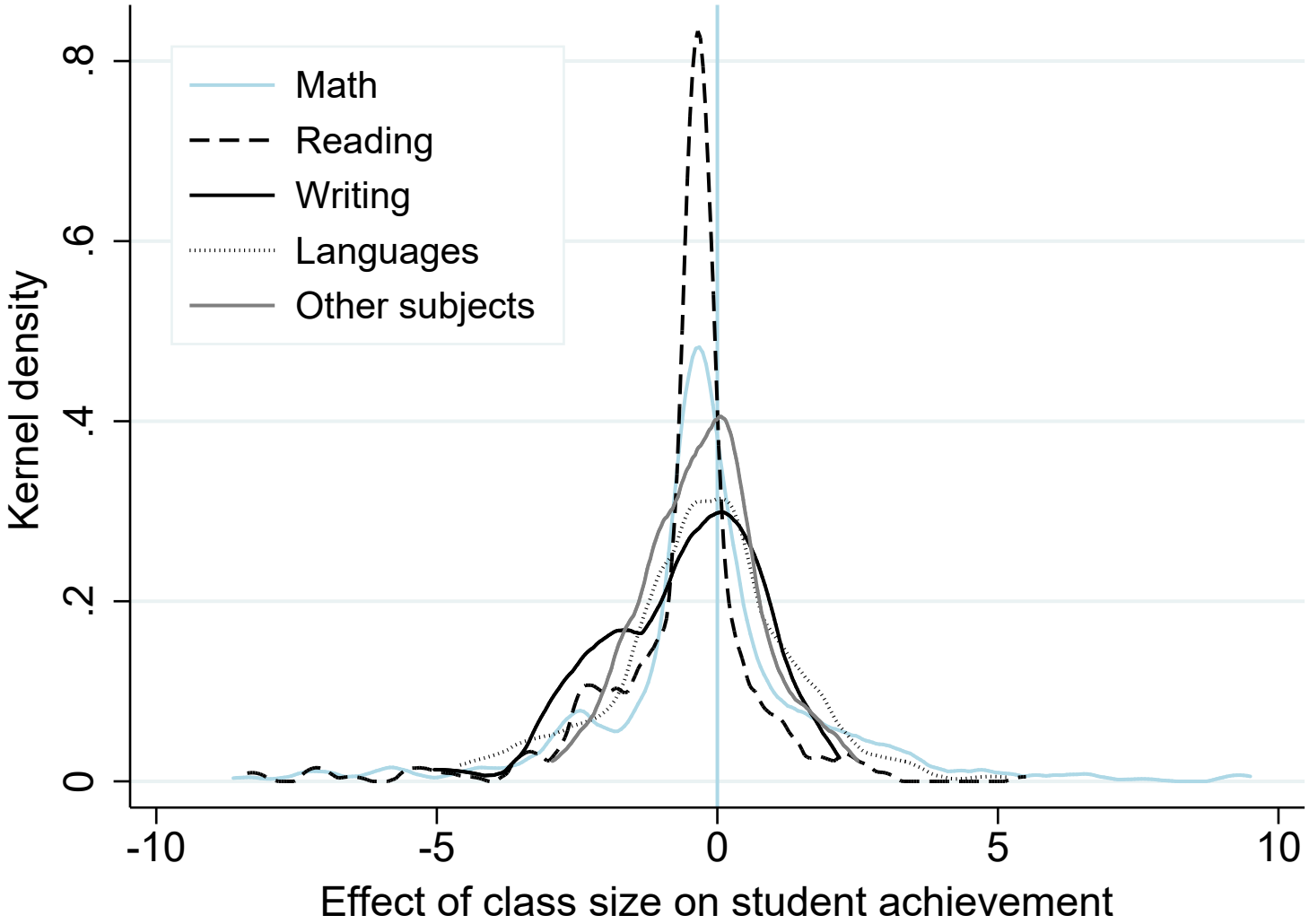
Box plot


author (publication year)
data year
Fredriksson et al. (2013)
Krueger (1999)
Shin & Raudenbush (2011)
Akerhielm (1995)
Borland et al. (2005)
Boozer & Rouse (2001)
Angrist & Lavy (1999)
Bressoux et al. (2009)
Rivkin et al. (2005)
Levin (2001)
Wossmann & West (2006)
Hoxby (2000)
Wossmann (2005a)
Wossmann (2005b)
Urquiola (2006)
Cho et al. (2012)
Gottfried (2014)
Lindahl (2005)
Milesi & Gamoran (2006)
Asadullah (2005)
Bonesronning (2003)
Jepsen & Rivkin (2009)
Sims (2008)
Sims (2009)
Gerritsen et al. (2017)
Babcock & Betts (2009)
Engin-Demir (2009)
Leuven & Ronning (2016)
Leuven et al. (2008)
Urquiola & Verhoogen (2009)
Bosworth (2014)
Heinesen (2010)
Shen & Konstantopoulos (2017)
Altinok & Kingdon (2012)
Hojo (2013)
Jakubowski & Sakowski (2006)
Shen & Konstantopoulos (2021)
Suryadarma et al. (2006)
Iversen & Bonesronning (2013)
Konstantopoulos & Shen (2016)
Bruhwiler & Blatchford (2011)
Krassel & Heinesen (2014)
Angrist et al. (2019)
Hojo & Oshio (2012)
Kedagni et al. (2021)
Shen & Konstantopoulos (2022)
Akabayashi & Nakamura (2014)
Angrist et al. (2017)
Nandrup (2016)
Li & Konstantopoulos (2017)
Surianshah (2022)
Connolly & Haeck (2022)
Hojo & Senoh (2019)
Datta & Kingdon (2023)
1990
2000
2010
-10
-5
0
5
10
Estimated effect of class size on student achievement

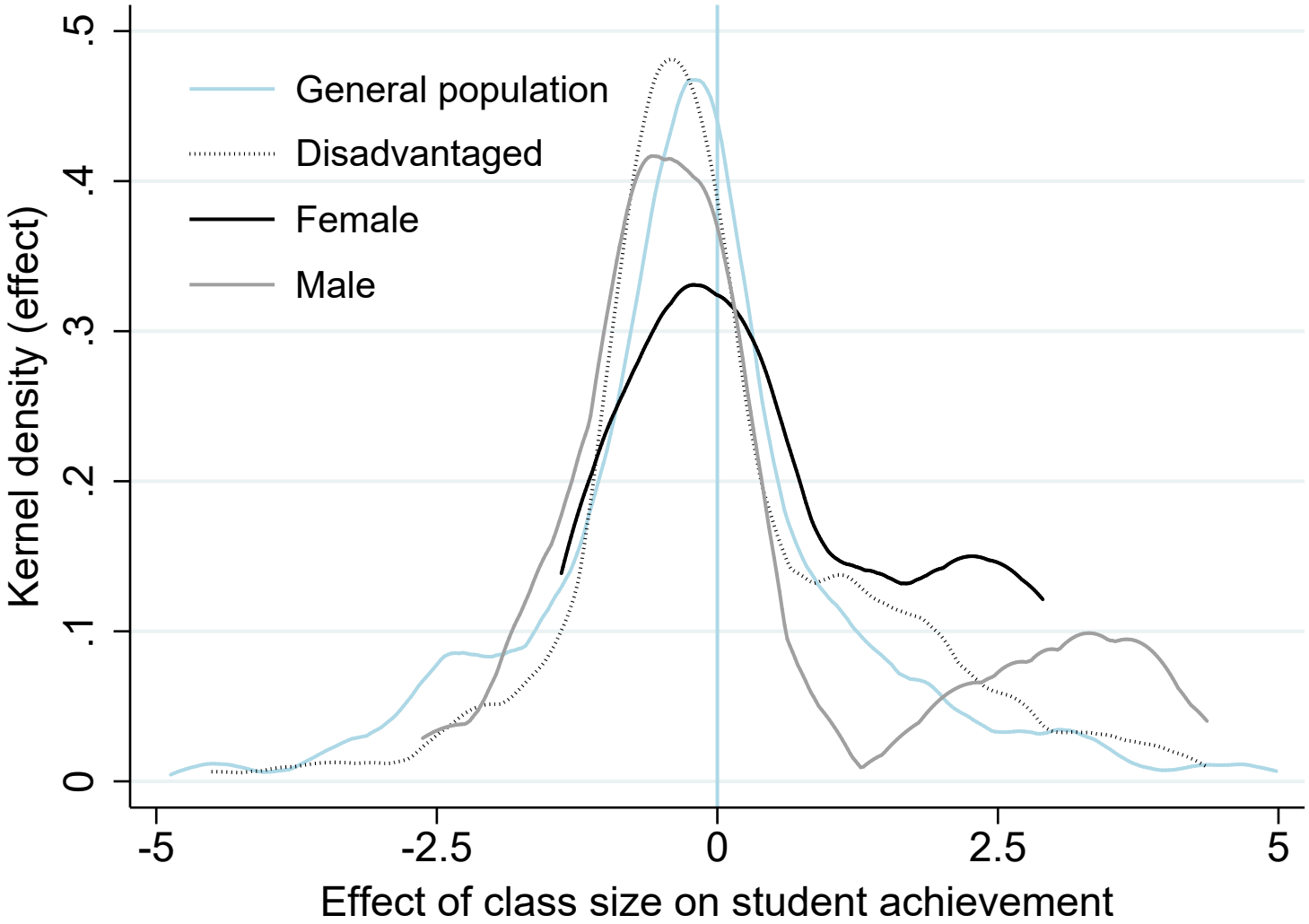
Box plot


Australia
Belgium
Canada
Denmark
Estonia
France
Germany
Ireland
Latvia
New Zealand
Norway
Russia
Spain
United States
United Kingdom
-.2
-.1
0
.1
Estimated effect of student employment on educational outcomes
(partial correlation coefficient)

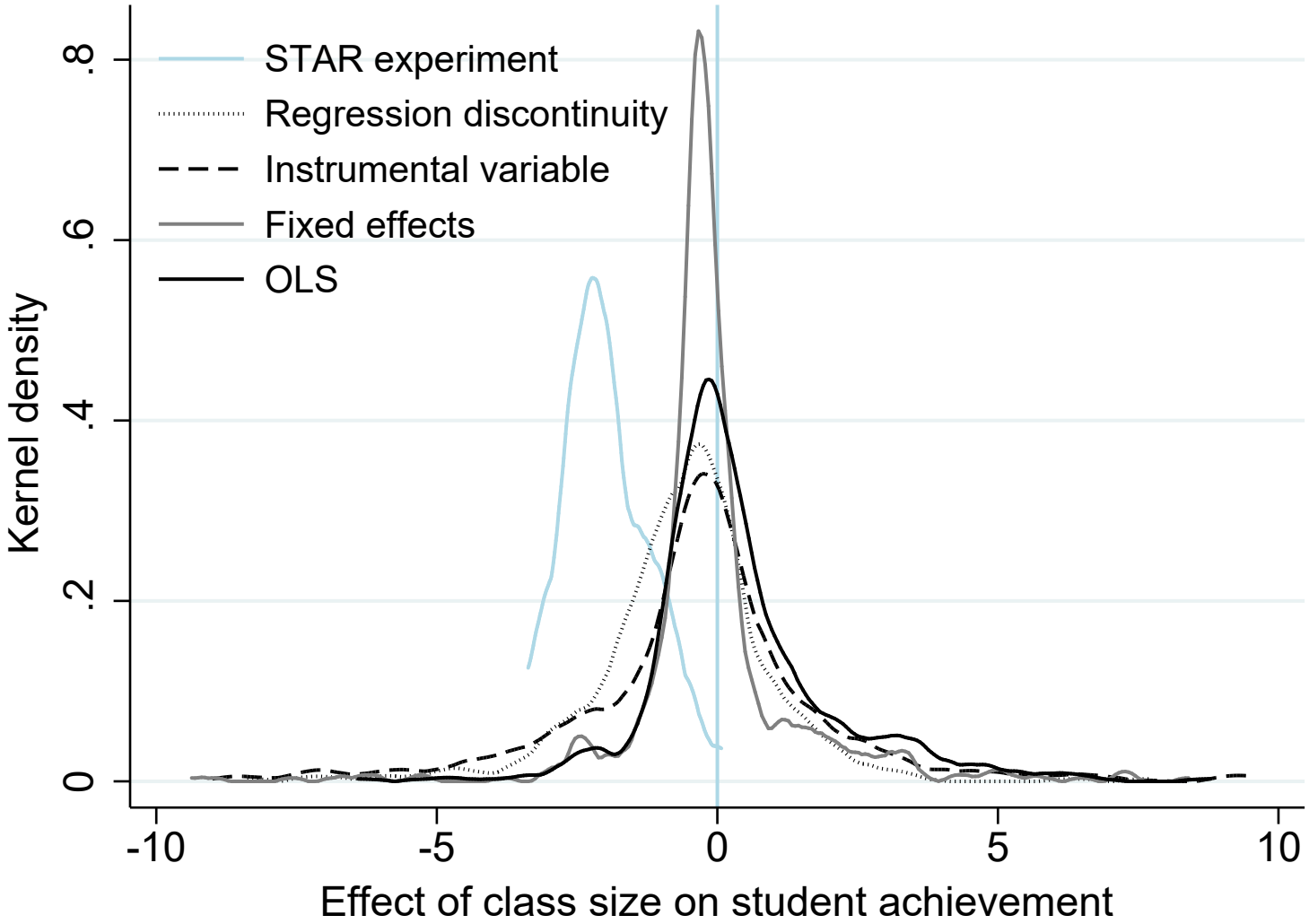
Histogram


Math
Reading
Writing
Languages
Other subjects
Kernel density
0
.2
.4
.6
.8
-10
-5
0
5
10
Effect of class size on student achievement

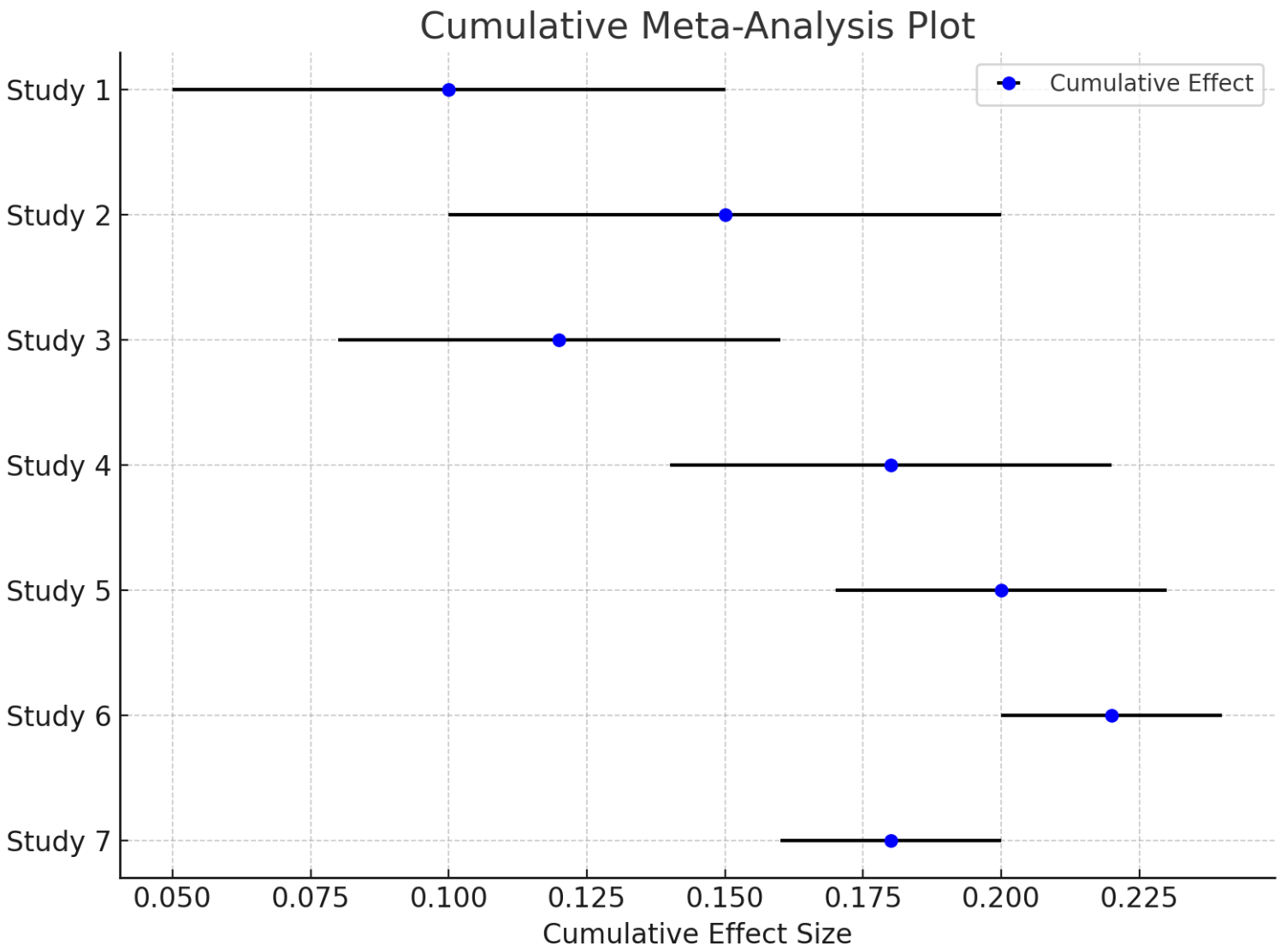
Histogram


General population
Disadvantaged
Female
Male
Kernel density (effect)
0
.1
.2
.3
.4
.5
-5
-2.5
0
2.5
5
Effect of class size on student achievement

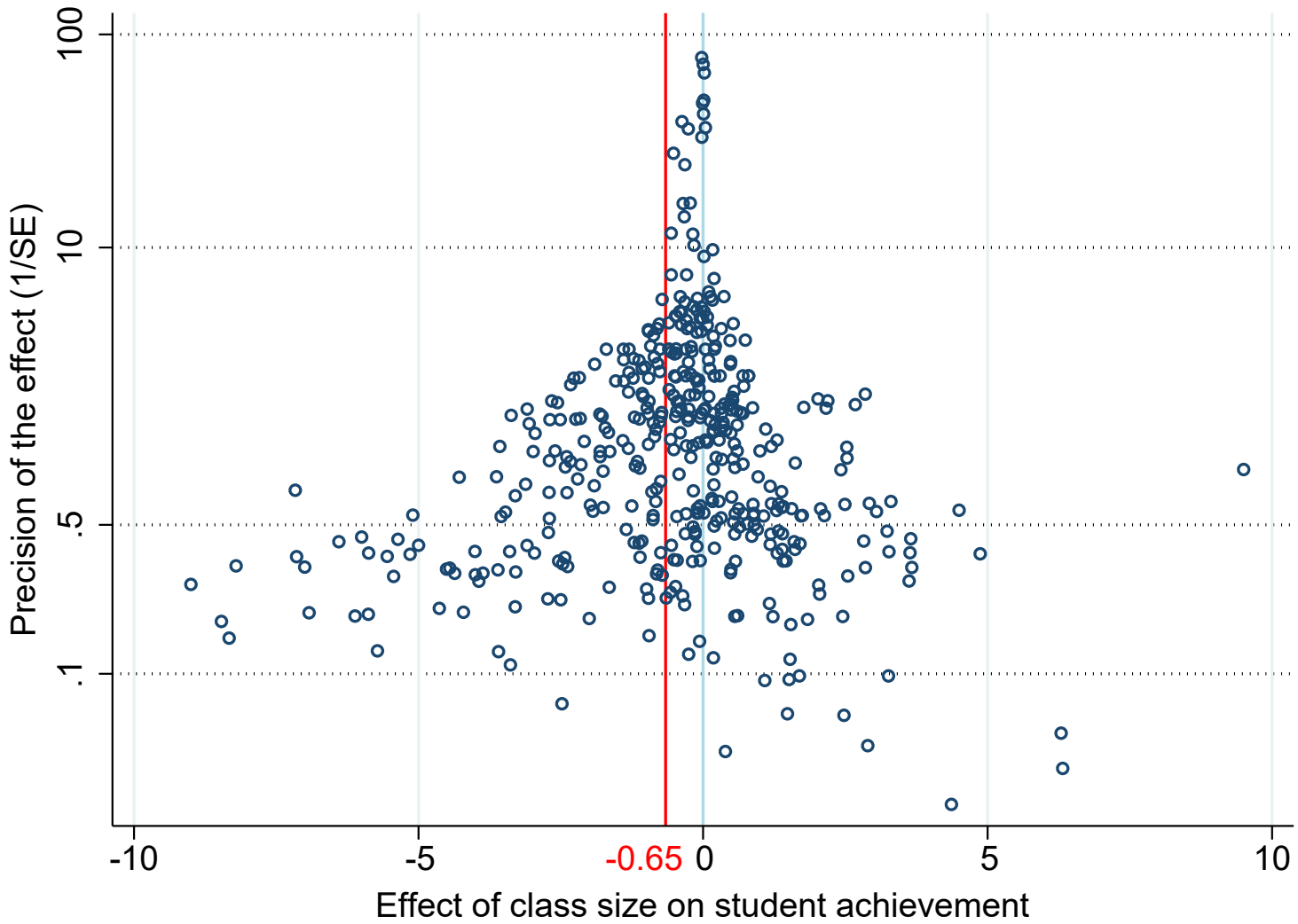
Histogram


STAR experiment
Regression discontinuity
Instrumental variable
Fixed effects
OLS
Kernel density
0
.2
.4
.6
.8
-10
-5
0
5
10
Effect of class size on student achievement

Cumulative meta-analysis plot



Funnel plot (topic of the next session)



Workshop Outline

Paxil scandal

- **Background:**
  - Paxil (paroxetine): Antidepressant approved for adolescents, developed by GlaxoSmithKline (GSK).
  - Study 329 (1998-2001): Clinical trial funded by GSK to test Paxil's safety and efficacy in adolescents.

- **What Happened?**
  - Published study claimed Paxil was *"well-tolerated and effective."*
  - Internal documents later revealed:
    - Negative results (e.g., increased suicidal ideation) were downplayed.
    - Positive outcomes were selectively reported.

Paxil scandal

- **Discovery of Bias:**
  - Legal proceedings in 2012 forced disclosure of internal data.
  - Independent reanalysis of Study 329 showed:
    - No significant benefit over placebo.
    - Serious safety concerns (e.g., suicidal ideation).

- **Key Lessons:**
  - Publication bias can harm patients and public trust.
  - Highlights the need for:
    - Transparency in research data.
    - Independent replication and reanalysis.

No bias, no E-SE correlation (?)

Researchers estimate

$$Earnings_j = \gamma Beauty_j + \ldots + w_j.$$

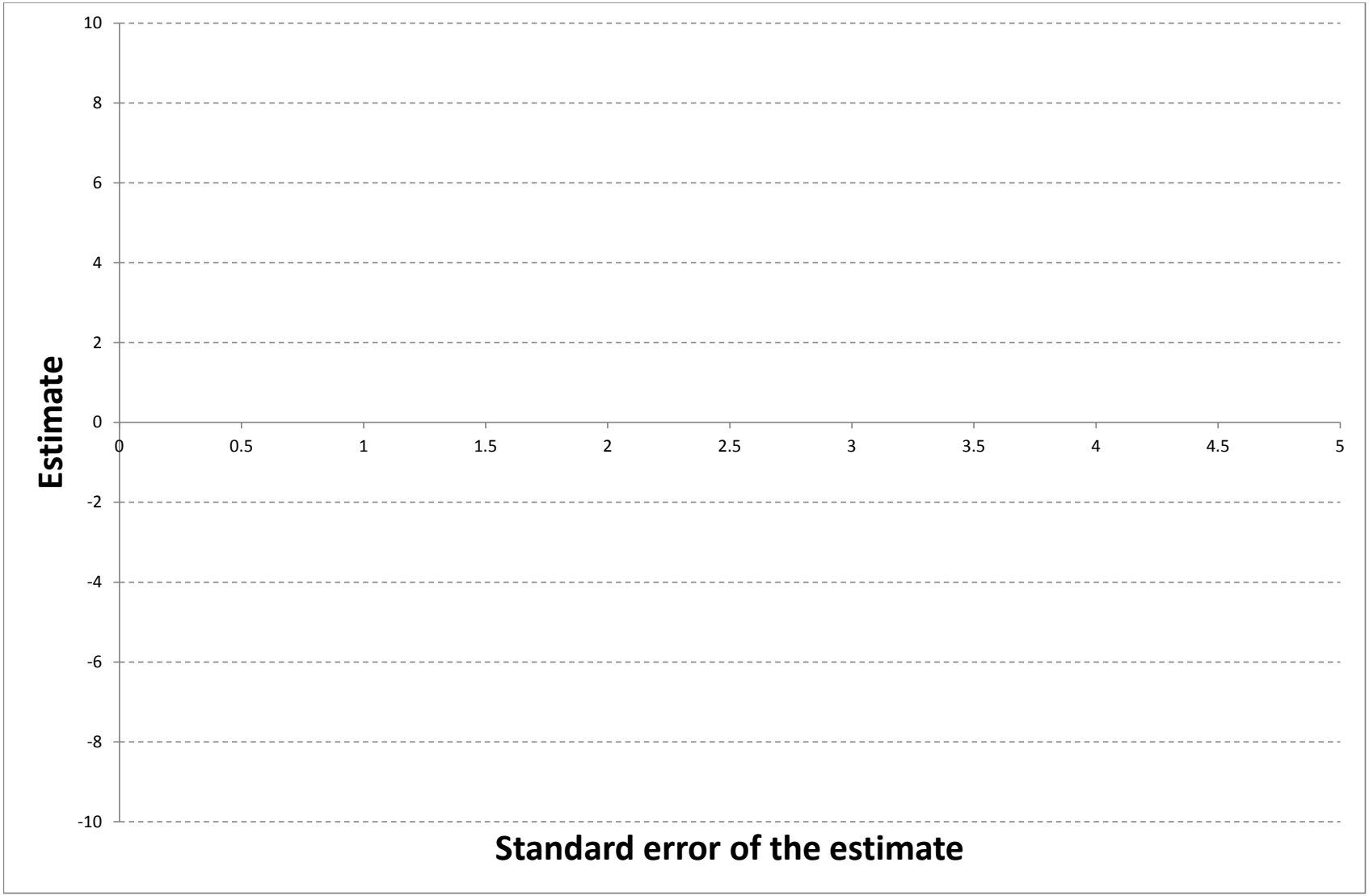
They assume that $\widehat{\gamma}/SE_{\widehat{\gamma}}$ has a t-distribution.

$\rightarrow$ estimates should not be correlated with standard errors
(Card & Krueger, 1995)

Types of publication bias

Type I: Selection for the "correct sign."
Type II: Selection for statistical significance.

Simulation


10
8
6
4
2
0
-2
-4
-6
-8
-10
Estimate
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Standard error of the estimate

Simulation


Estimate
Standard error of the estimate
10
8
6
4
2
0
-2
-4
-6
-8
-10
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Simulation


10
8
6
4
2
0
-2
-4
-6
-8
-10
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Estimate
Standard error of the estimate

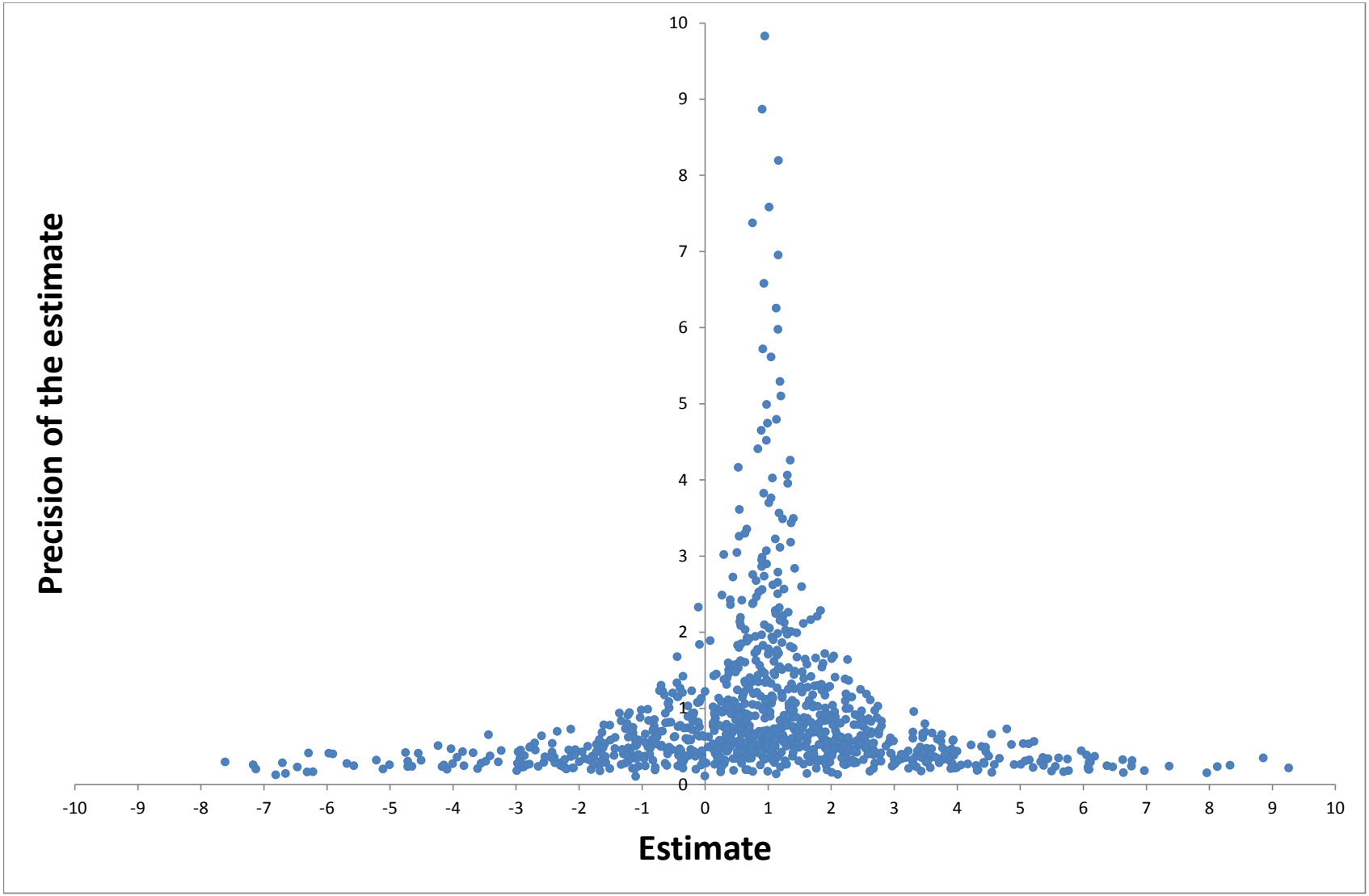
Simulation


10
8
6
4
2
0
-2
-4
-6
-8
-10
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Estimate
Standard error of the estimate

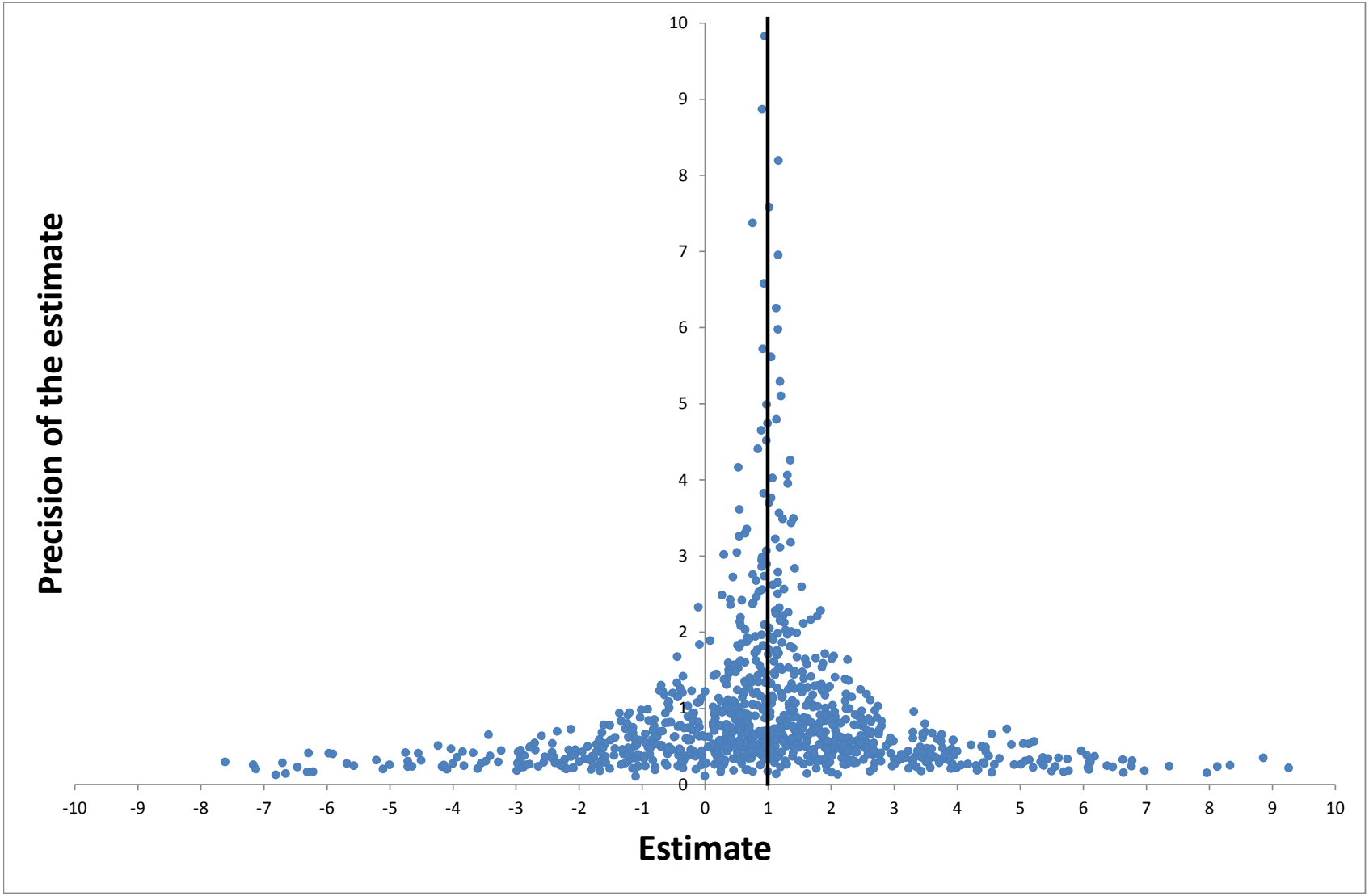
Simulation


10
8
6
4
2
0
-2
-4
-6
-8
-10
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Estimate
Standard error of the estimate

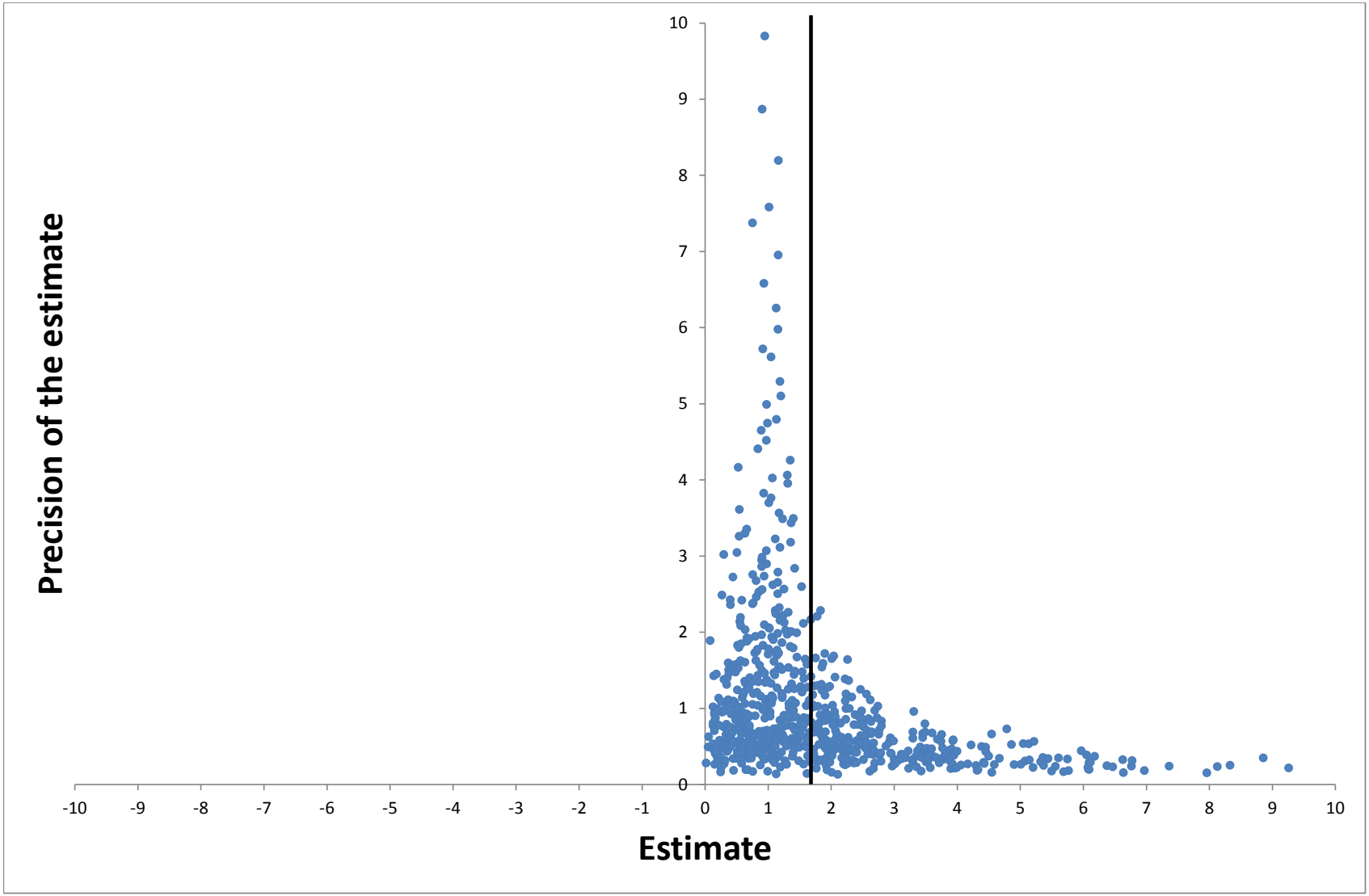
Simulation


Precision of the estimate
Estimate
10
9
8
7
6
5
4
3
2
1
0
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10

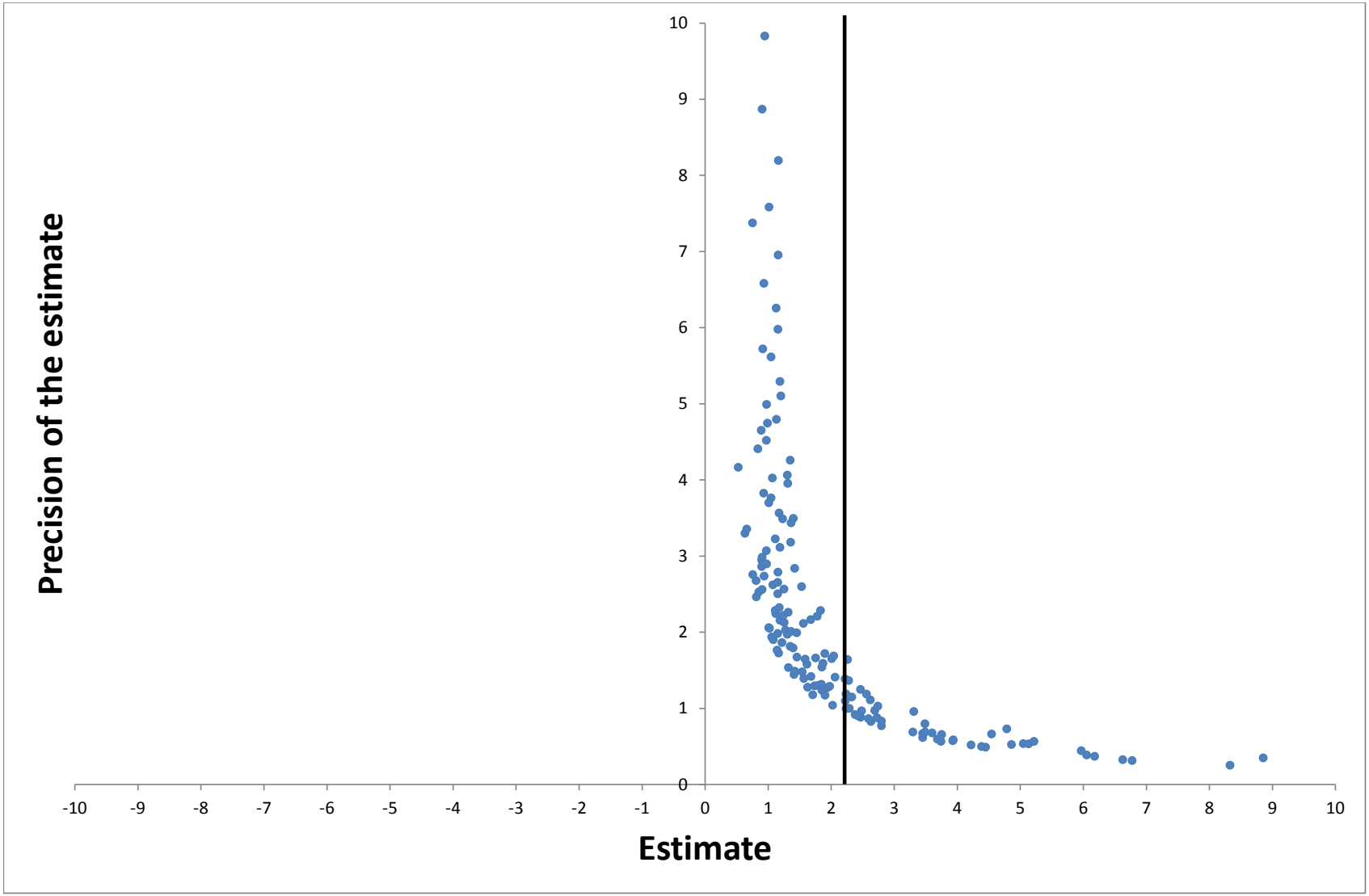
Simulation


Precision of the estimate
Estimate
10
9
8
7
6
5
4
3
2
1
0
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10

Simulation


Precision of the estimate
Estimate
10
9
8
7
6
5
4
3
2
1
0
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10

Simulation


Precision of the estimate
Estimate
10
9
8
7
6
5
4
3
2
1
0
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10

Funnel asymmetry test

In the absence of publication bias the funnel should be symmetrical:

$$\text{estimate}_i = \underbrace{\beta}_{\text{true effect}} + \underbrace{\beta_0 SE_i}_{\text{publication bias}} + \mu_i.$$

The equation is heteroscedastic. Weighted least squares yield

$$t_i = \beta_0 + \beta \left(\frac{1}{SE_i} \right) + \vartheta_i.$$

This is call the funnel asymmetry test—precision effect test (FAT-PET).

Hands on

Open Stata

Default nonlinear correction model (PEESE)

Publication bias unlikely to be linear in SE. Precision effect test with standard error:

$$\text{estimate}_i = \underbrace{\beta}_{\text{true effect}} + \underbrace{\beta_0 SE_i^2}_{\text{publication bias}} + \mu_i.$$

Weighted least squares yield

$$t_i = \beta_0 SE_i + \beta \left(\frac{1}{SE_i} \right) + \vartheta_i.$$

Works well under most circumstances.

WAAP (Ioannidis et al., 2017)

Weighted **A**verage of **A**dequately **P**owered (WAAP) estimates.

- Focus on studies with adequate statistical power to reduce the impact of selective reporting.
- Combines results only from studies with at least 80% power.

Issues:

- Simple and robust.
- Need to estimate the true effect first and then compute power.

Stem-based correction technique (Furukawa, 2019)

Stem-based method:

- A non-parametric estimator that exploits the variance-bias trade-off.
- Focuses on the most precise estimates, referred to as the "stem" of the funnel plot.

Key features:

- Robust to various publication selection processes.
- Does not rely on specific distributional assumptions.

Application:

- Filters studies with the highest precision (lowest standard errors) to estimate the true effect.
- Reduces influence from noisier, potentially biased studies in the broader funnel plot.

Endogenous kink model (Bom & Rachinger, 2019)

Overview:

- Fits a piecewise linear model with a kink determined by the data.
- Identifies where publication selection is less likely to occur.

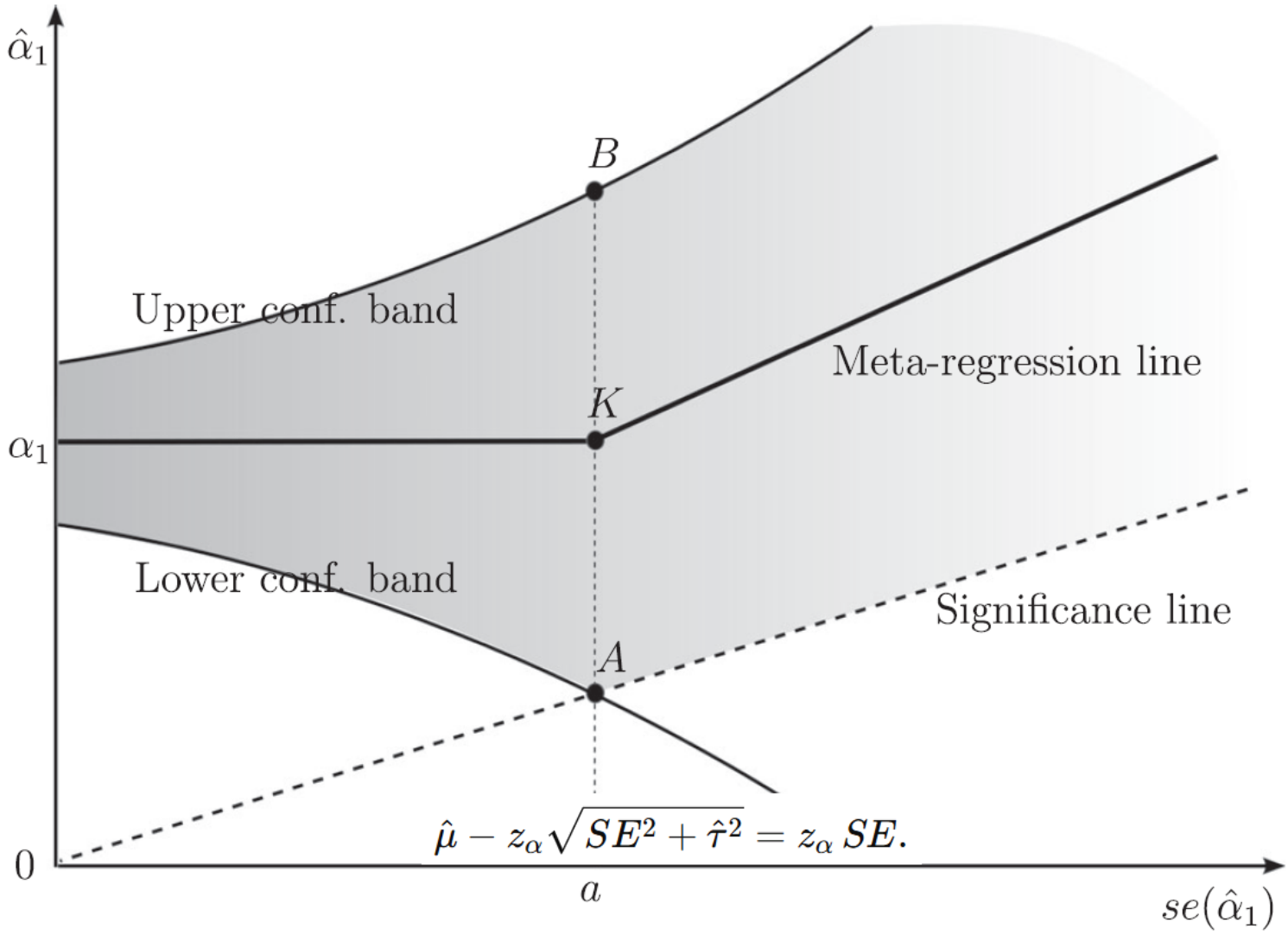
Key features:

- Endogenously estimates the cutoff based on standard errors.
- Often reduces bias and improves efficiency compared to traditional methods.

Application:

- Highlights genuine effects by focusing on less selective segments of the data.
- Effective in meta-analyses with substantial heterogeneity in study precision.

Endogenous kink



Selection model (Andrews & Kasy, 2019)

Overview:

- Models the selection process explicitly to address publication bias.
- Assumes researchers are more likely to publish statistically significant results.

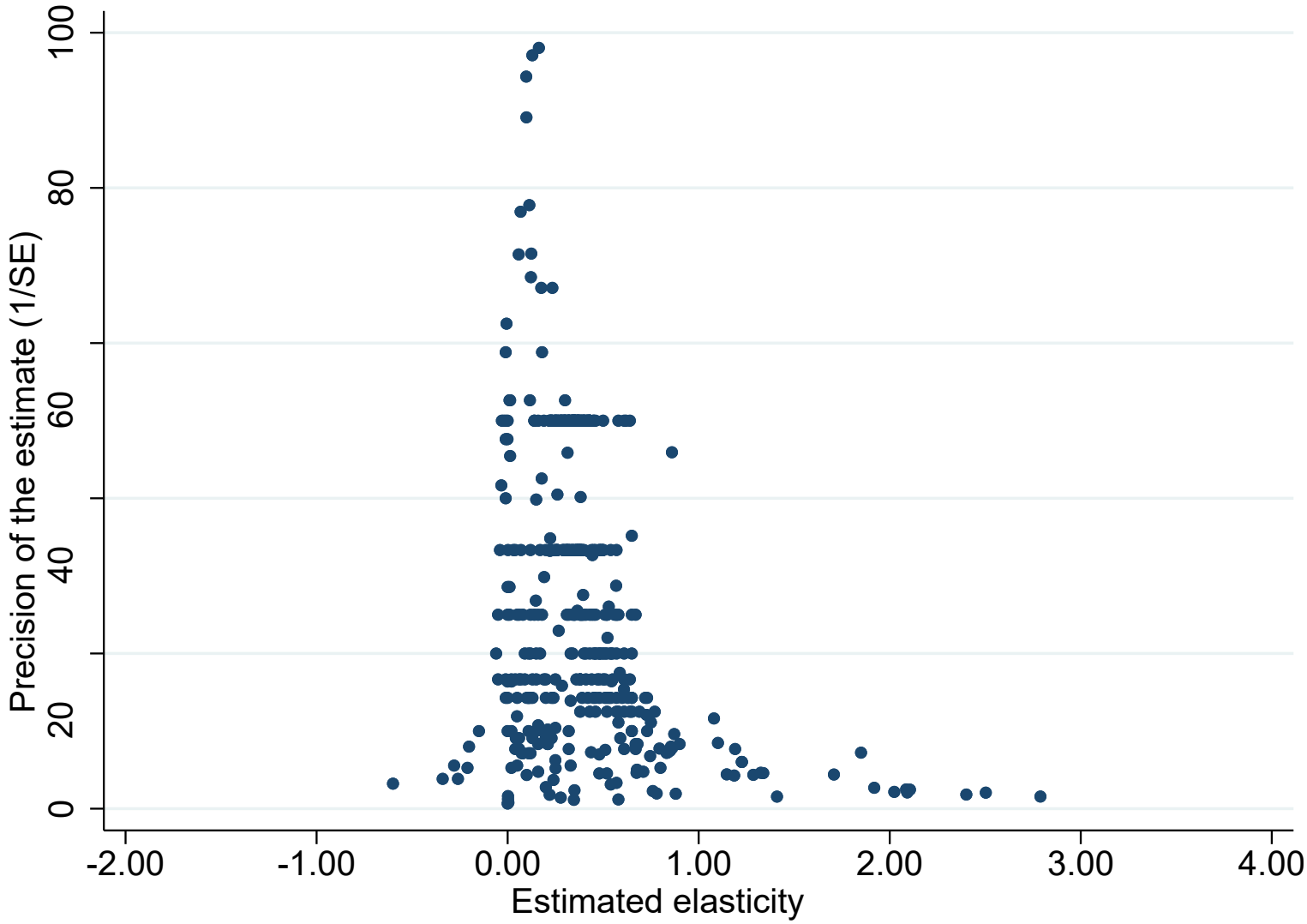
Key features:

- Uses the likelihood of publication as a function of statistical significance.
- Adjusts estimates to reflect the missing studies.

Comparison to funnel-based models:

- Selection models are more rigorous but require stronger assumptions about the publication process.
- Funnel methods are simpler and more robust to model misspecification.

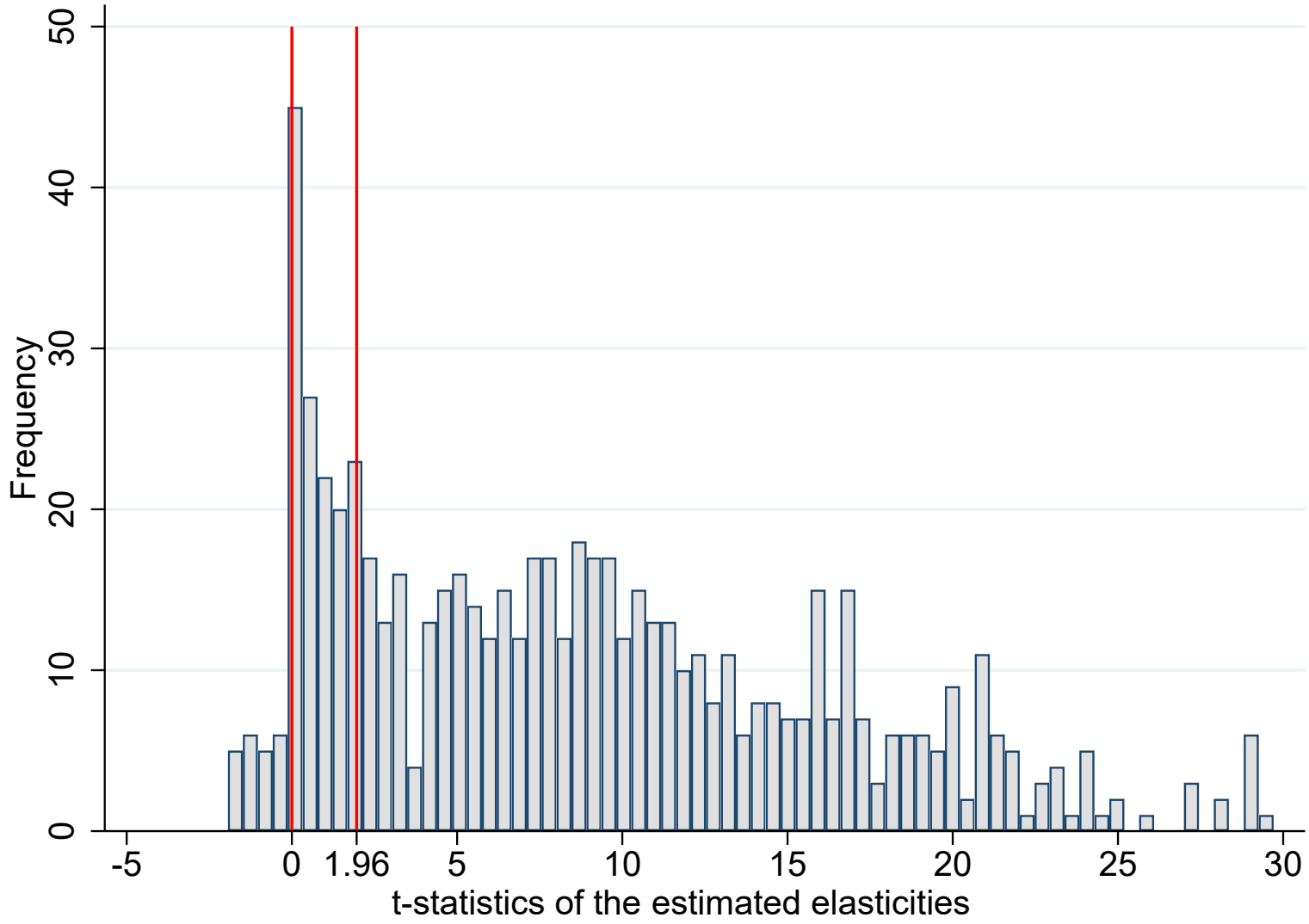
Example: labor supply elasticity



Tests often tell the same story qualitatively

Panel A: Linear tests					
	OLS	FE	Precision	Study	IV
Standard error	1.595***	0.788***	2.222***	2.106***	2.706
(*publication bias*)	(0.262)	(0.283)	(0.484)	(0.258)	(2.103)
	[0.93, 2.20]	-	[1.23, 3.31]	[1.53, 2.56]	-
Constant	0.301***	0.370***	0.247***	0.252***	0.130*
(*mean beyond bias*)	(0.044)	(0.026)	(0.065)	(0.109)	(0.066)
	[0.12, 0.42]	-	[0.11, 0.31]	[0.13, 0.38]	-
Observations	723	723	723	723	522
Studies	36	36	36	36	23
Panel B: Nonlinear tests					
	Ioannidis *et al.* (2017)	Andrews & Kasy (2019)	Bom & Rachinger (2019)	Furukawa (2021)	van Aert & van Assen (2021)
Effect beyond bias	0.260***	0.356***	0.207**	0.187*	0.374***
	(0.041)	(0.061)	(0.094)	(0.112)	(0.020)
Observations	723	723	723	723	723
Studies	36	36	36	36	36

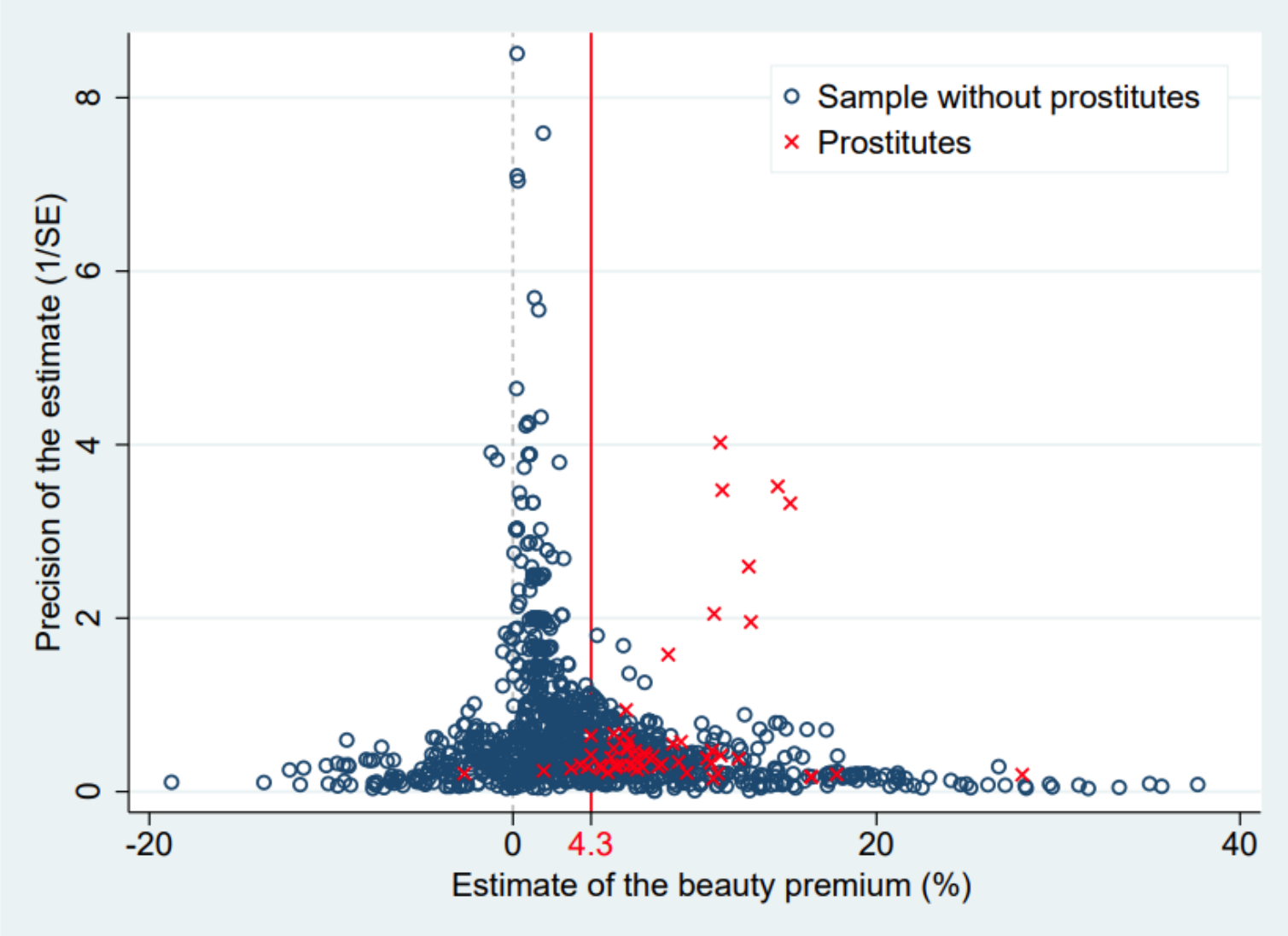
Bias caused by preference for positive estimates



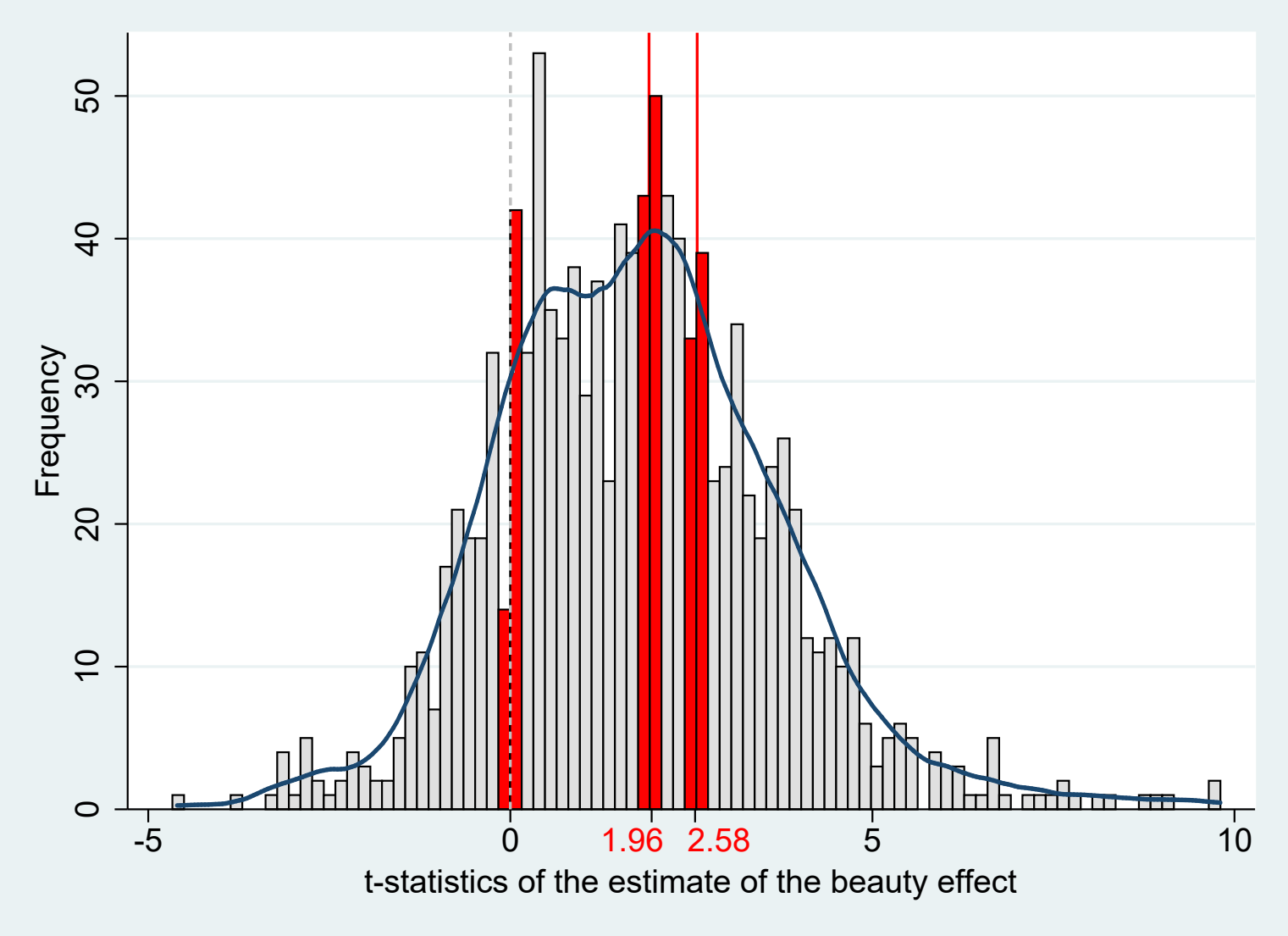
Beauty effect: corrected mean probably below 3

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.377***	0.208*	0.732***	0.758	0.656**
(*standard error*)	(0.118)	(0.119)	(0.164)	(0.489)	(0.272)
	[0.116, 0.634]			{-0.064, 2.160}	[-0.009, 1.268]
Effect beyond bias	2.865***	3.497***	2.243**	1.436	3.424***
(*constant*)	(0.464)	(0.446)	(0.871)	(1.866)	(1.156)
	[1.916, 3.781]			{-0.378, 3.251}	[0.776, 6.339]
First-stage robust F-stat				15.8	
Observations	1,159	1,159	1,159	1,159	1,159
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.720***			1.720***	P = 0.142
	(0.084)			(0.084)	(0.038)
	[1.262, 2.166]			[1.262, 2.166]	
Effect beyond bias	0.343	0.323**	0.055	0.343	0.493
	(0.118)	(0.147)	(1.276)	(0.118)	(0.410)
	[-0.039, 1.635]			[-0.039, 1.635]	
Observations	1,159	1,159	1,159	1,159	1,159

Sex workers stand apart



Bias caused by selection for a positive sign



Workshop Outline

Recall publication bias

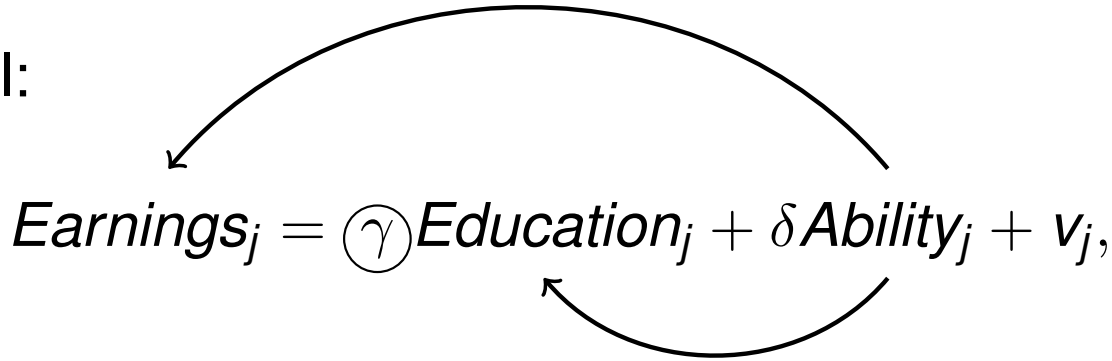
Recall publication bias

Key difference

- **Publication bias**: all estimates are individually unbiased
- The entire literature is biased because estimates are selectively reported
- If we know publication probabilities, we can recover the true effect (selection models)
- **P-hacking:** individual estimates can be biased
- Classical selection models fail
- Funnel-based models can work (with adjustments)

Example: education premium

True model:



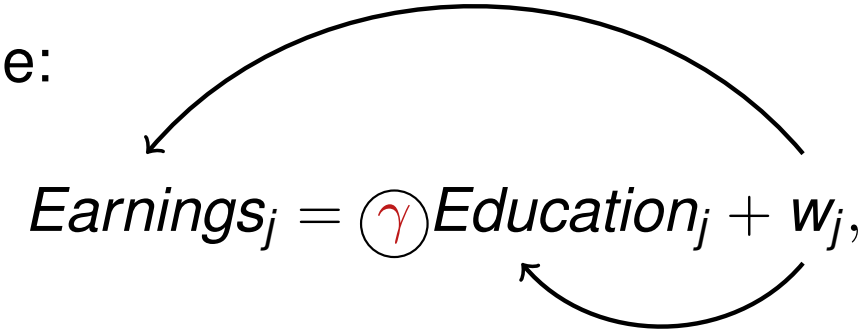
Ability not observed.

Primary studies:

1. ignore ability $\rightarrow \hat{\gamma}$ too large, SE($\hat{\gamma}$) too small.
2. include a proxy $\rightarrow \hat{\gamma}$ smaller, SE($\hat{\gamma}$) larger.
3. quasi-experiment $\rightarrow \hat{\gamma}$ even smaller, SE($\hat{\gamma}$) even larger.

Example: education premium

Omitted variable:



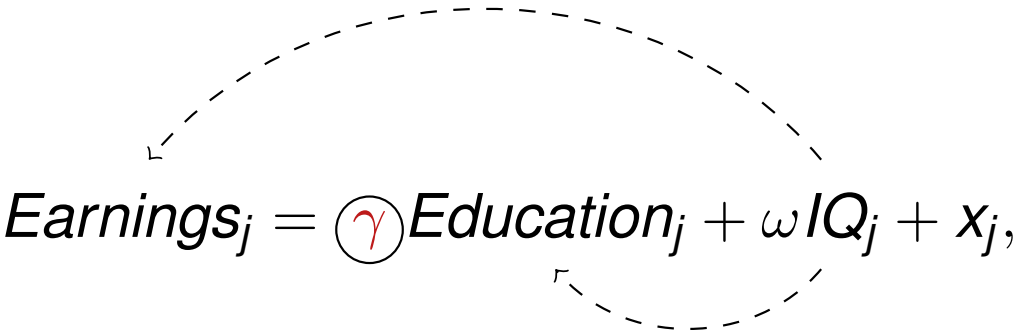
Ability not observed.

Primary studies:

1. ignore ability $\rightarrow \hat{\gamma}$ too large, SE($\hat{\gamma}$) too small.
2. include a proxy $\rightarrow \hat{\gamma}$ smaller, SE($\hat{\gamma}$) larger.
3. quasi-experiment $\rightarrow \hat{\gamma}$ even smaller, SE($\hat{\gamma}$) even larger.

Example: education premium

Proxy:



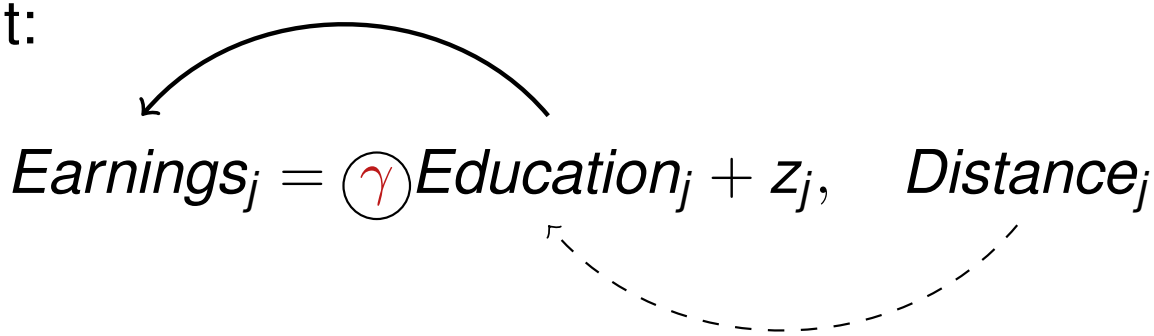
Ability not observed.

Primary studies:

1. ignore ability $\rightarrow \hat{\gamma}$ too large, SE($\hat{\gamma}$) too small.
2. include a proxy $\rightarrow \hat{\gamma}$ smaller, SE($\hat{\gamma}$) larger.
3. quasi-experiment $\rightarrow \hat{\gamma}$ even smaller, SE($\hat{\gamma}$) even larger.

Example: education premium

Instrument:

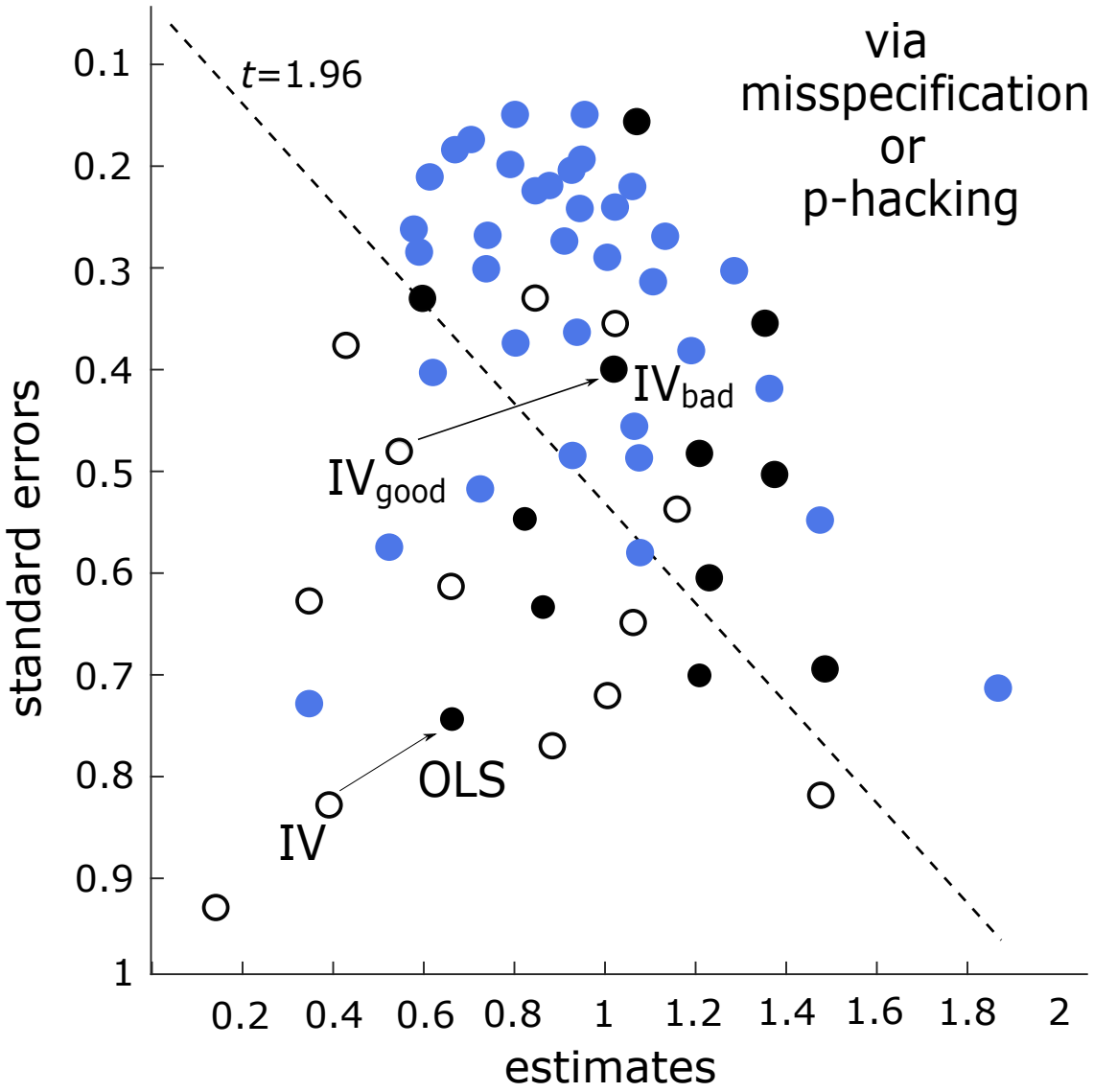


Ability not observed.

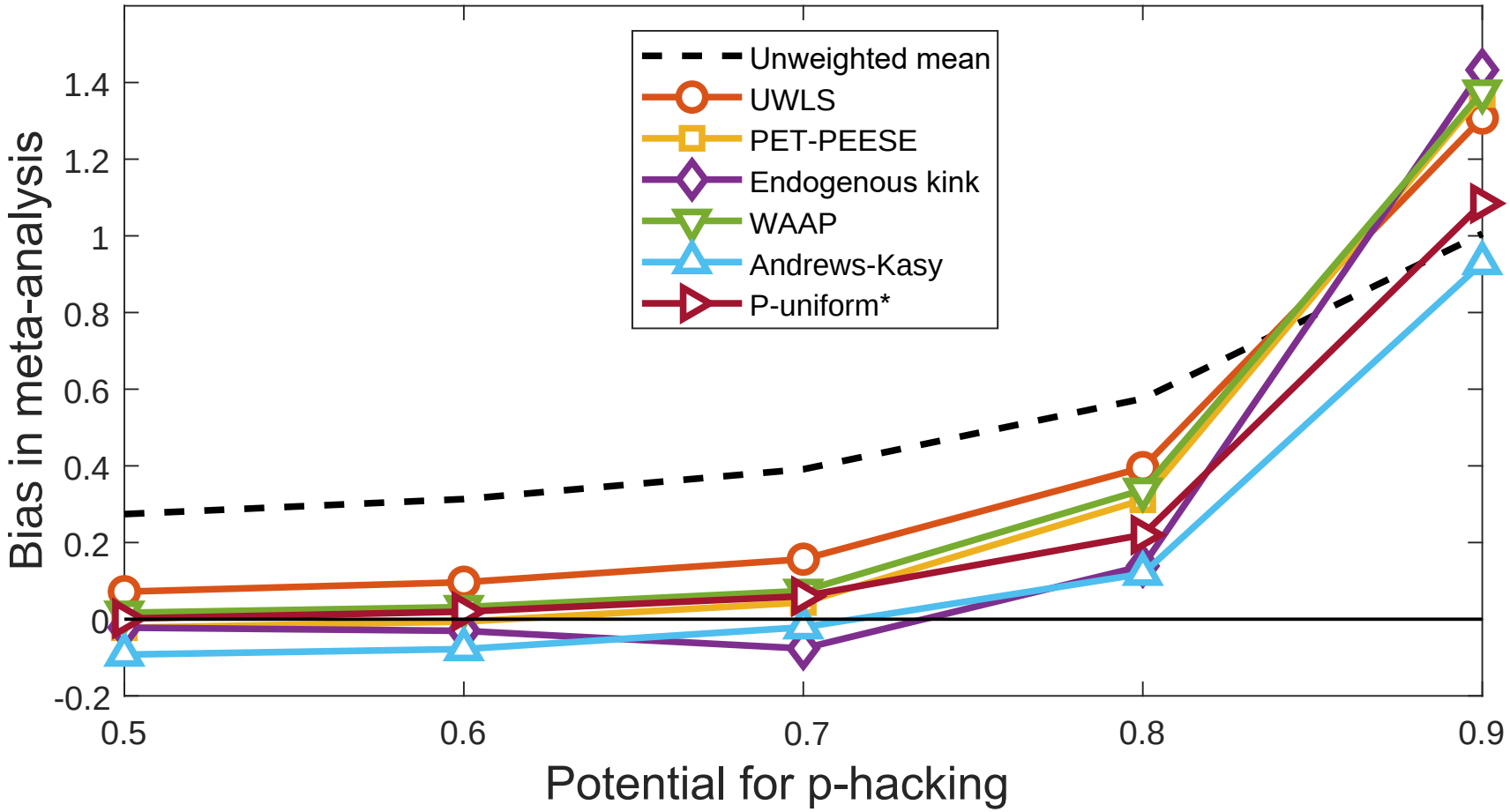
Primary studies:

1. ignore ability $\rightarrow \hat{\gamma}$ too large, SE($\hat{\gamma}$) too small.
2. include a proxy $\rightarrow \hat{\gamma}$ smaller, SE($\hat{\gamma}$) larger.
3. quasi-experiment $\rightarrow \hat{\gamma}$ even smaller, SE($\hat{\gamma}$) even larger.

Some estimates spuriously large & precise



All estimators biased upwards



Example: Changing controls changes precision

Table: Regression of test scores on class size

	(1)	(2)	(3)	(4)
Small class (treatment)	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
White/Asian			8.35	8.44
			(1.35)	(1.36)
Girl			4.48	4.39
			(0.63)	(0.63)
Free lunch			-13.15	-13.07
			(0.77)	(0.77)
White teacher				-0.57
				(2.1)
Teacher experience				0.26
				(0.10)
Master's degree				-0.51
				(1.06)
School intercepts	No	Yes	Yes	Yes
Sample	5,861	5,861	5,861	5,861

Notes: Adapted from Krueger (1999). Dependent variable: test score percentile.
Standard errors in parentheses.

Changing clustering changes precision

Table: Regression of test scores on class size

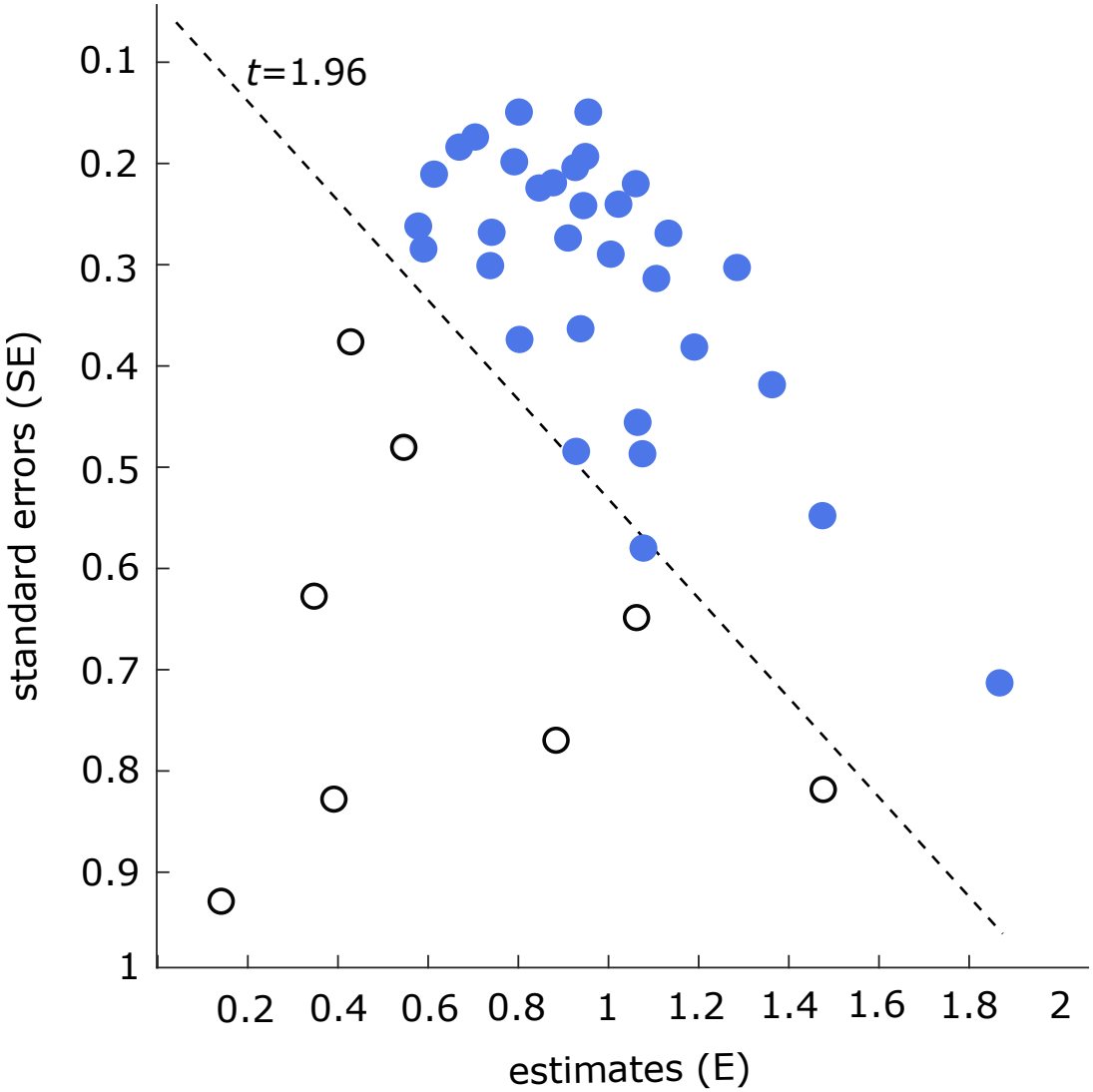
	Krueger (1999)	Replications using different computations of SE				
	(1)	(2) Bootstrap of class clusters	(3) Class clusters	(4) School clusters	(5) Huber-White SE	(6) Plain vanilla SE
Small class	4.82	4.71	4.71	4.71	4.71	4.71
	(2.19)	(2.00)	(1.88)	(1.38)	(0.79)	(0.76)
Sample	5,861	5,743	5,743	5,743	5,743	5,743

Notes: Dependent variable: test score percentile. Standard errors (SE) in parentheses.

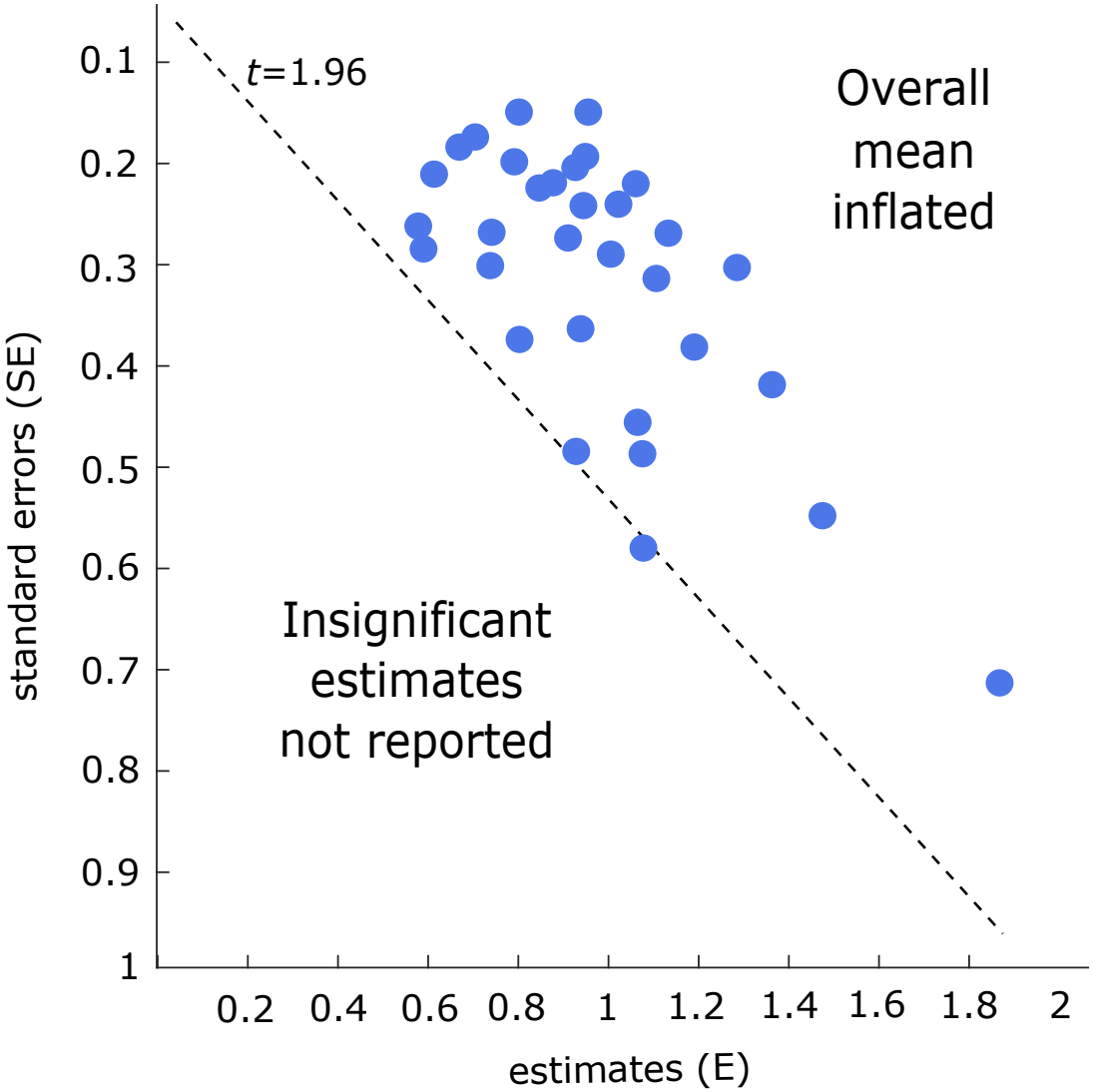
Meta-analysis of the class size effect

Available at meta-analysis.cz/class. (forthcoming in *JOLE*)

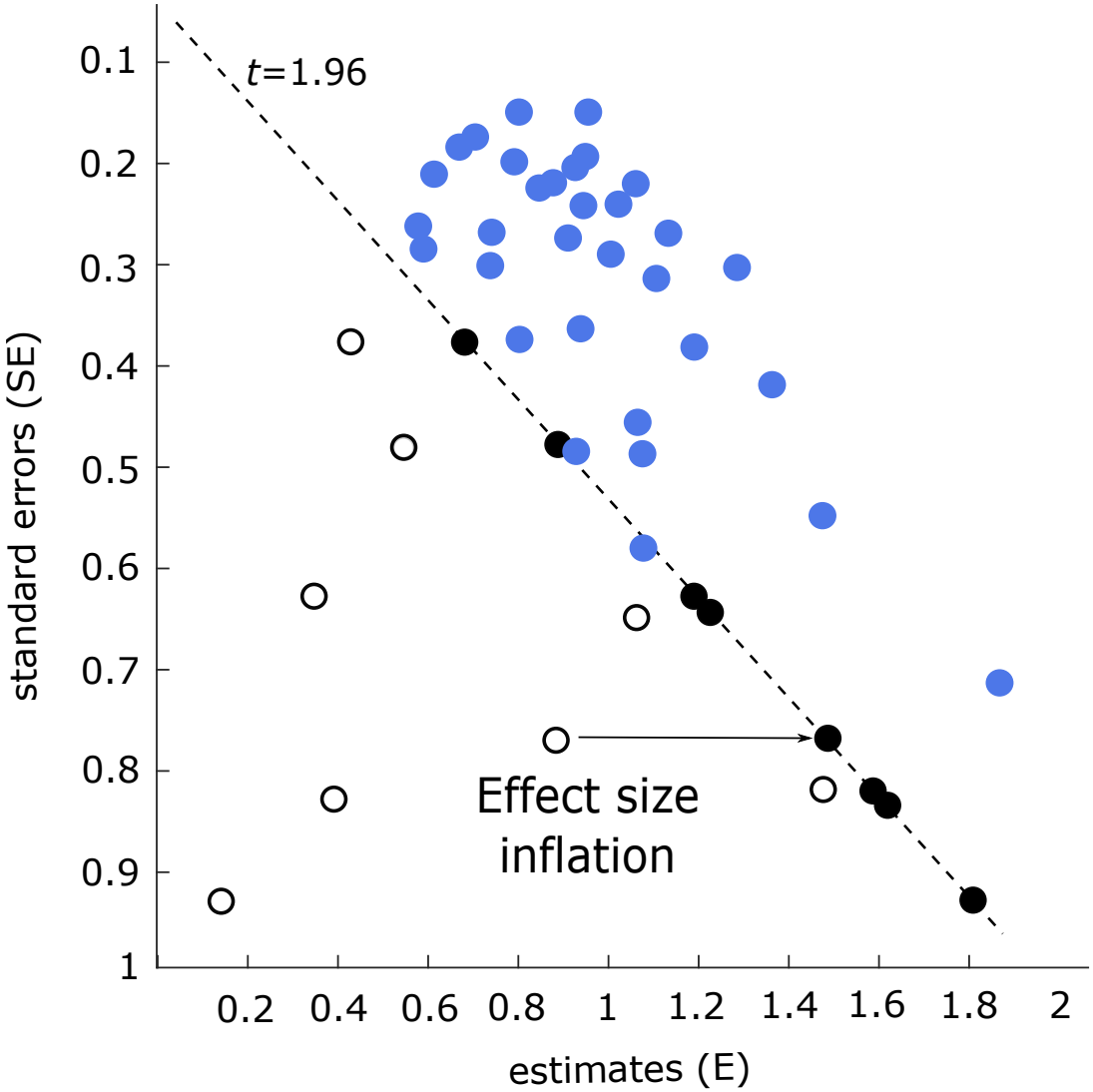
Funnel plot


t=1.96
standard errors (SE)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
estimates (E)

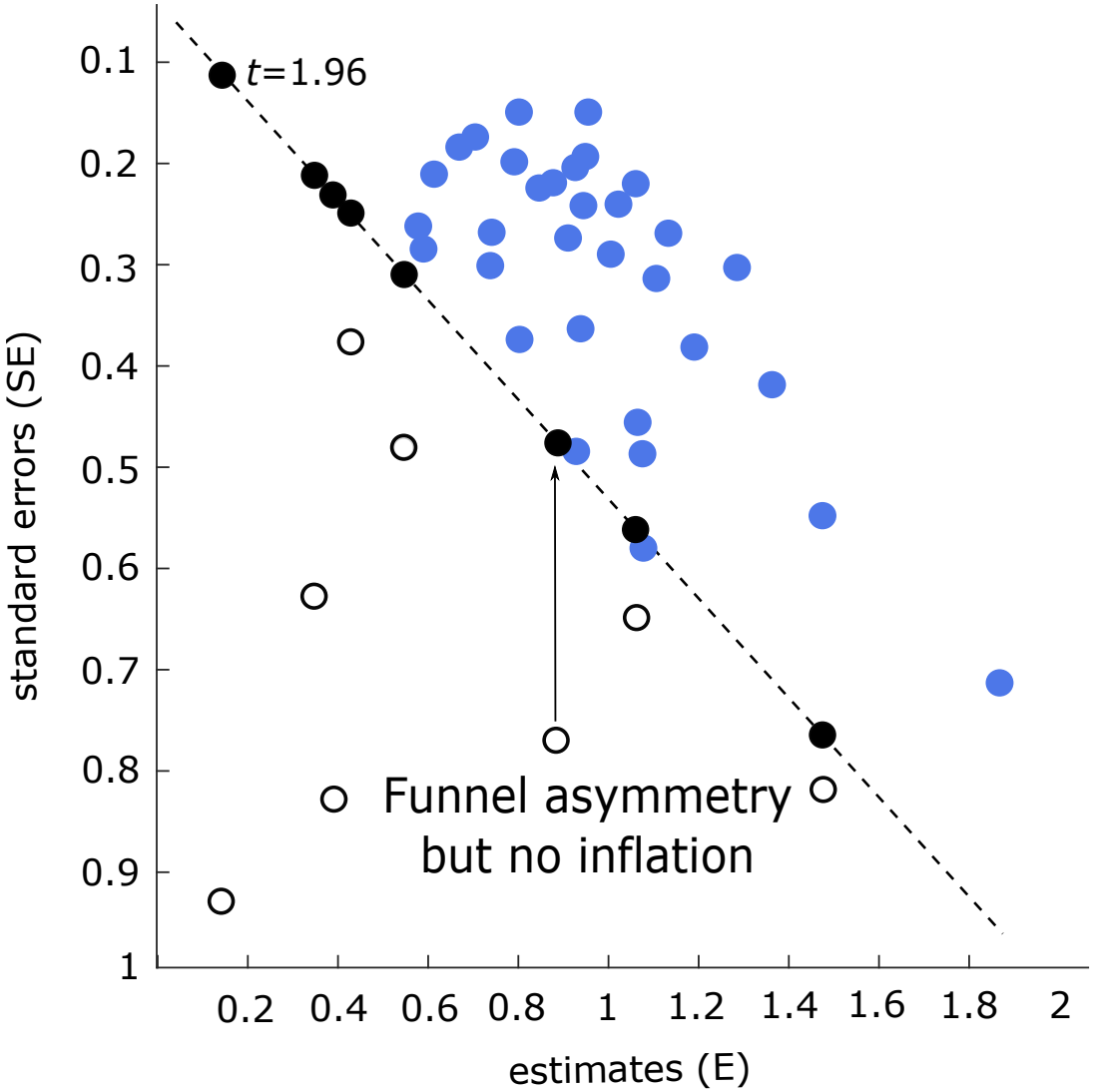
Publication bias



Conventional p-hacking

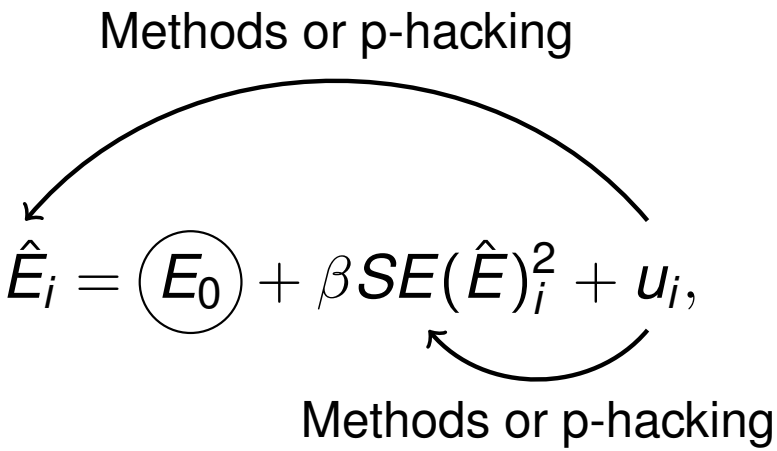


Spurious precision



Key meta assumption broken

PEESE:



$\text{corr}(SE, u) \neq 0 \Rightarrow \hat{\beta}$ and $\hat{E}_0$ biased.

Natural solution: N instrumenting $SE(\hat{E})_i^2 \rightarrow$ MAIVE.

Options for the meta-analyst

1. Use only quasi-experimental studies.
2. Include controls (dummies for OLS, DID, . . .).

   But in observational research
   we never know the true model!

3. Remove “bad” variation from SE $\rightarrow$ MAIVE.

Meta-analysis instrumental variable estimator

MAIVE intuition:

Definition of SE

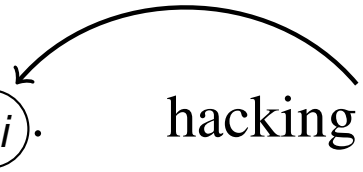
$$\hat{E}_i = \boxed{E_0} + \beta SE(\hat{E})_i^2 + u_i. \quad 1/N_i$$

- MAIVE + PEESE = classical IV.
- Can add controls in the 2nd stage.
- Plug MAIVE-adjusted SEs into other estimators.

Meta-analysis instrumental variable estimator

MAIVE first stage:

$$SE(\hat{E})_i^2 = \alpha_0 + \alpha_1 (1/N_i) + \pi_i.$$



hacking or misspecifications

- MAIVE + PEESE = classical IV.
- Can add controls in the 2nd stage.
- Plug MAIVE-adjusted SEs into other estimators.

Meta-analysis instrumental variable estimator

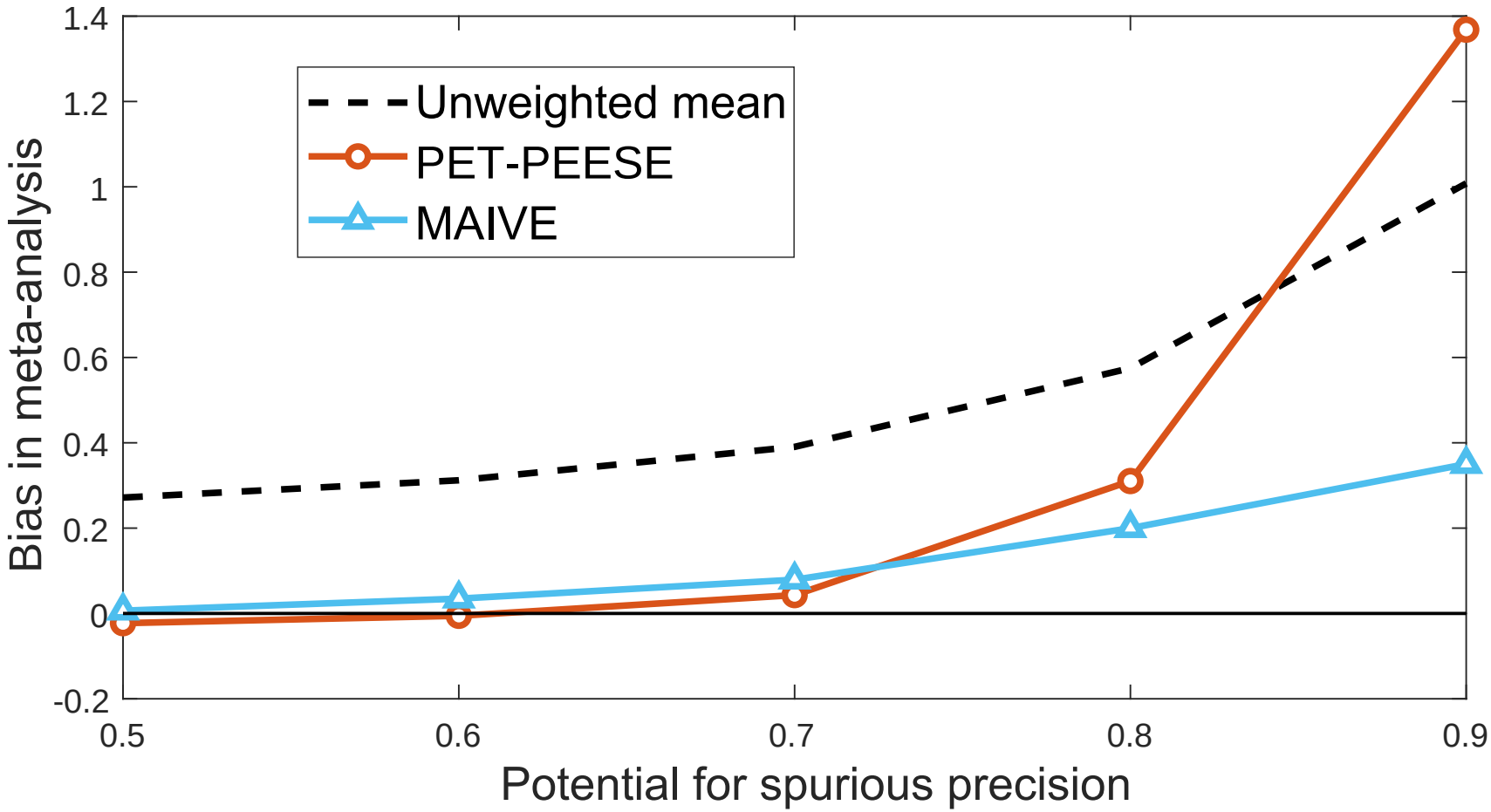
MAIVE adjustment:

$$SE(\hat{E})^2_{adj,i} = \hat{\alpha}_0 + \hat{\alpha}_1 (1/N_i) + \cancel{\pi_i}.$$

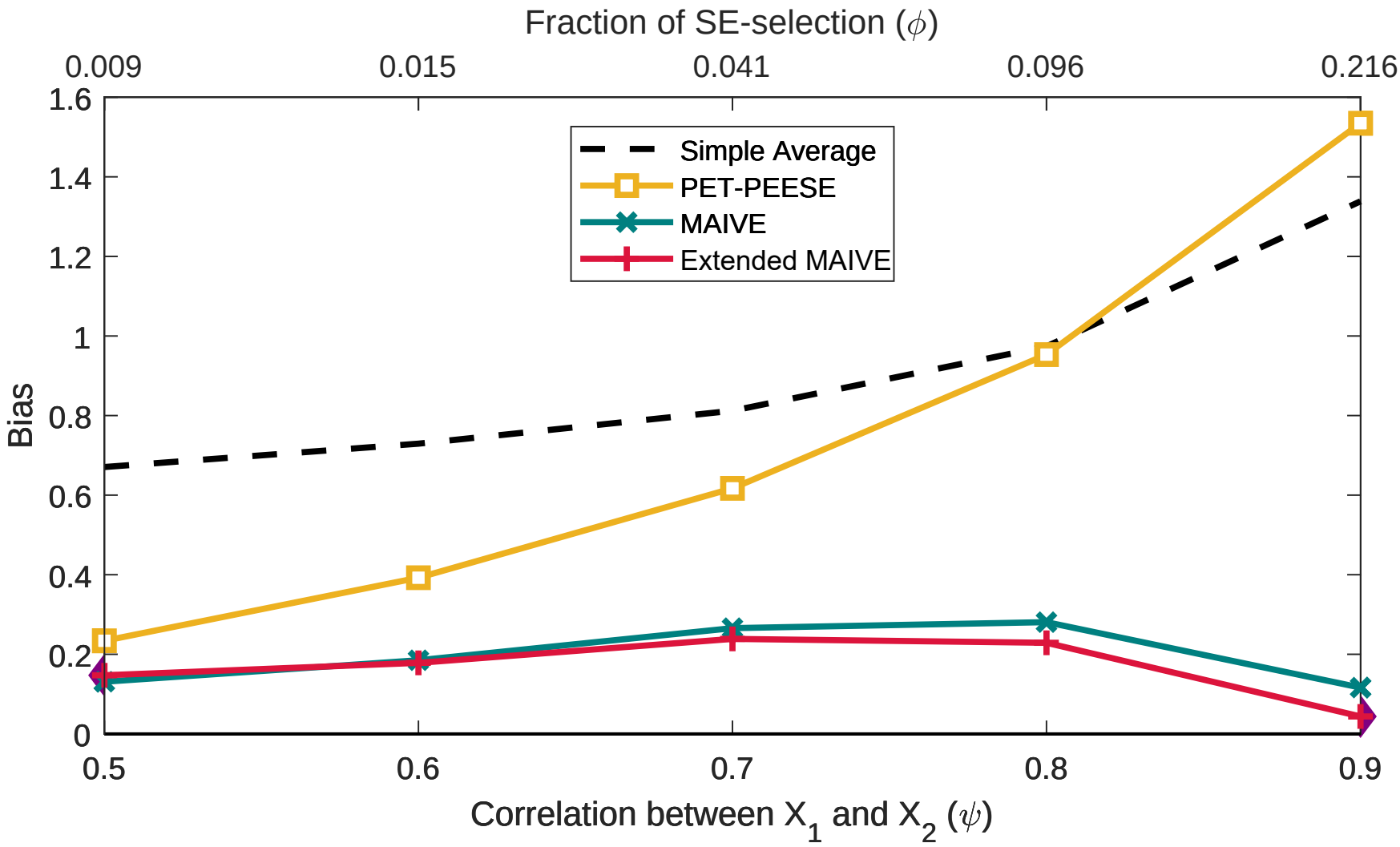
~~hacking~~ or ~~misspecifications~~

- MAIVE + PEESE = classical IV.
- Can add controls in the 2nd stage.
- Plug MAIVE-adjusted SEs into other estimators.

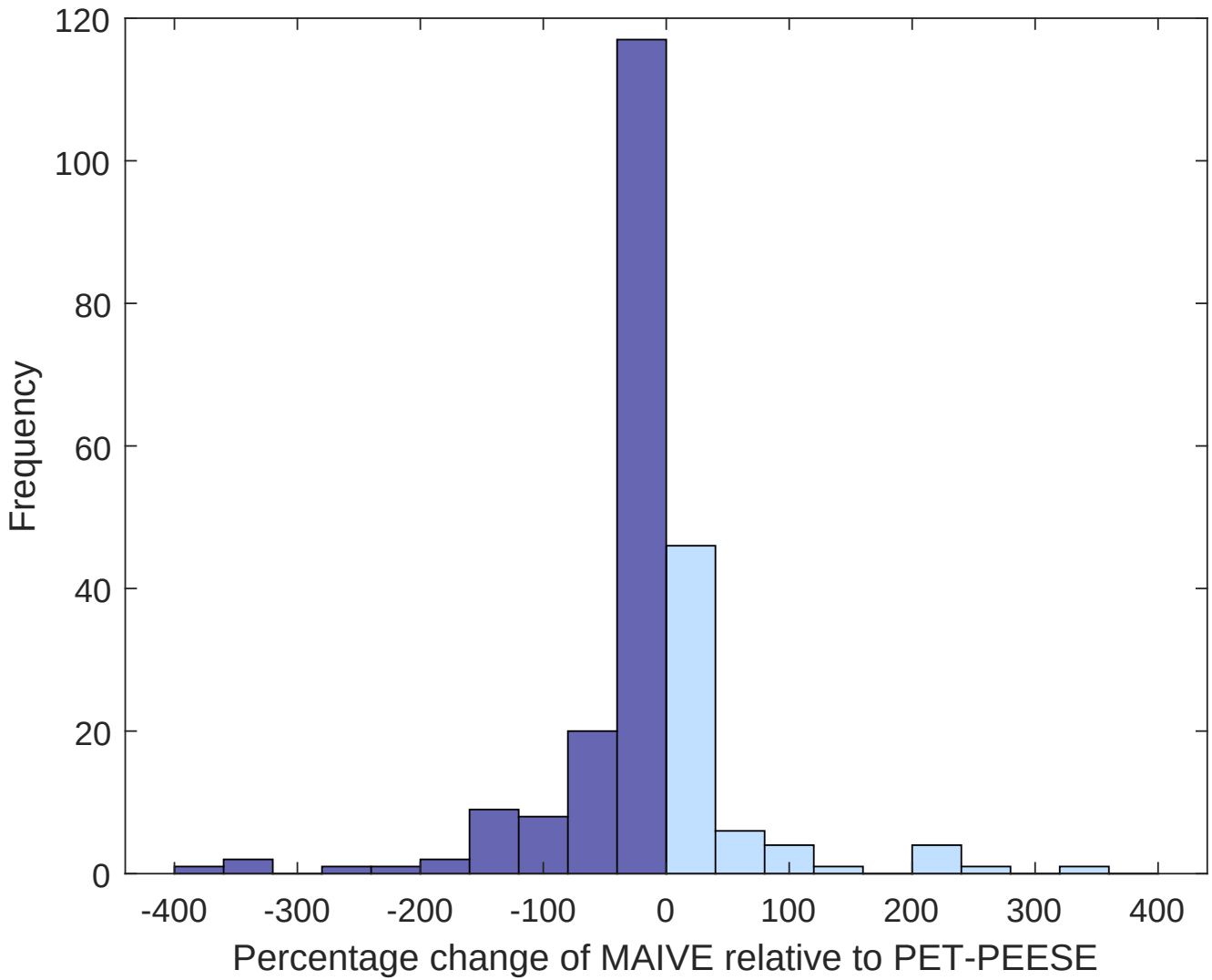
MAIVE alleviates the bias



Extended MAIVE



MAIVE reduces PET-PEESE in 70% of the cases



Practical issues

Main consideration

- Can methods or p-hacking influence both $\hat{E}$ and SE?
- Yes $\rightarrow$ MAIVE helps.
- No $\rightarrow$ MAIVE doesn't hurt much.

Project Website

meta-analysis.cz/maive

Forthcoming in *Nature Communications*.

Hands on

Open Stata

(MAIVE web app at maive.eu)

RTMA: Maya Mathur's p-hacking correction

- Right-truncated meta-analysis (Mathur 2024, RSM)
- Assumption 1: insignificant estimates are NOT p-hacked
- Assumption 2: estimates are normally distributed
- Assumption 3: there are enough insignificant estimates published that we can recover the underlying distribution
- Bayesian techniques used for computation feasibility (ML would typically need huge samples)
- Digression: Robust Bayesian Meta-Analysis (RoBMA), still unavailable for economics data

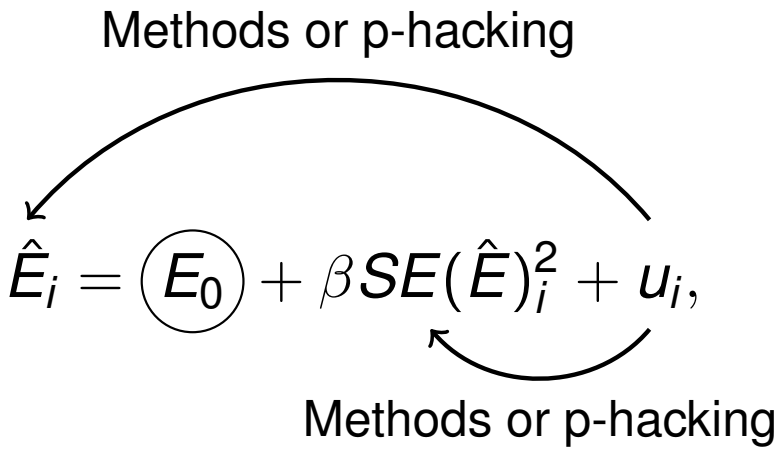
Hands on

Open R

Workshop Outline

Identification of publication bias

PEESE:



corr(SE, u) $\neq 0 \Rightarrow \hat{\beta}$ and $\hat{E}_0$ biased.

Natural solution: N instrumenting $SE(\hat{E})_i^2 \rightarrow$ MAIVE.

Or add covariates!

Classical heterogeneity measure

Common measure: estimate the relative heterogeneity of underlying effects *(Higgins and Thompson, 2002)*

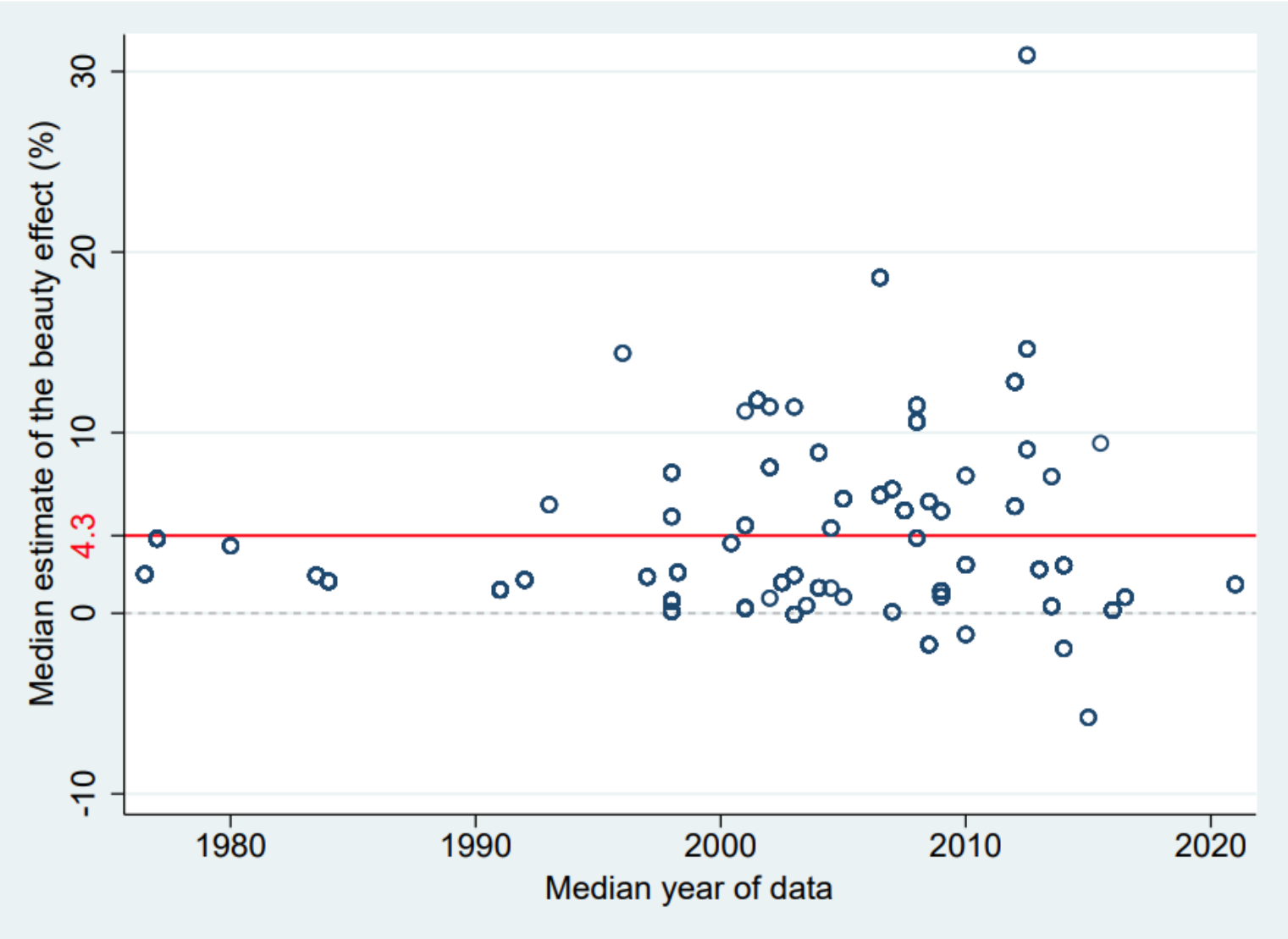
$$I^2 \equiv \frac{\text{underlying variance of effects}}{\text{observed variance of estimates}} = \frac{\tau^2}{\tau^2 + \bar{v}}$$

- $I^2 = 0.75 \sim 1.0$ . . . *"considerable heterogeneity" in medicine (Cochrane Review Guideline)*
- $I^2 \simeq 0.9$ . . . common in economics meta-analyses (e.g., *Vivalt, 2021*)

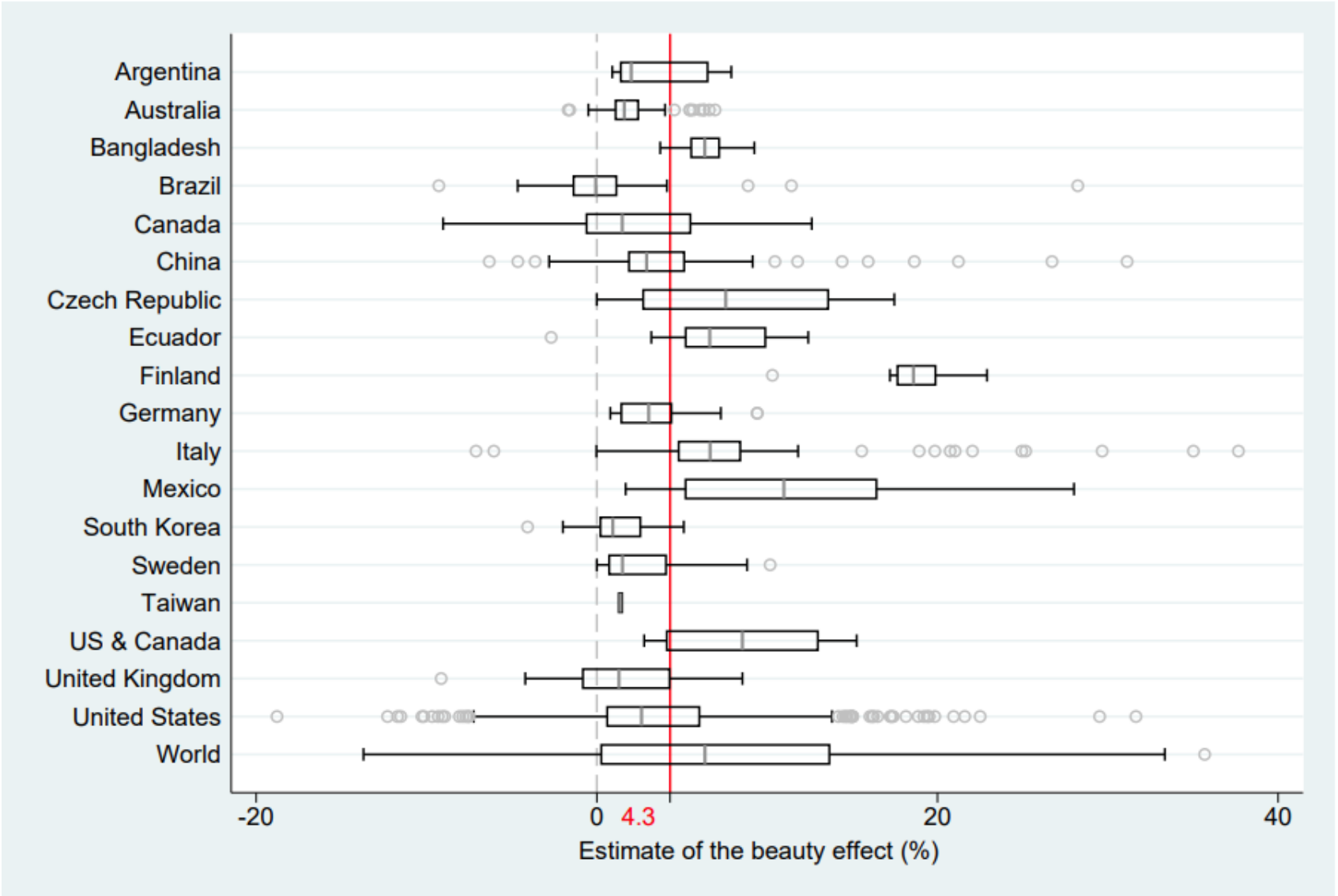
Unresolved concern:

- no universally agreed criteria for "too much" heterogeneity
- undesirable pattern: with more precise studies, higher I^2

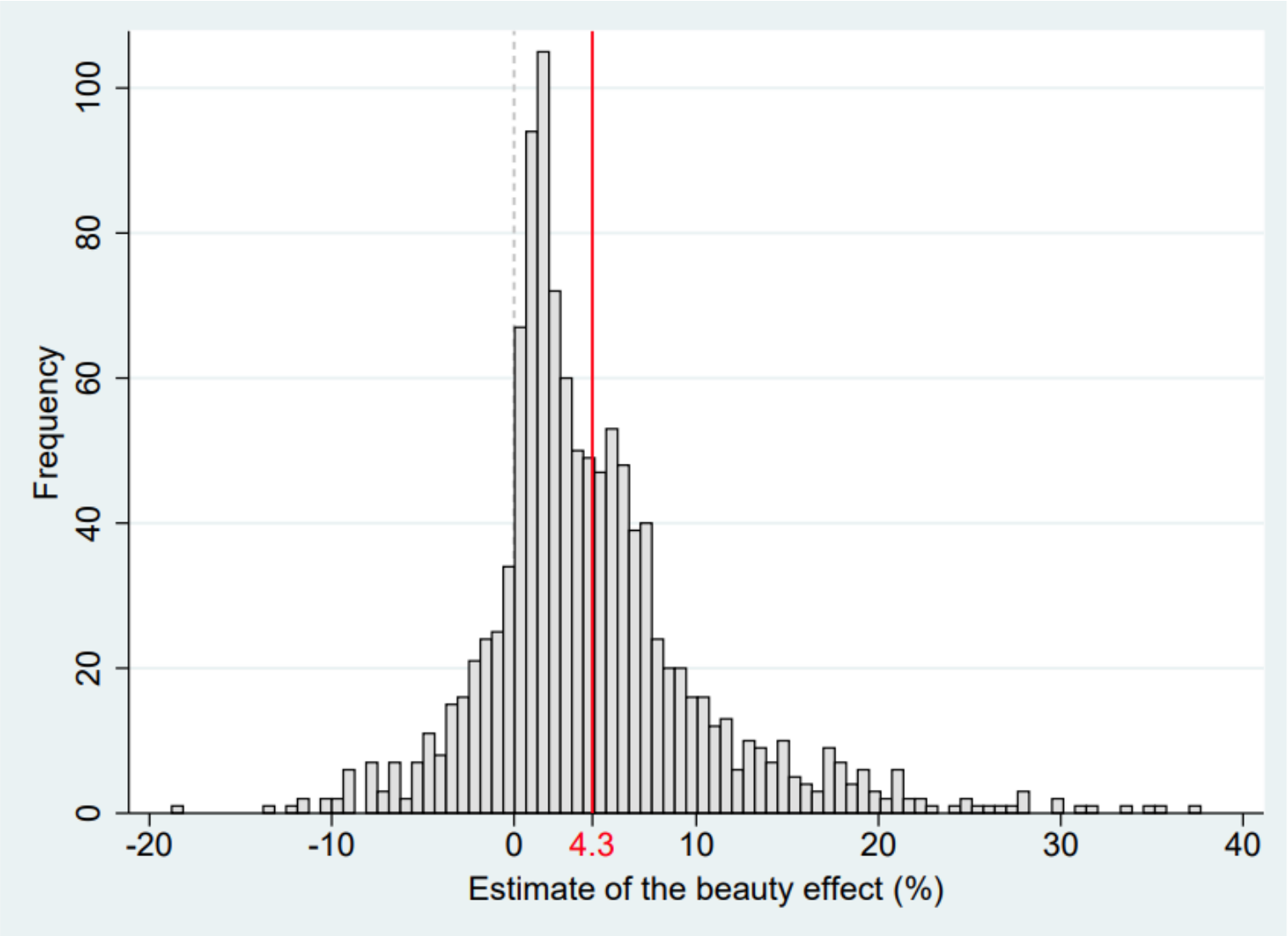
Guiding example: beauty and success



Heterogeneity both within and across



Wide dispersion



Measurement of beauty

			Unweighted		
	Est.	Stud.	Mean	95% conf.	int.
Measurement of beauty					
Interviewer-rated beauty	526	24	4.33	3.80	4.86
Photo-rated beauty	481	37	4.28	3.71	4.85
Software-rated beauty	104	8	5.49	4.15	6.83
Self-rated beauty	56	6	2.04	0.79	3.29
Dummy beauty	460	23	3.63	3.05	4.22
Categorical beauty	699	52	4.70	4.24	5.16
Beauty premium	954	67	4.48	4.07	4.89
Beauty penalty	205	16	3.33	2.56	4.10

Measurement of success

			Unweighted		
	Est.	Stud.	Mean	95% conf. int.	
Measurement of success					
Earnings	688	43	4.33	3.92	4.75
Study outcomes	189	9	3.19	1.98	4.41
Teaching & research outcomes	139	8	4.95	4.05	5.85
Athletic success	33	2	6.28	3.03	9.54
Electoral success	37	4	7.00	4.26	9.75
Other outcomes	73	6	2.98	2.24	3.73

Data characteristics

			Unweighted		
	Est.	Stud.	Mean	95% conf.	int.
Data characteristics					
Male subjects	375	44	3.66	3.10	4.22
Female subjects	447	48	4.10	3.45	4.75
Mixed gender	337	36	5.21	4.57	5.85
High-skilled workers	335	27	4.30	3.78	4.81
Prostitutes	55	4	8.55	7.27	9.82
Other dressy occupations	138	15	4.92	3.89	5.96
Non-dressy occupations	966	51	3.94	3.55	4.34
Western culture	874	47	3.85	3.46	4.23
Other cultures	285	21	5.60	4.72	6.48
Panel data	998	55	3.86	3.47	4.25
Cross-section	161	13	6.87	5.90	7.83

Estimation technique

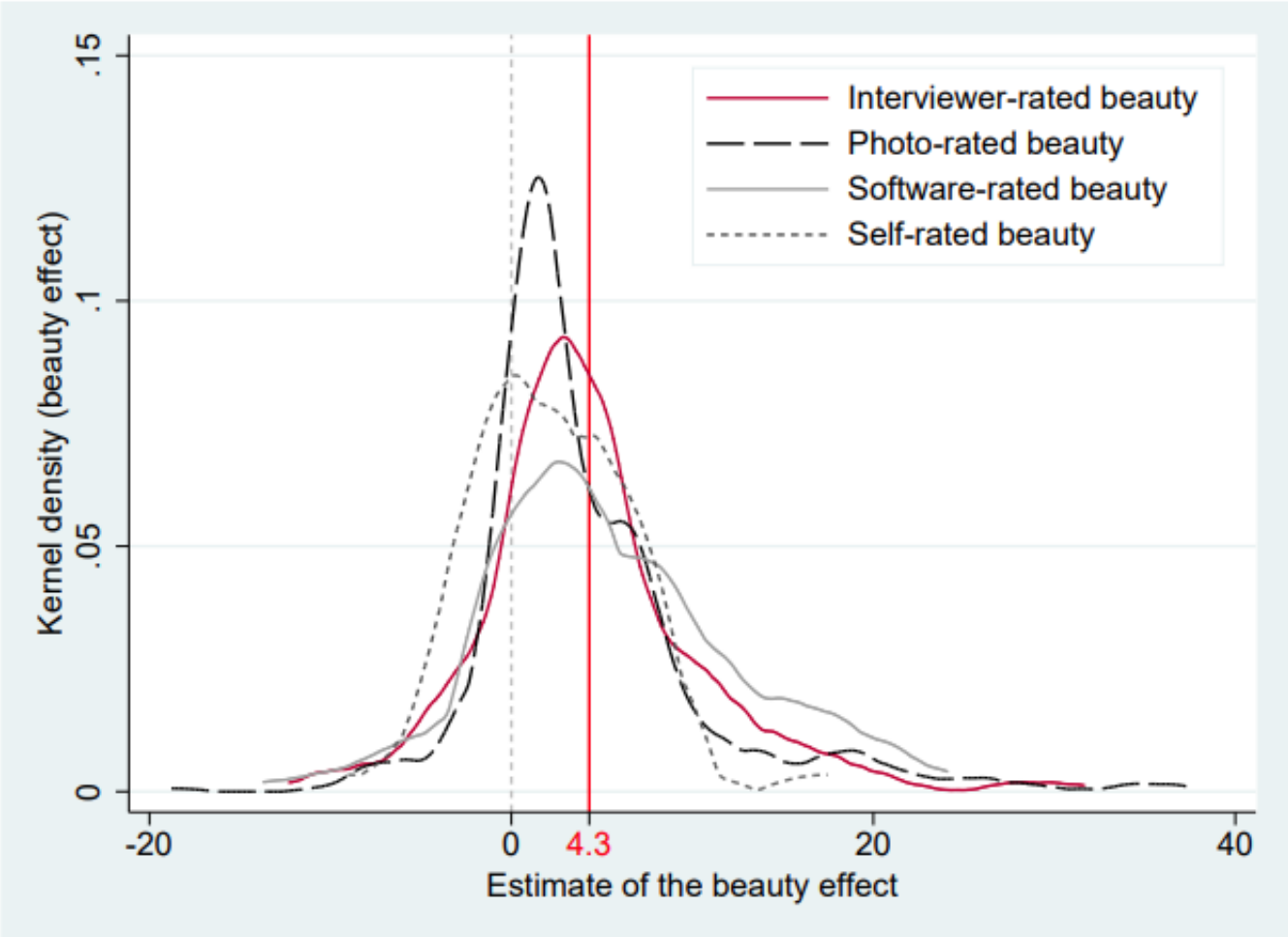
			Unweighted		
	Est.	Stud.	Mean	95% conf. int.	
Estimation technique					
Ordinary least squares	903	60	4.17	3.77	4.56
Instrumental variables	83	10	4.86	3.07	6.64
Difference-in-differences	12	2	1.24	0.63	1.84
Other method	161	13	4.84	3.78	5.90
Cognitive skill control	445	26	2.77	2.17	3.36
No cognitive skill control	714	53	5.22	4.78	5.66

Publication characteristics

			Unweighted		
	Est.	Stud.	Mean	95% conf.	int.
Publication characteristics					
Published study	1,027	58	4.19	3.80	4.57
Unpublished study	132	9	4.99	3.96	6.03
High-quality peer review	151	7	5.40	4.56	6.25

Beauty measurement doesn't matter

(a) Beauty measurement

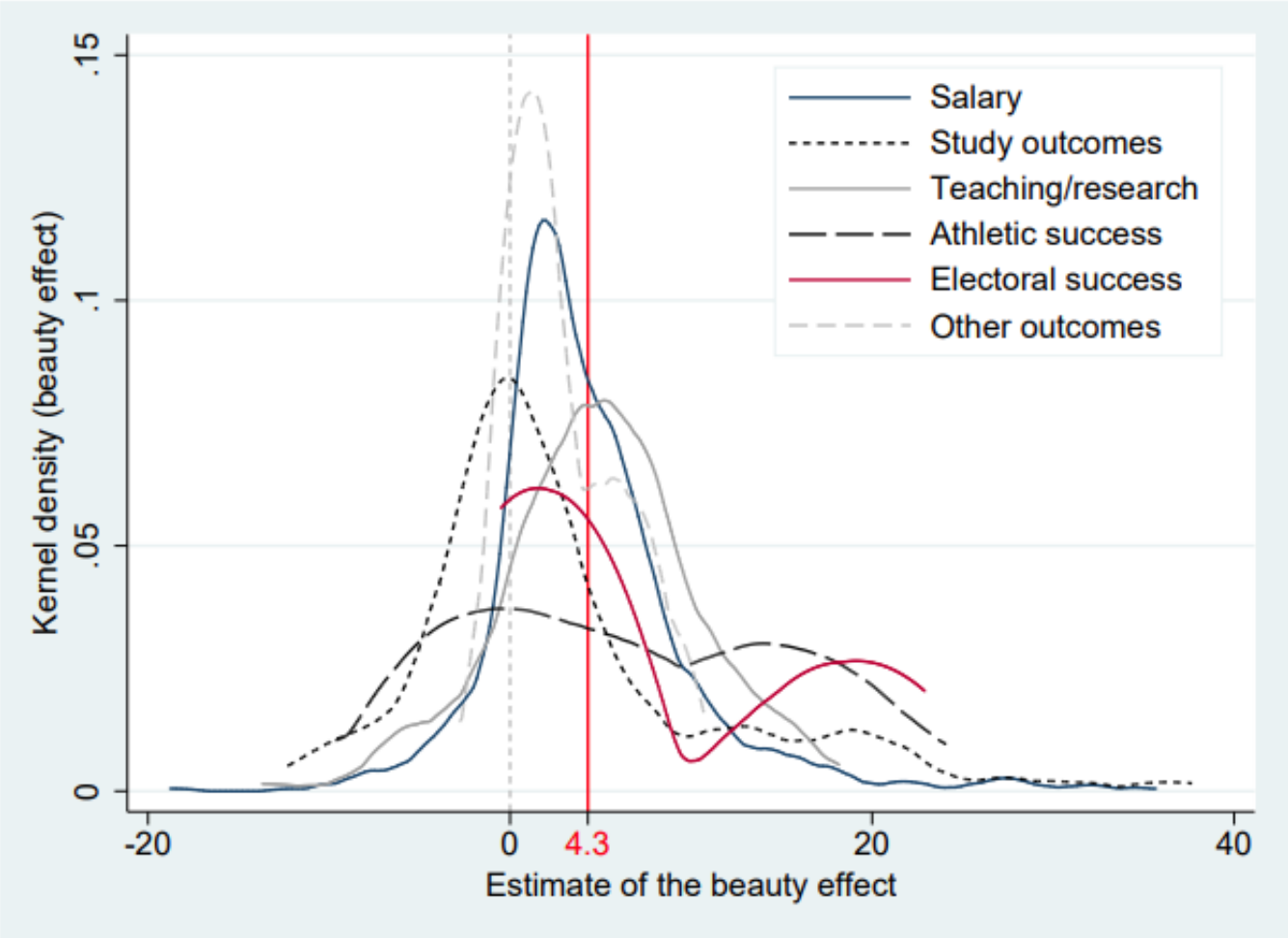


Hands on

Open Stata

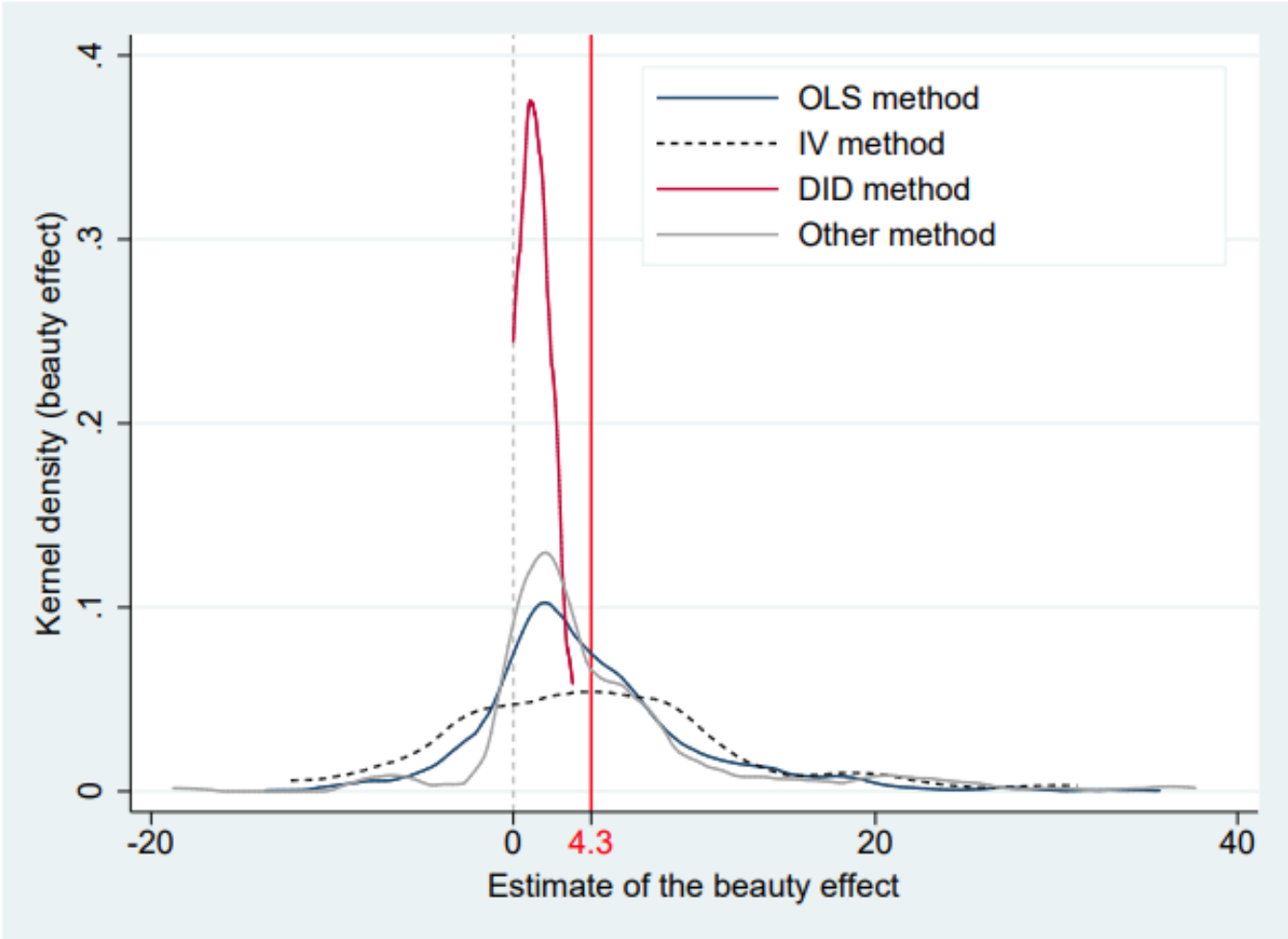
Success measurement doesn't matter

(b) Success measurement



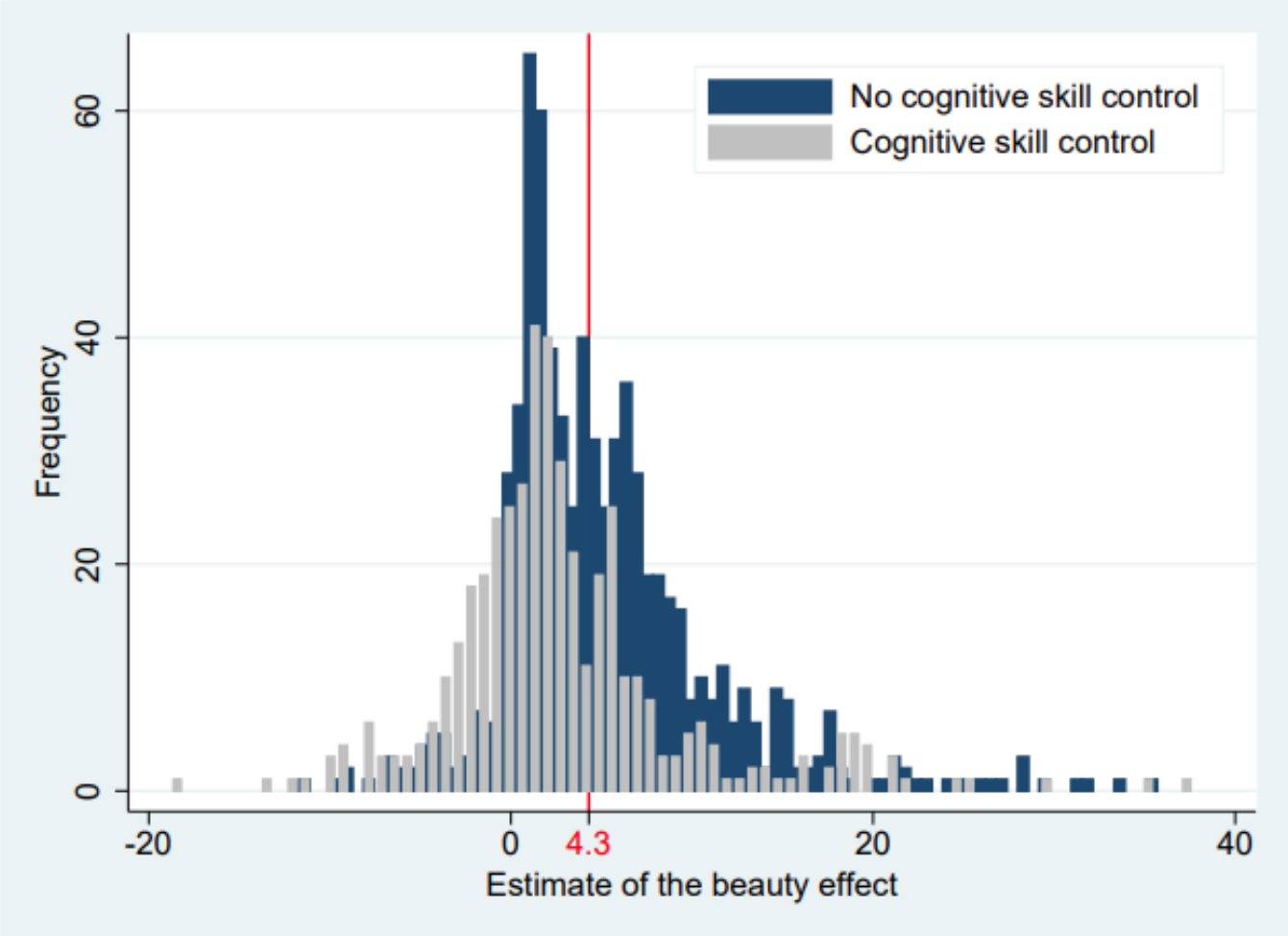
Method doesn't matter

(c) Method choice



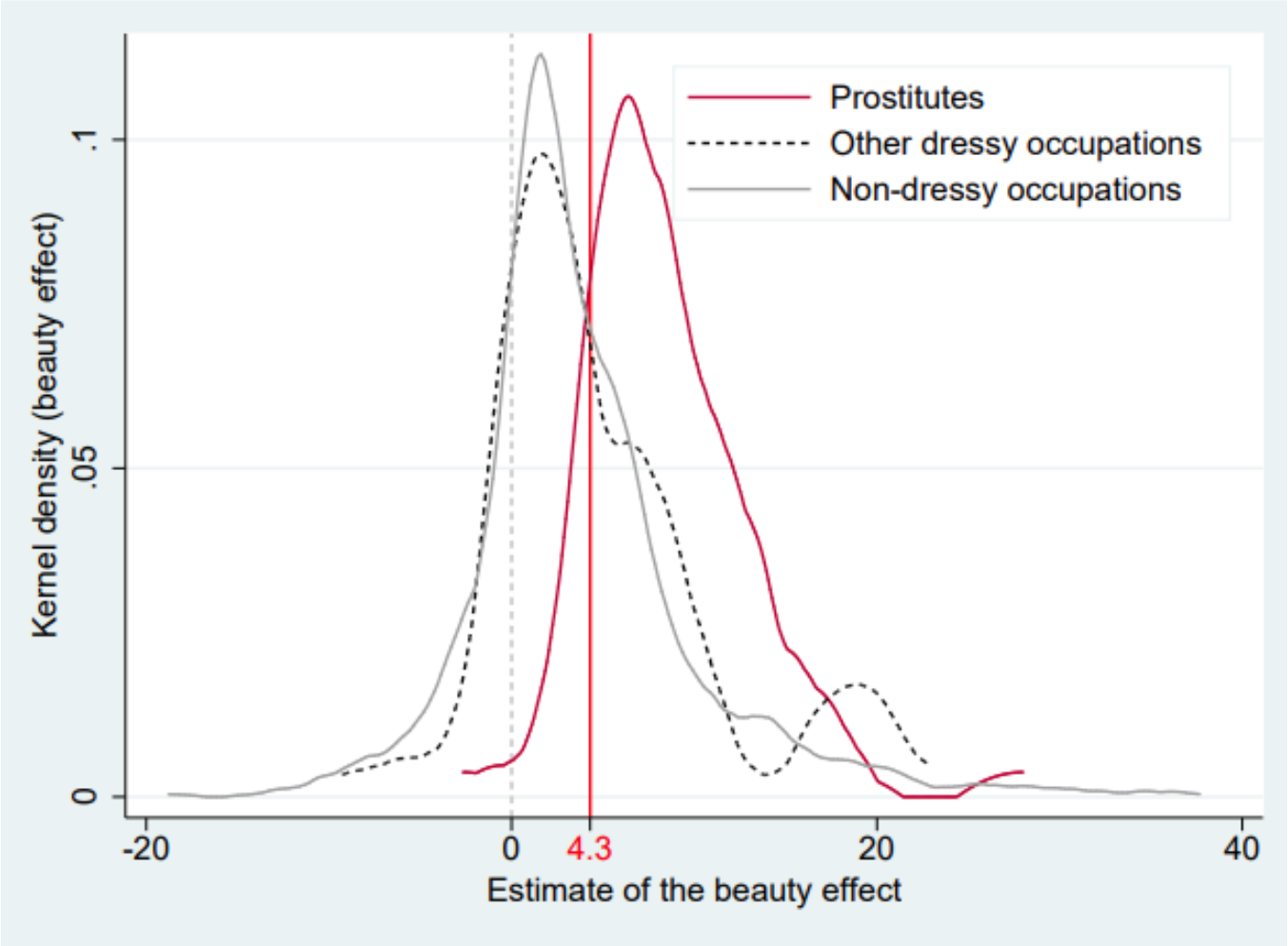
Ability control matters

(g) Ability control

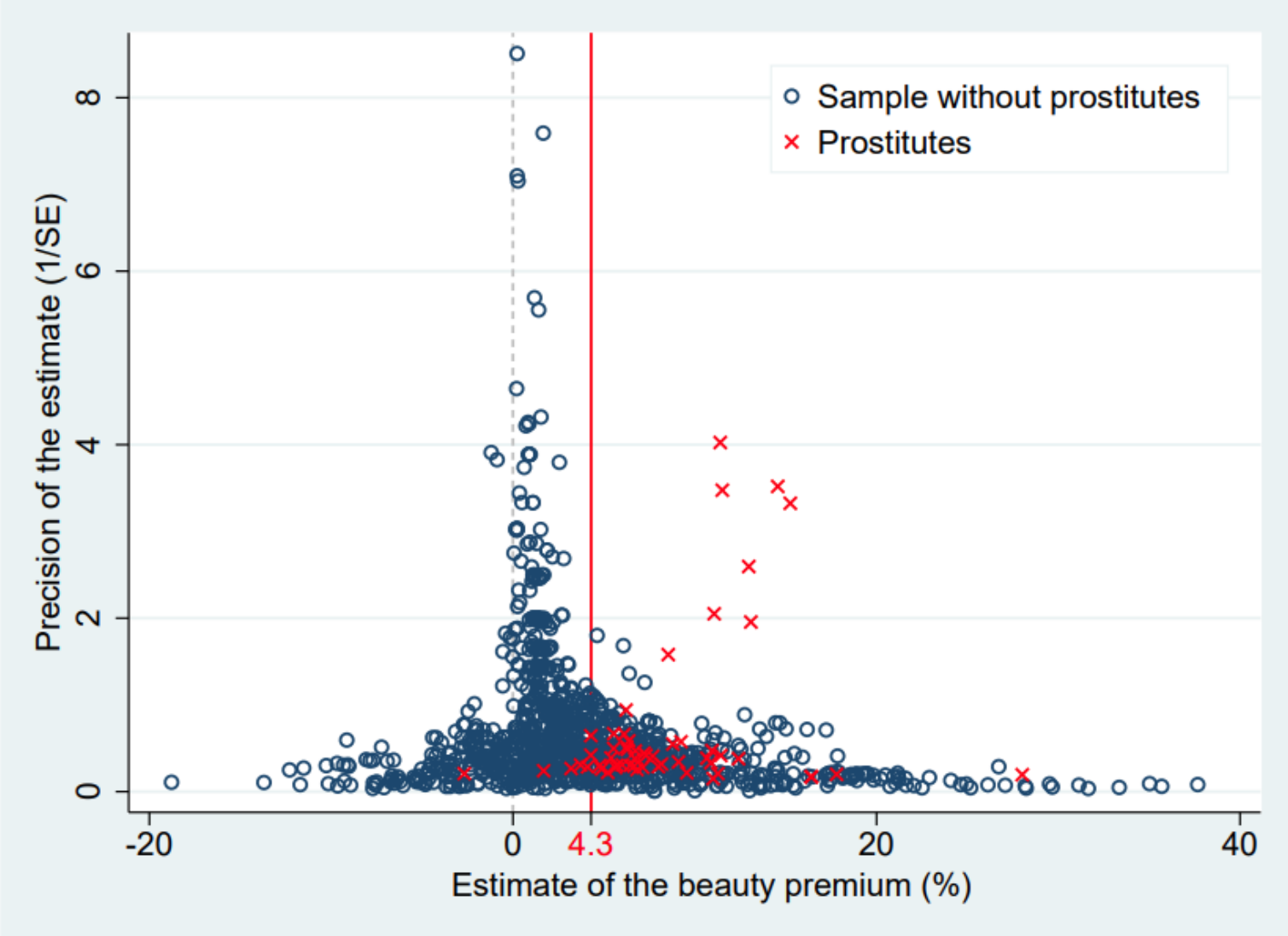


Occupation matters

(d) Occupation



Sex workers stand apart



Option 1: subsamples

Part 1. Subsample of prostitutes

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias (*standard error*)	0.148 (0.776) [-2.956, 2.671]	1.467 (0.782)	-0.756 (1.032)	-1.567 (0.967) {-7.405, -0.514}	-0.468** (0.210) [-3.545, 0.383]
Effect beyond bias (*constant*)	8.136*** (2.767) [1.525, 15.17]	4.488 (2.164)	11.260* (2.686)	12.880*** (1.392) {1.110, 17.173 }	11.760*** (0.414) [5.245, 13.97]
First-stage robust F-stat				15.8	
Observations	55	55	55	55	55
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	-2.154*** (0.745) [-4.057, -0.568]			-2.126 (1.878) [-4.326, -0.755]	P = 1.215 (0.091)
Effect beyond bias	13.290*** (0.298) [-22.03, 16.87]	12.168*** (0.372)	11.205*** (0.905)	12.214*** (0.349) [-22.49, 17.96]	8.490*** (1.691)
Observations	55	55	55	55	55

Option 1: subsamples

Part 2. Sample without prostitutes

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.391***	0.204*	0.800***	0.875*	0.718***
(*standard error*)	(0.117)	(0.119)	(0.161)	(0.450)	(0.271)
	[0.120, 0.642]			{0.118, 2.078}	[0.061, 1.331]
Effect beyond bias	2.581***	3.289***	1.603*	0.741	2.617**
(*constant*)	(0.447)	(0.453)	(0.875)	(1.728)	(1.061)
	[1.660, 3.501]			{0.048, 3.696}	[0.333, 5.113]
First-stage robust F-stat				16.2	
Observations	1,104	1,104	1,104	1,104	1,104
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.610***			1.610***	P = 0.307
	(0.202)			(0.202)	(0.039)
	[1.184, 2.051]			[1.184, 2.051]	
Effect beyond bias	0.152	0.229***	0.008	0.152	0.669***
	(0.118)	(0.07)	(0.798)	(0.118)	(0.231)
	[-0.050, 0.963]			[-0.050, 0.963]	
Observations	1,104	1,104	1,104	1,104	1,104

Option 1: subsamples

Part 1. Subsample of penalties

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.165**	-0.354	0.702***	0.578	0.354***
(*standard error*)	(0.0678)	(0.235)	(0.168)	(1.285)	(0.131)
	[0.021, 0.406]			{-1.941, 3.096}	[0.028, 1.395]
Effect beyond bias	2.647***	4.801***	0.447	0.933	2.655**
(*constant*)	(0.733)	(0.977)	(0.877)	(5.565)	(1.153)
	[1.225, 4.378]			{-3.133, 4.998}	[0.038, 5.782]
First-stage F-stat				0.8	
Observations	205	205	205	205	205
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.097***			1.097***	P = 0.369
	(0.164)			(0.12)	(0.019)
	[0.709, 1.489]			[0.709, 1.489]	
Effect beyond bias	0.139***	0.244***	0.245	0.139	0.236***
	(0.0531)	(0.021)	(0.323)	(0.097)	(0.067)
	[-0.405, 1.425]			[-0.405, 1.425]	
Observations	205	205	205	205	205

Option 1: subsamples

Part 2. Sample without penalties

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.437***	0.282*	0.745***	0.772*	0.666**
(*standard error*)	(0.139)	(0.160)	(0.172)	(0.440)	(0.281)
	[0.125, 0.745]			{0.032, 1.949}	[-0.044, 1.295]
Effect beyond bias	2.881***	3.448***	2.228**	1.652	3.470***
(*constant*)	(0.555)	(0.585)	(0.914)	(1.580)	(1.202)
	[1.745, 4.058]			{0.068, 4.171}	[0.845, 6.305]
First-stage robust F-stat				17.3	
Observations	954	954	954	954	954
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.881***			1.881***	P = 0.300
	(0.246)			(0.246)	(0.037)
	[1.351, 2.378]			[1.351, 2.378]	
Effect beyond bias	0.346	0.38*	0.013	0.346	0.200
	(0.294)	(0.229)	(1.323)	(0.294)	(0.828)
	[-0.051, 2.285]			[-0.051, 2.285]	
Observations	954	954	954	954	954

Option 2: Bayesian model averaging

- We want to regress estimates of the beauty effect on variables reflecting heterogeneity
- But too many variables; model uncertainty
- Solution: run regressions with different combinations of the variables
- Give more weight to those that fit the data well and are parsimonious

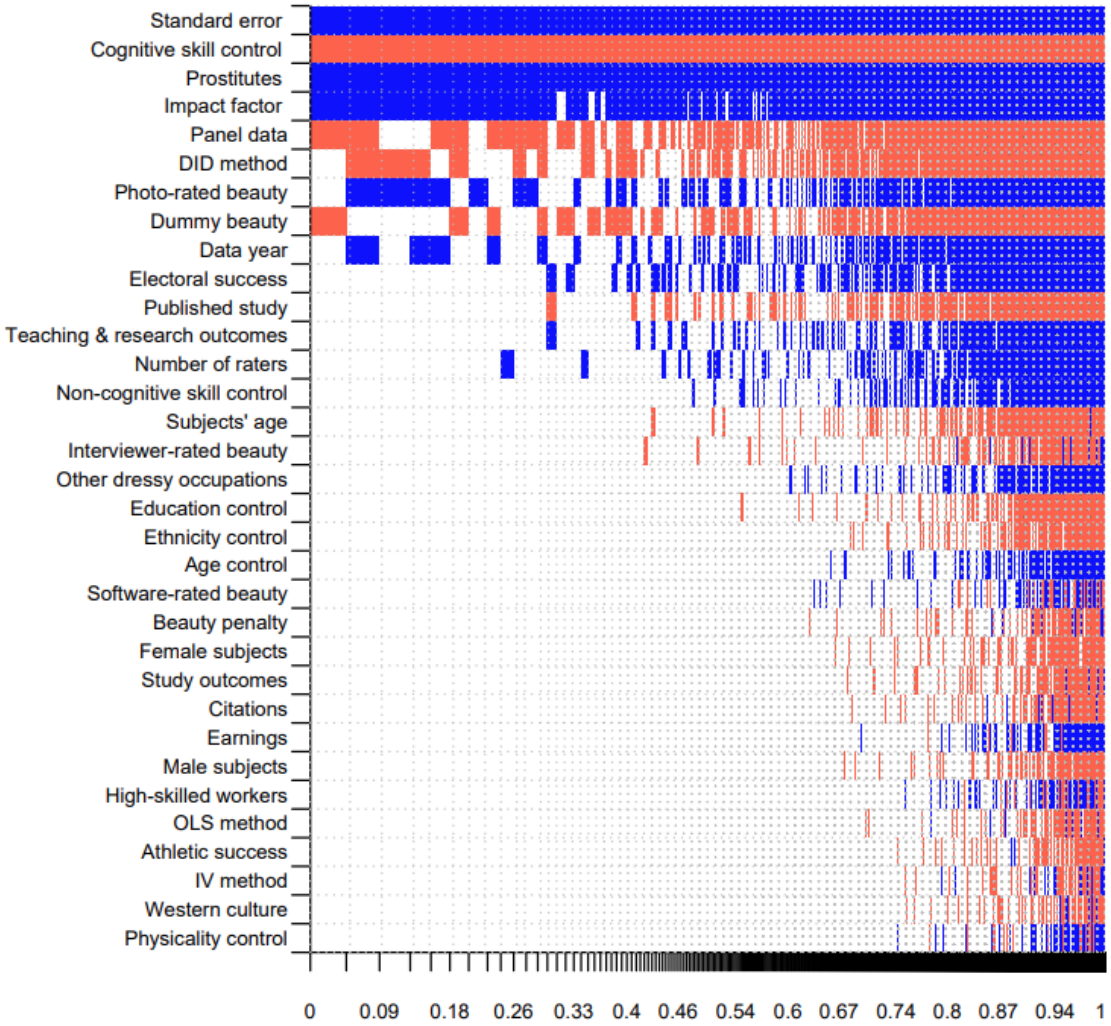
Option 2: Bayesian model averaging

- Model averaging can be frequentist but computationally difficult
- Intuition: run different combinations and weight by adjusted R^2 or information criteria
- In the Bayesian setting we can use the MCMC algorithm
- G-prior: all coefficients are zero. Weight? UIP = the prior has the same weight as one observation
- Model prior: uniform = all models have the same weight
- Dilution prior: models with collinearity get less weight

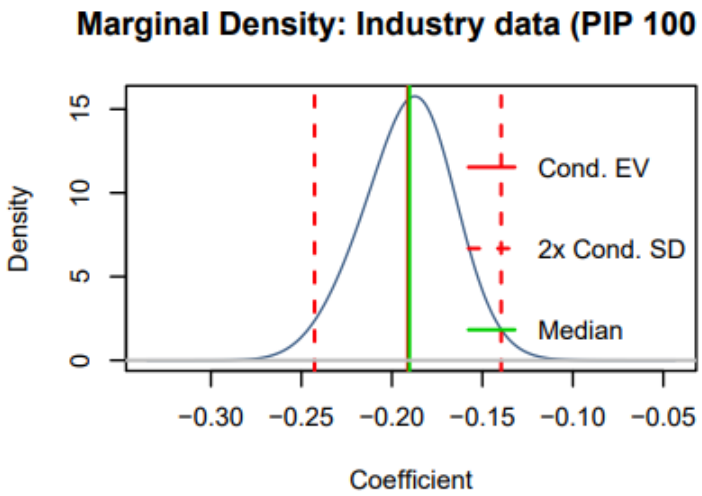
Hands on

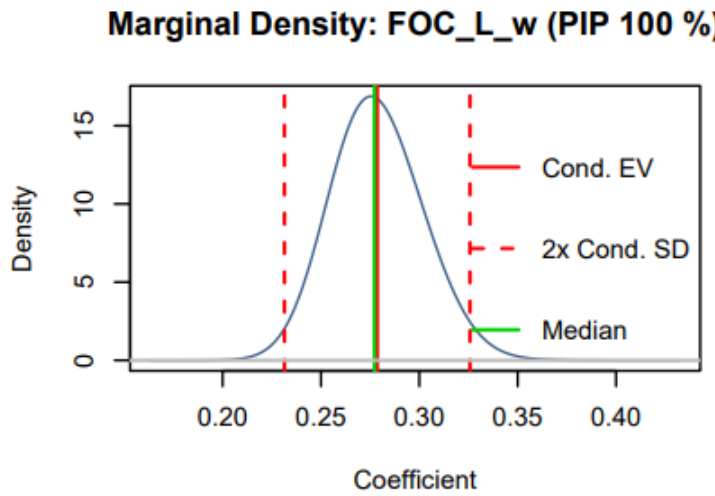
Open R

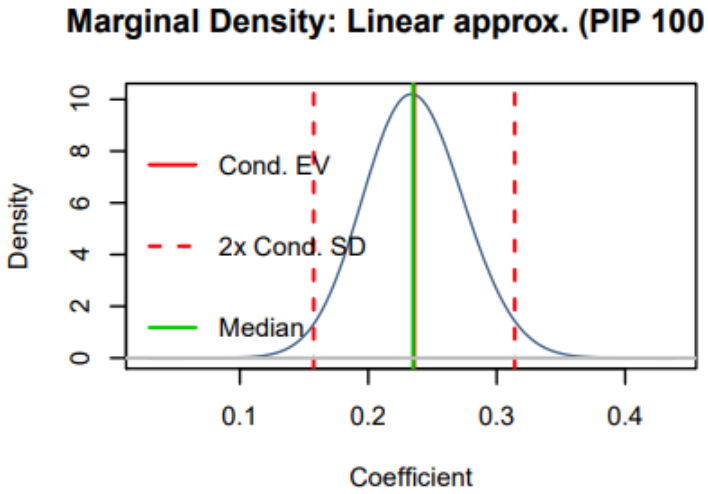
Model inclusion in BMA

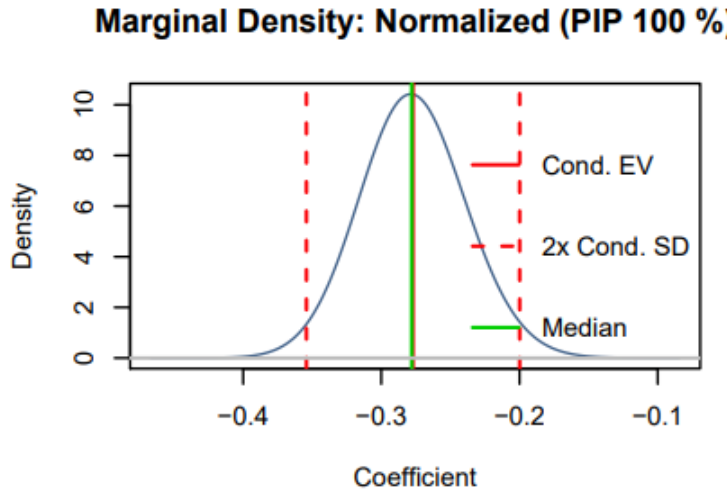


Posterior densities









Regression results

Table 5: Why reported beauty premiums vary

Response variable: Beauty premium	Bayesian model averaging (baseline model)			Stepwise regression (frequentist check)		
	P. mean	P. SD	PIP	Mean	SE	p-value
Constant	2.759	NA	1.000	3.582	0.562	0.000
Standard error	0.435	0.042	1.000	0.426	0.112	0.000
Measurement of beauty						
Interviewer-rated beauty	-0.033	0.204	0.036			
Photo-rated beauty	0.591	0.745	0.424			
Software-rated beauty	0.009	0.149	0.014			
Dummy beauty	-0.485	0.664	0.387			
Beauty penalty	-0.007	0.089	0.012			
Number of raters	0.051	0.140	0.134			
Measurement of success						
Earnings	0.006	0.098	0.010			
Study outcomes	-0.006	0.114	0.011			
Teaching & research	0.238	0.645	0.141			
Athletic success	-0.004	0.104	0.007			
Electoral success	0.670	1.361	0.224			
Data characteristics						
Male subjects	-0.005	0.064	0.010			
Female subjects	-0.006	0.069	0.012			
Subjects' age	-0.075	0.332	0.061			
High-skilled workers	0.002	0.064	0.009			
Prostitutes	4.793	1.058	0.999	4.386	1.199	0.001
Other dressy occupations	0.035	0.227	0.032			
Western culture	-0.001	0.039	0.006			
Panel data	-1.131	1.016	0.606			
Data year	0.337	0.504	0.346			

Best practice (implied estimate)

- From BMA results we know how estimates depend on study design
- Some study designs are better: identification, data, generally quality
- Select best practice based on the literature (quasi-experimental studies, etc.)
- Compute fitted values and confidence intervals (lincom command in Stata)

Hands on

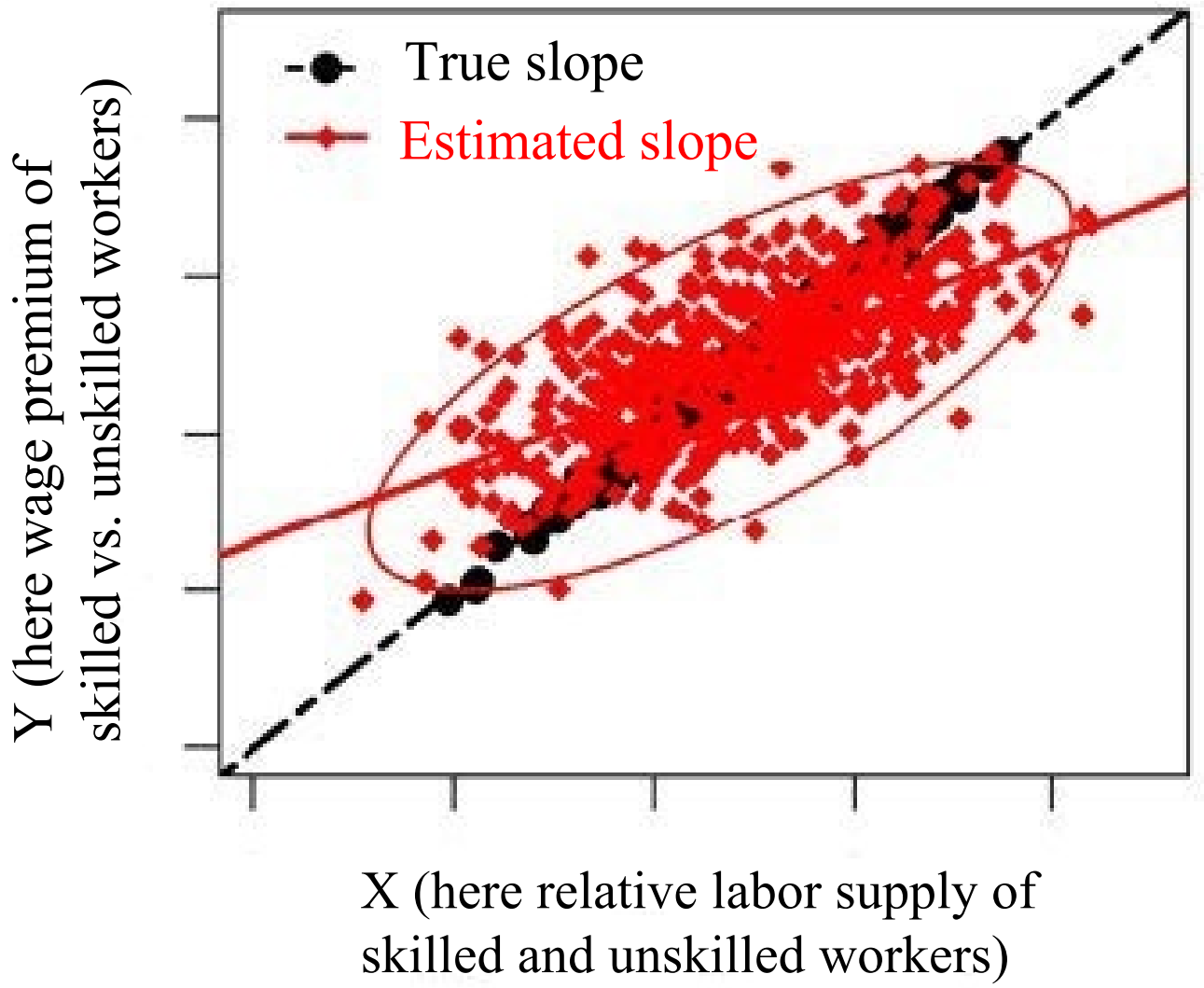
Open Stata

Best practice (class size effect)

	Effect	95% cred. int.	
STAR experiment	-1.50	-3.04	0.04
Regression discontinuity	-0.18	-1.36	1.00
Instrumental variable	-0.42	-1.59	0.75
Fixed effects	0.01	-1.17	1.19
OLS	0.01	-1.19	1.21
Kindergarten	0.18	-1.10	1.46
Primary school	-0.46	-1.65	0.72
Secondary school	0.18	-0.99	1.35
Disadvantaged students	-0.19	-1.42	1.03
USA	-0.21	-1.30	0.88
Scandinavian countries	-0.21	-1.47	1.05
Other countries	-0.21	-1.43	1.02
Math	-0.21	-1.38	0.97
Reading	-0.21	-1.39	0.96
Writing	-0.20	-1.50	1.11
Languages	-0.21	-1.40	0.99
Other subject	-0.21	-1.43	1.01
Class size = 15	-0.99	-2.18	0.21
Class size = 20	-0.45	-1.62	0.72
Class size = 25	-0.18	-1.35	1.00
Class size = 30	0.01	-1.18	1.19
Class size = 35	0.15	-1.05	1.34
Overall corrected mean	-0.21	-1.38	0.97

Workshop Outline

Issue 1: Attenuation bias, aka regression dilution



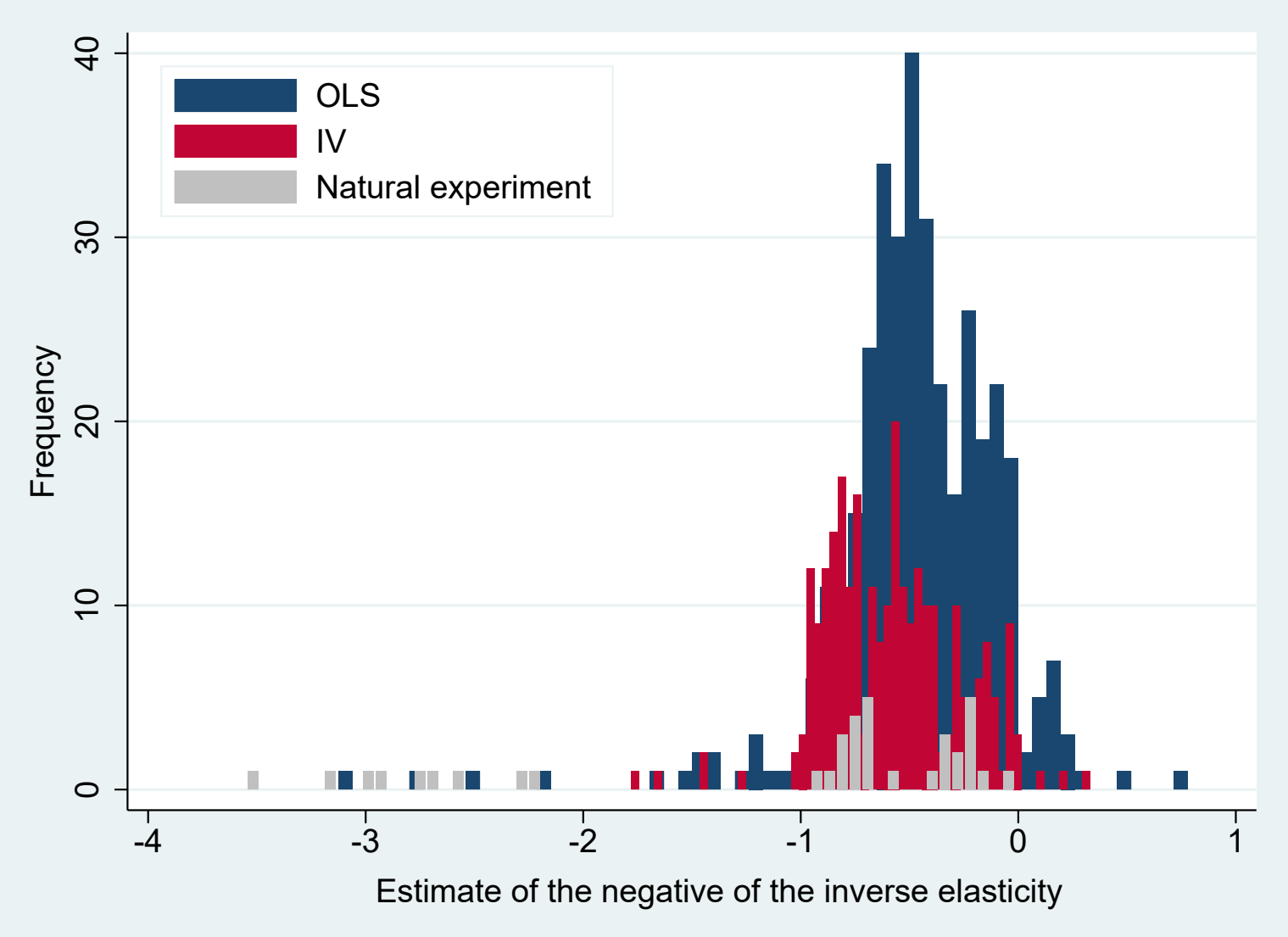
Elasticity of skill substitution

Important complication

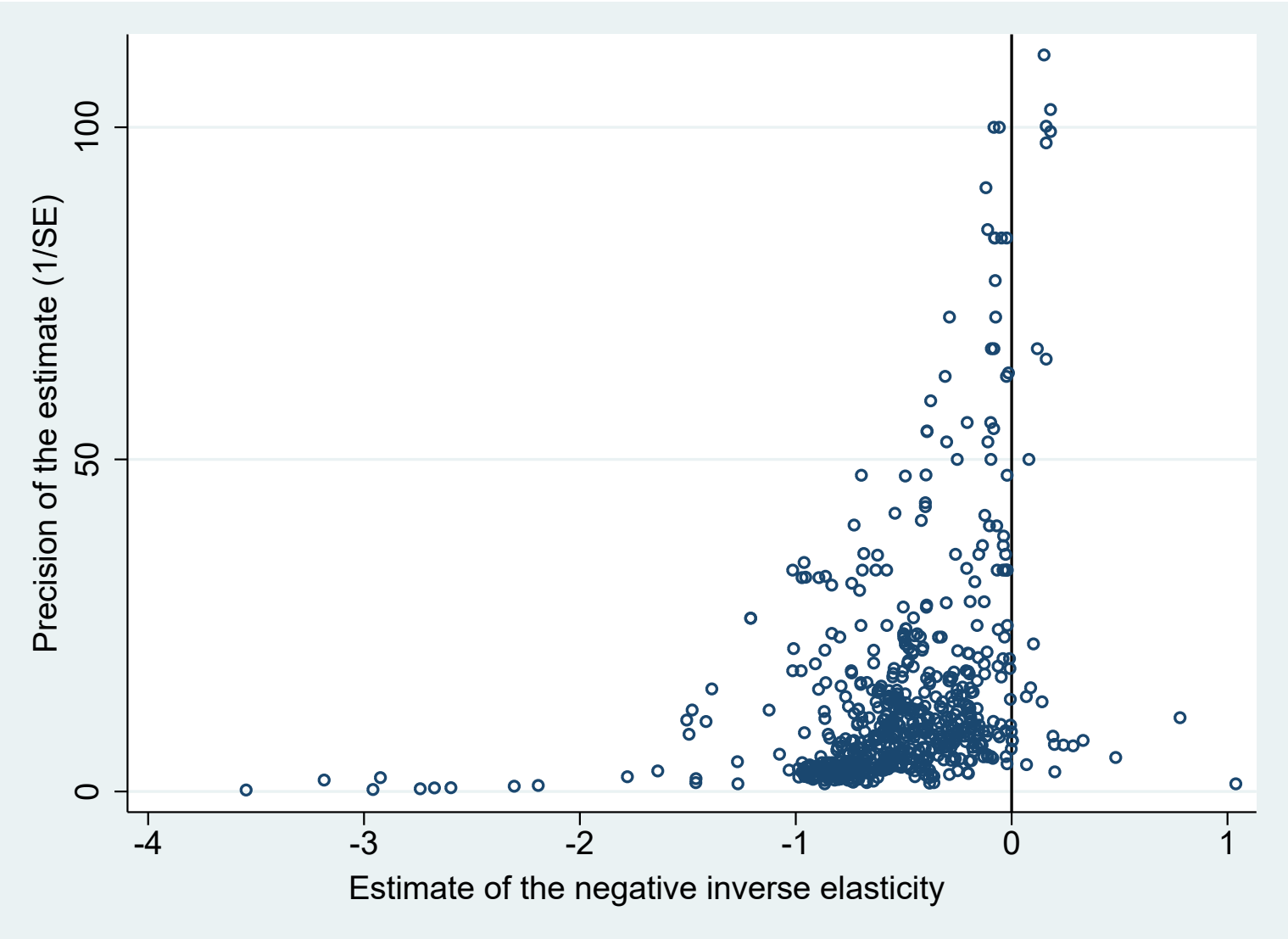
Researchers estimate negative inverse elasticity: they regress skill premium on relative labor supply (not vice versa!)

- Measures substitutability between skilled (college) and unskilled (high school) workers
- Important for wage inequality, explaining cross-country differences in productivity, ...
- 682 estimates from 77 studies, "consensus" elasticity = 1.5 (negative inverse = -0.67)

Three estimation frameworks



Strong publication bias (or *p*-hacking?)



OLS: no effect (but attenuation & endogeneity biases)

Panel A: OLS estimates

	FE	BE	IV	EK	AK
Publication bias	-5.804***	-4.277***	-6.962***	-5.465***	P=0.468
	(1.999)	(1.266)	(1.694)	(0.540)	(0.139)
			[-11.770, -2.494]		
			{-11.972, -3.133}		
Effect beyond bias	-0.0207	-0.0965	0.0103	-0.0361**	-0.289**
	(0.103)	(0.0627)	(0.104)	(0.0191)	(0.113)
			[-0.331, 0.214]		
First-stage robust F-stat			46.17		
Observations	347	347	251	347	347

IV: strong effect (no biases beyond publication)

Panel B: IV estimates

	FE	BE	IV	EK	AK
Publication bias	-2.287**	-0.923	-0.553	-1.485***	P=0.336
	(0.843)	(1.365)	(0.681)	(0.268)	(0.093)
			[-1.913, 1.078]		
			{-1.991, 0.748}		
Effect beyond bias	-0.149	-0.297**	-0.400***	-0.252***	-0.333***
	(0.109)	(0.115)	(0.114)	(0.0246)	(0.058)
			[-0.719, 0.175]		
First-stage robust F-stat			69.98		
Observations	264	264	212	264	264

Natural experiments: no effect (but attenuation bias)

Panel C: Natural experiment estimates

	FE	BE	IV	EK	AK
Publication bias	-3.557***	-1.874*	-3.176***	-3.115***	P=0.187
	(0.0178)	(0.682)	(0.853)	(0.343)	(0.075)
			[-4.854, -1.407]		
			{-4.653, -1.444}		
Effect beyond bias	0.0496***	-0.121	-0.00307	0.00302	-0.009
	(0.00246)	(0.0824)	(0.0297)	(0.0280)	(0.066)
			[NA, NA]		
First-stage robust F-stat			260.41		
Observations	40	40	40	40	40

Publication bias $>$ attenuation bias

- $|\text{IV}| > |\text{OLS}| \rightarrow$ attenuation or other endogeneity biases important
- OLS = Natural experiments $\rightarrow$ other endogeneity biases negligible

$\rightarrow$ $|\text{IV} - \text{OLS}|$ measures attenuation bias, which is strong

- Uncorrected inverted elasticity = -0.67
- Inverted elasticity corrected for pub and attenuation bias (IV) = -0.25

$\rightarrow$ Pub bias dominates attenuation

$\rightarrow$ Implied elasticity = 4

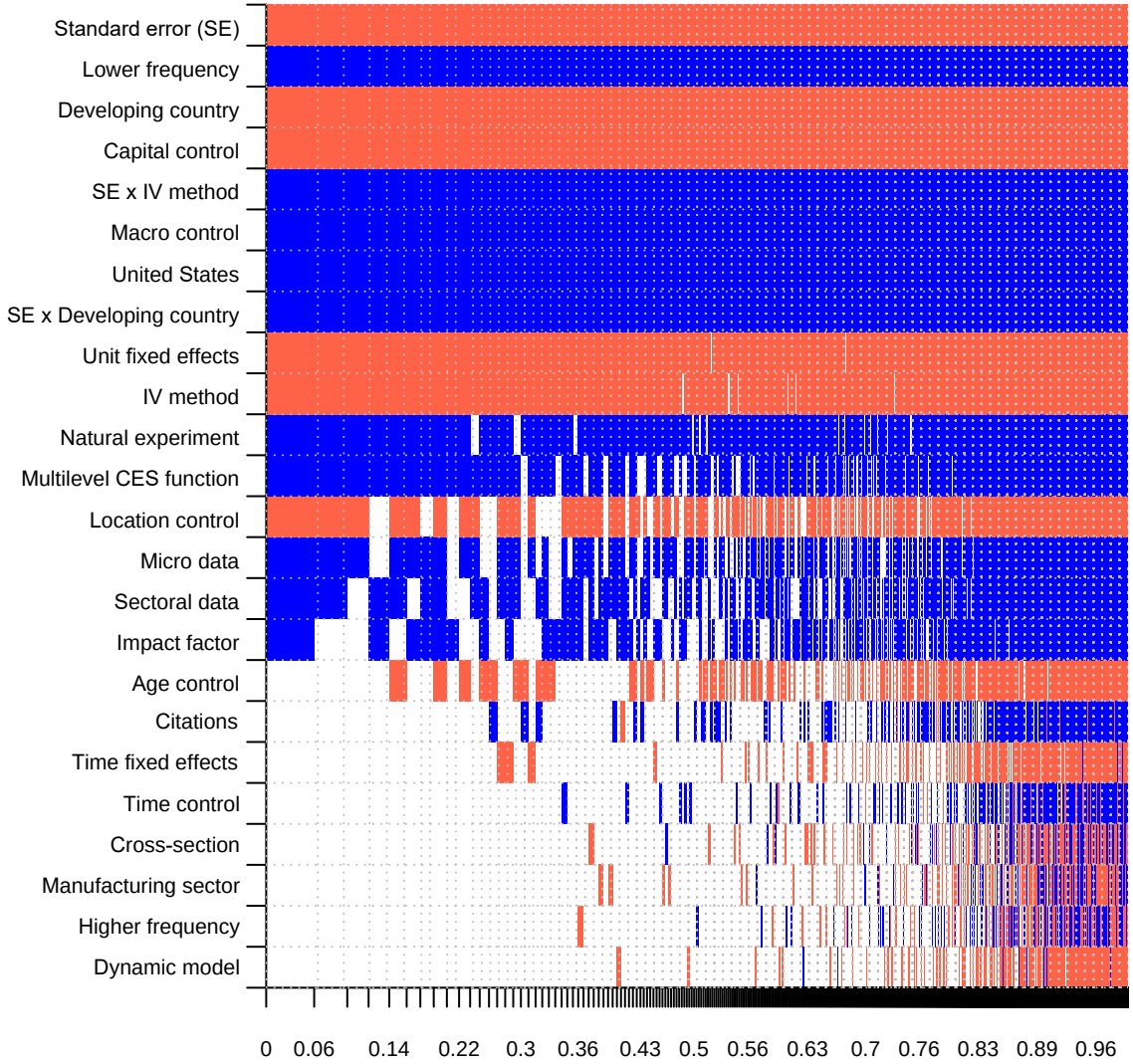
Assumptions of Andrews & Kasy not satisfied

Table C7: Specification test for the Andrews & Kasy (2019) model

	All inverse	OLS method	IV method	Natural experiment	Developed country
Correlation	0.515 [0.436, 0.605]	0.265 [0.160, 0.392]	0.635 [0.499, 0.743]	0.947 [0.928, 0.976]	0.558 [0.462, 0.667]
Observations	654	347	264	40	418
	Developing country	Country estimate	Region estimate	One-level CES function	Multilevel CES function
Correlation	0.330 [0.159, 0.482]	0.856 [0.775, 0.95]	0.476 [0.384, 0.568]	0.199 [0.047, 0.342]	0.592 [0.498, 0.709]
Observations	151	555	93	198	444

Notes: Following Kranz & Putz (2022), the table shows, for various subsets of the literature, the correlation coefficient between the logarithm of the absolute value of the estimated inverse elasticity and the logarithm of the corresponding standard error, weighted by the inverse publication probability estimated by the Andrews & Kasy (2019) model. If the assumptions of the model hold, the correlation is zero. Bootstrapped 95% confidence interval in parentheses.

Model uncertainty; biases confirmed



Best practice estimate around 4

Table 5: Implied elasticities

	Subjective best practice	Autor (2014)	Card (2009)	Carneiro *et al.* (2022)
All countries	-0.27	-0.13	-0.24	0.05
	(-0.48, -0.05)	(-0.24, -0.02)	(-0.39, -0.09)	(-0.12, 0.23)
	$\sigma = 3.7$	$\sigma = 7.7$	$\sigma = 4.2$	$\sigma = -18.4$
USA	-0.16	-0.02	-0.13	0.16
	(-0.38, 0.06)	(-0.12, 0.07)	(-0.28, 0.02)	(-0.02, 0.34)
	$\sigma = 6.3$	$\sigma = 45.0$	$\sigma = 7.8$	$\sigma = -6.2$
Developing countries	-0.47	-0.33	-0.44	-0.15
	(-0.70, -0.24)	(-0.47, -0.19)	(-0.60, -0.27)	(-0.33, 0.04)
	$\sigma = 2.1$	$\sigma = 3.0$	$\sigma = 2.3$	$\sigma = 6.8$

Results

Main Findings

1. Attenuation bias is important ($|\text{IV}| > |\text{OLS}| = |\text{natural experiments}|$).
2. Publication bias is more important (inverse elasticity: simple mean -0.67 ; corrected IV mean -0.25).
3. Implied elasticity around 4 (compared to consensus 1.5).

Project Website

www.meta-analysis.cz/skill

Published in the *Review of Economics and Statistics*, 2024.

Issue 2: How much heterogeneity is too much?

Common measure: estimate the relative heterogeneity of underlying effects *(Higgins and Thompson, 2002)*

$$I^2 \equiv \frac{\text{underlying variance of effects}}{\text{observed variance of estimates}}$$

- $I^2 = 0.75 \sim 1.0$ ... *"considerable heterogeneity" in medicine (Cochrane Review Guideline)*
- $I^2 \simeq 0.9$ ... common in economics meta-analyses (e.g., *Vivalt, 2021*)

Unresolved concern:

- no universally agreed criteria for "too much" heterogeneity
- undesirable pattern: with more precise studies, higher I^2

This paper (still under construction)

1. Propose a new test for external validity in meta-analysis:
   - Based on the tail index of the underlying distribution, without contamination by estimate precision.
   - Test whether the implied variance of underlying effect is infinite (i.e., too heterogeneous).

2. Show that the implied variance is infinite in meta-analysis of 126 nudge RCTs *(DellaVigna and Linos, 2020)*:
   - **Nudge:** "choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives." *(Thaler and Sunstein, 2008)*
   - The possibility of infinite variance can be important in real-world applications.

Test of infinite variance

Pareto tail:

$$F(b) = 1 - Cb^{-\alpha}, \quad b > b_{\min} \quad \text{and} \quad C > 0$$

- e.g. earthquakes, word frequency, city size, wealth distribution
- If $\alpha < 2$, then $Var(b) = \infty$ because

$$\int_{b_{\min}}^{\infty} b^2 f(b)\, db \propto \int_{b_{\min}}^{\infty} b^2 b^{-3}\, db = \int_{b_{\min}}^{\infty} b^{-1}\, db = \log b \Big|_{b_{\min}}^{\infty} = \infty$$

Tail index test of external validity:

- **Criteria 1:** if $\hat{\alpha} > 2$, then pass the test
- **Criteria 2:** if $\mathbb{P}(\hat{\alpha} \leq 2) < p$, then pass the test

Estimation methods (i) inner problem of tail index

Following *Hill (1975)*, use Maximum Likelihood Estimation for $\hat{\alpha}_{MLE}$: given $b_{\min}$,

$$\hat{\alpha}_{MLE} = \arg\max_{\alpha} \sum_{i=1}^{n} \ln l\left(\alpha \mid \beta_i, SE_i, b_{\min}\right),$$

where

$$l\left(\alpha \mid \beta_i, SE_i, b_{\min}\right) \equiv \int_{\theta_{\min}}^{\infty} \frac{\phi\left(\frac{\beta_i - \theta}{SE_i}\right)}{1 - \Phi\left(\frac{b_{\min} - \theta}{SE_i}\right)} \alpha b_{\min}^{\alpha} b^{-(\alpha+1)}\, db.$$

- Numerical integration

Estimation methods (ii) outer problem of cut-off

Following *Clauset et al. (2006)*, use the goodness-of-fit metric to choose $b^*_{\min}$:

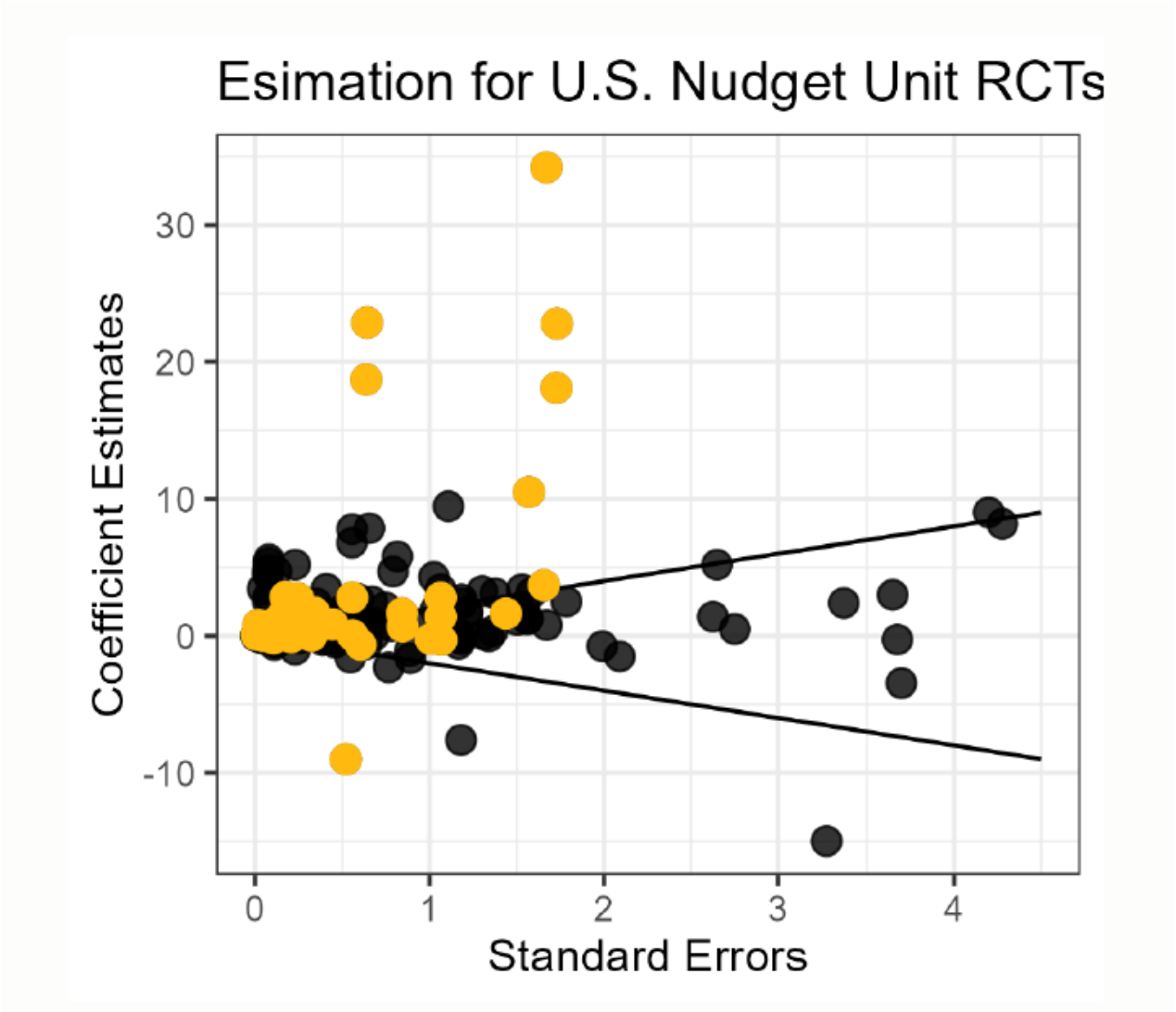
$$b^*_{\min} \in \arg\min_{b_{\min}} \tilde{KS}\left(b_{\min} \mid \hat{\alpha}_{MLE}\right),$$

where the reweighted Kolmogorov-Smirnov statistic is

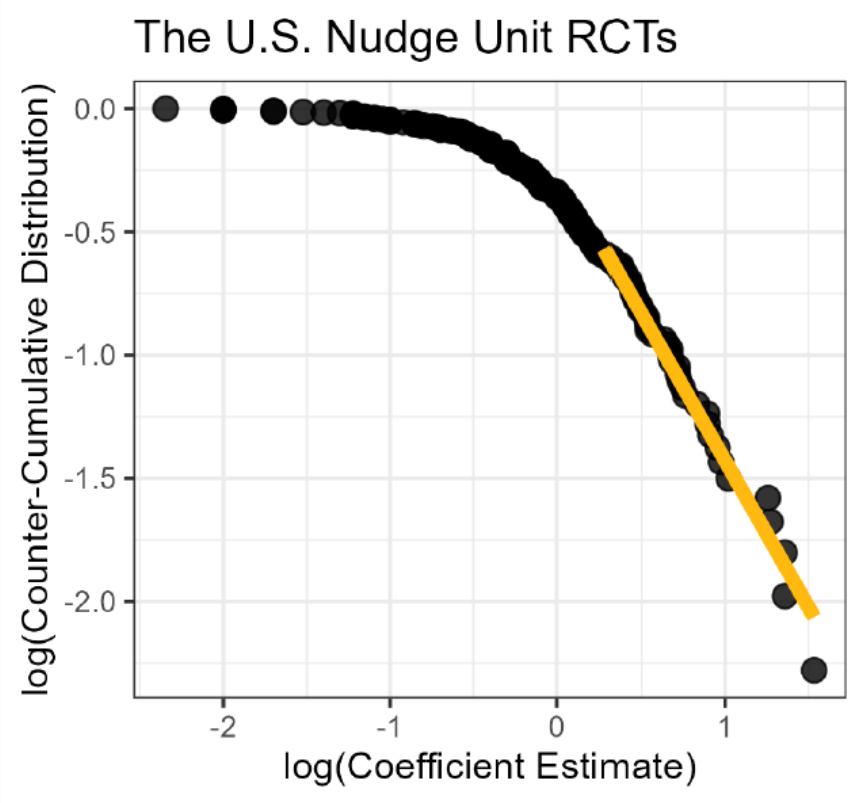
$$\tilde{KS}\left(b_{\min} \mid \hat{\alpha}_{MLE}\right) \equiv \sup_{\beta} \frac{\left|G(\beta \mid \beta \geq b_{\min}) - \hat{G}(\beta \mid \cdot)\right|}{\sqrt{\hat{G}(\beta \mid \cdot)\left[1 - \hat{G}(\beta \mid \cdot)\right]}},$$

and the estimated distribution is $\hat{G}(\beta \mid \hat{\alpha}_{MLE}, \beta \geq b_{\min})$ and the empirical distribution is $G(\beta \mid \beta \geq b_{\min})$.

Nudges


Esimation for U.S. Nudget Unit RCTs
Coefficient Estimates
30
20
10
0
-10
0
1
2
3
4
Standard Errors

Nudges



$$\log\left(1 - G\left(\beta\right)\right)_i = -\underbrace{1.19}_{=\hat{\alpha}} \log \beta_i - \log \underbrace{1.90}_{=\hat{b}_{\min}}$$

Summary

1. Proposed the concept of *"infinite variance"* as representing the persistent criticism that *"too much" heterogeneity* underlies the overall mean estimate.

2. Developed an estimation method by extending the maximum likelihood estimation and goodness-of-fit statistics commonly used in tail estimation.

3. Applied the estimation method and its implied test to show that the influential nudge meta-analysis exhibits the infinite variance.

Workshop Outline

Summary: How to do a modern meta-analysis

1. Search for primary studies using Google Scholar, snowballing, AI agent
2. Use comparable estimates (elasticities, mean differences, correlations)
3. Code standard errors, sample sizes, main differences in data and methods
4. Winsorize estimates and standard errors at the 1% level
5. Correct for publication bias and p-hacking (MAIVE; if weak instrument $\rightarrow$ RTMA)
6. Examine heterogeneity using BMA with the dilution prior
7. Report implied effects for different contexts

Thank you for inviting me

Prepared for Tillväxtanalys. Published at meta-analysis.cz/teaching under CC BY 4.0.