TSE M1 Applied Econometrics (2021) Credibility and Replicability in Economics, Issues and Solutions

Victor Gay

Credibility and Replicability in Economics
Issues and Solutions

Victor Gay

University of Toulouse 1 Capitole

Toulouse School of Economics

Institute for Advanced Study in Toulouse

Fall 2021

Plan of the Session

- The credibility revolution.
- The replicability crisis.
- Replicability in practice: workflow, coding, and reporting principles.

The credibility revolution

The Credibility Revolution

- Trends toward empirical research. Angrist et al. (2017)
- Trends toward design-based methods. Angrist and Pischke (2010)

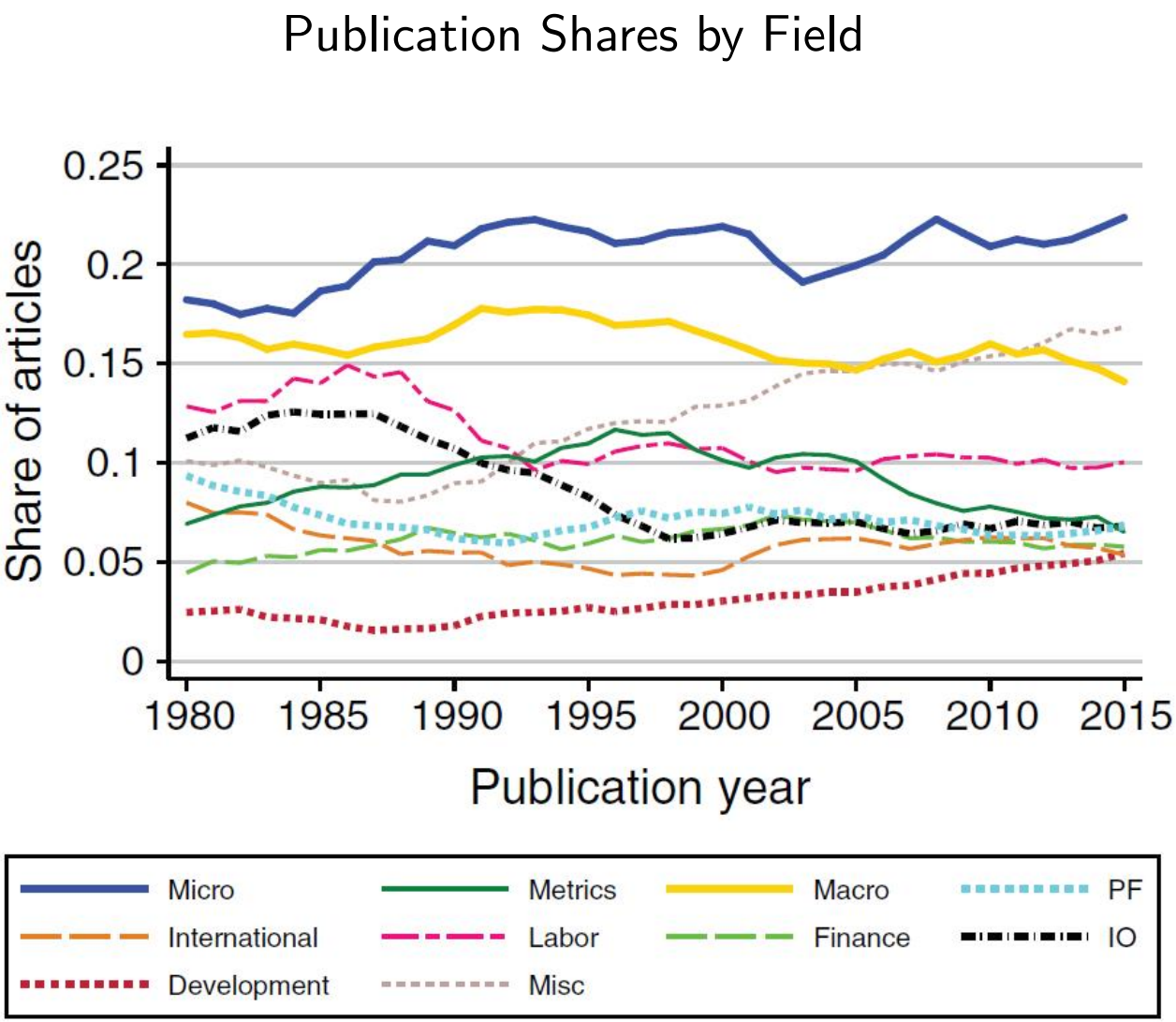
Trends toward empirical research

Trends Toward Empirical Research

Angrist et al. (2017)

- Analysis of 135k papers in 80 journals, 1980–2015.
- Slight trends toward microeconomics and non-traditional fields.
- Large empirical shift within fields.
- Hamermesh (2013) documents similar trends.

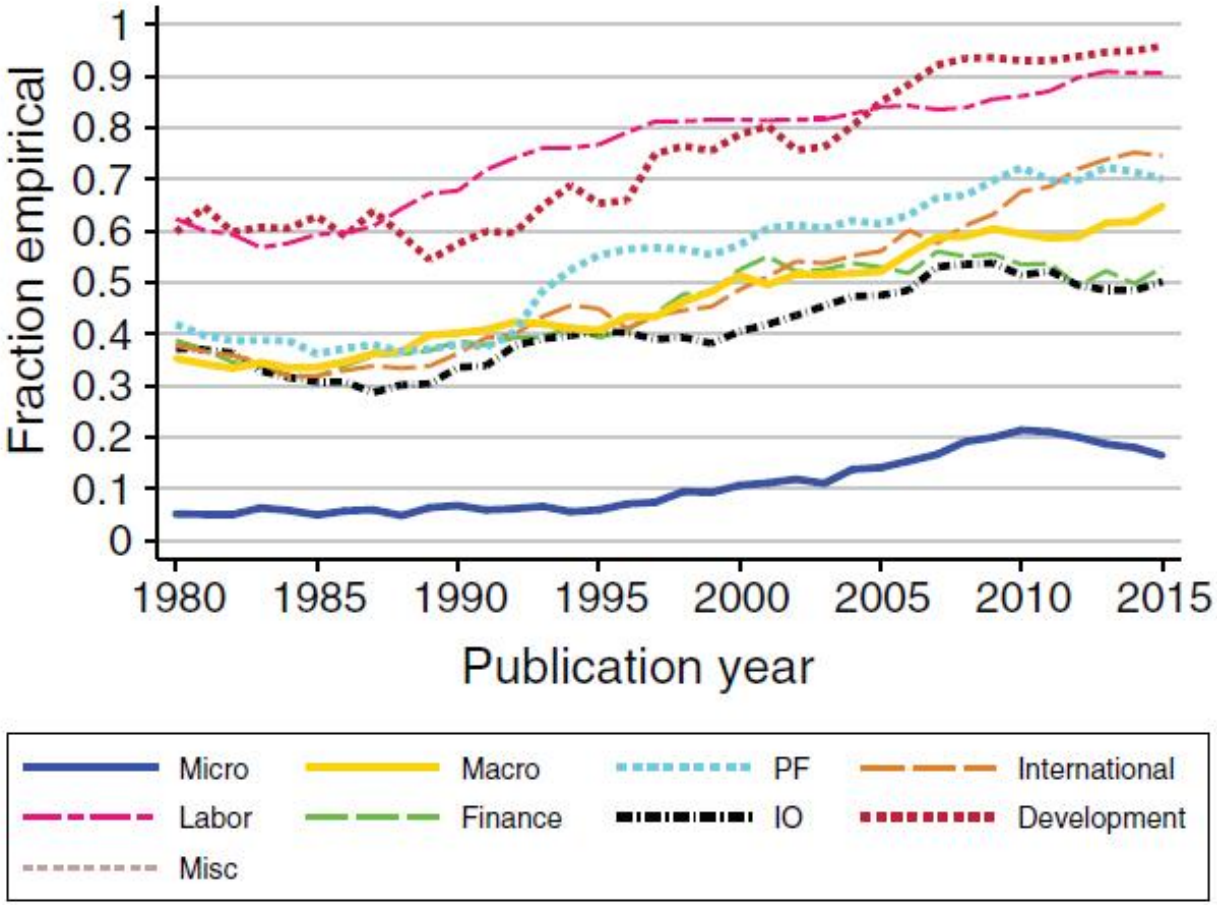
Trends Toward Empirical Research



Source: Angrist et al. (2017).

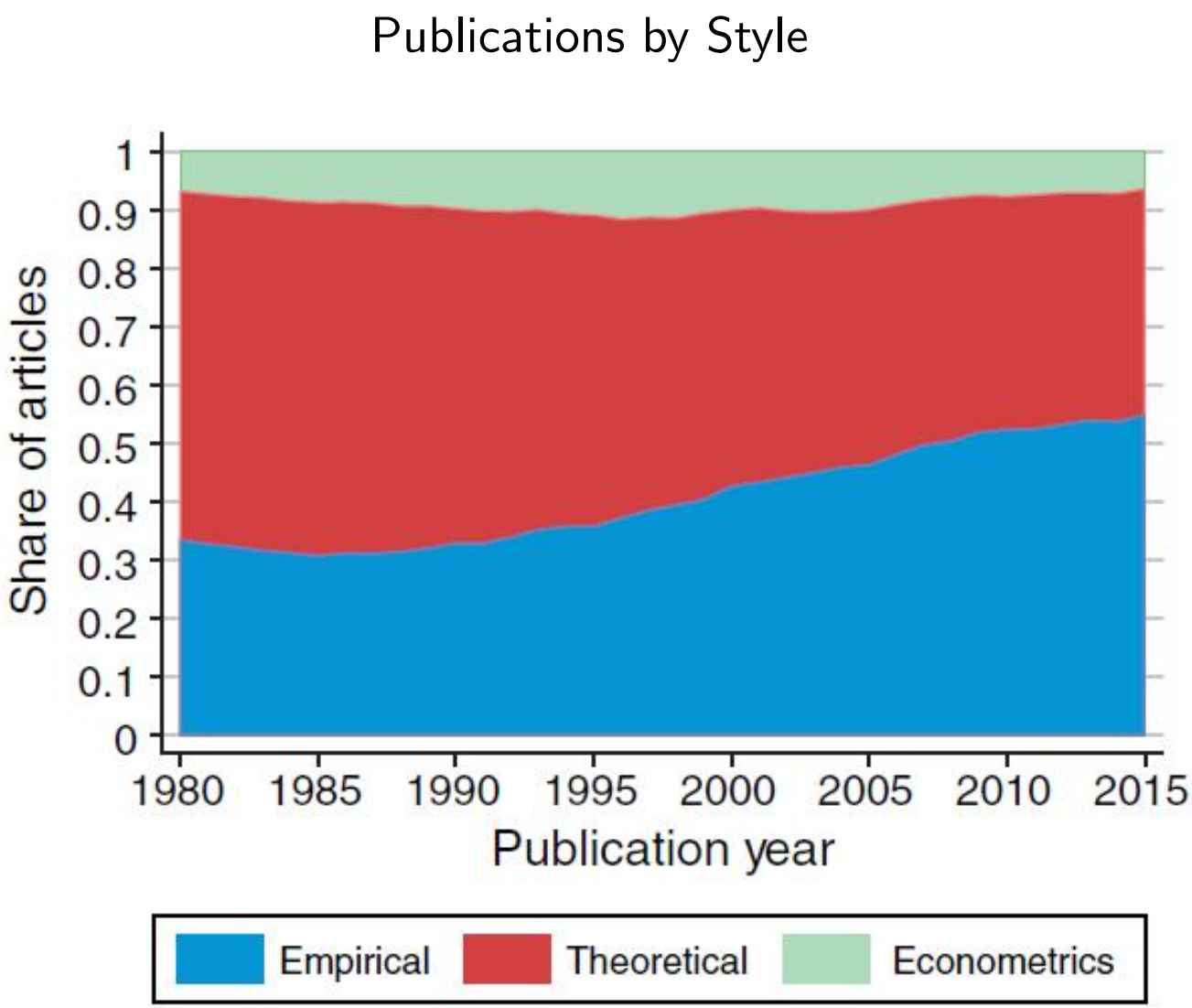
Trends Toward Empirical Research

Fraction Empirical by Field



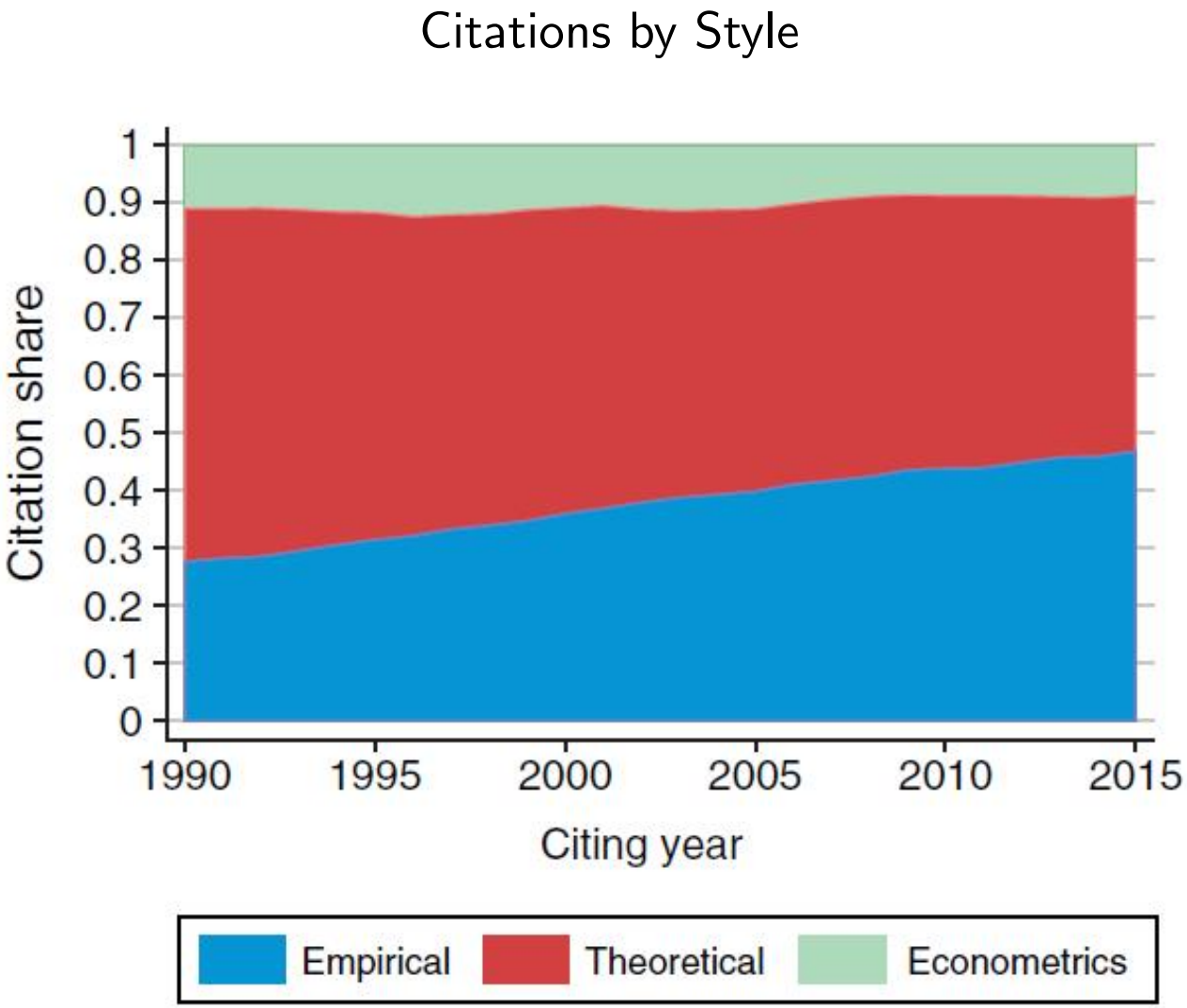
Source: Angrist et al. (2017).

Trends Toward Empirical Research



Source: Angrist et al. (2017).

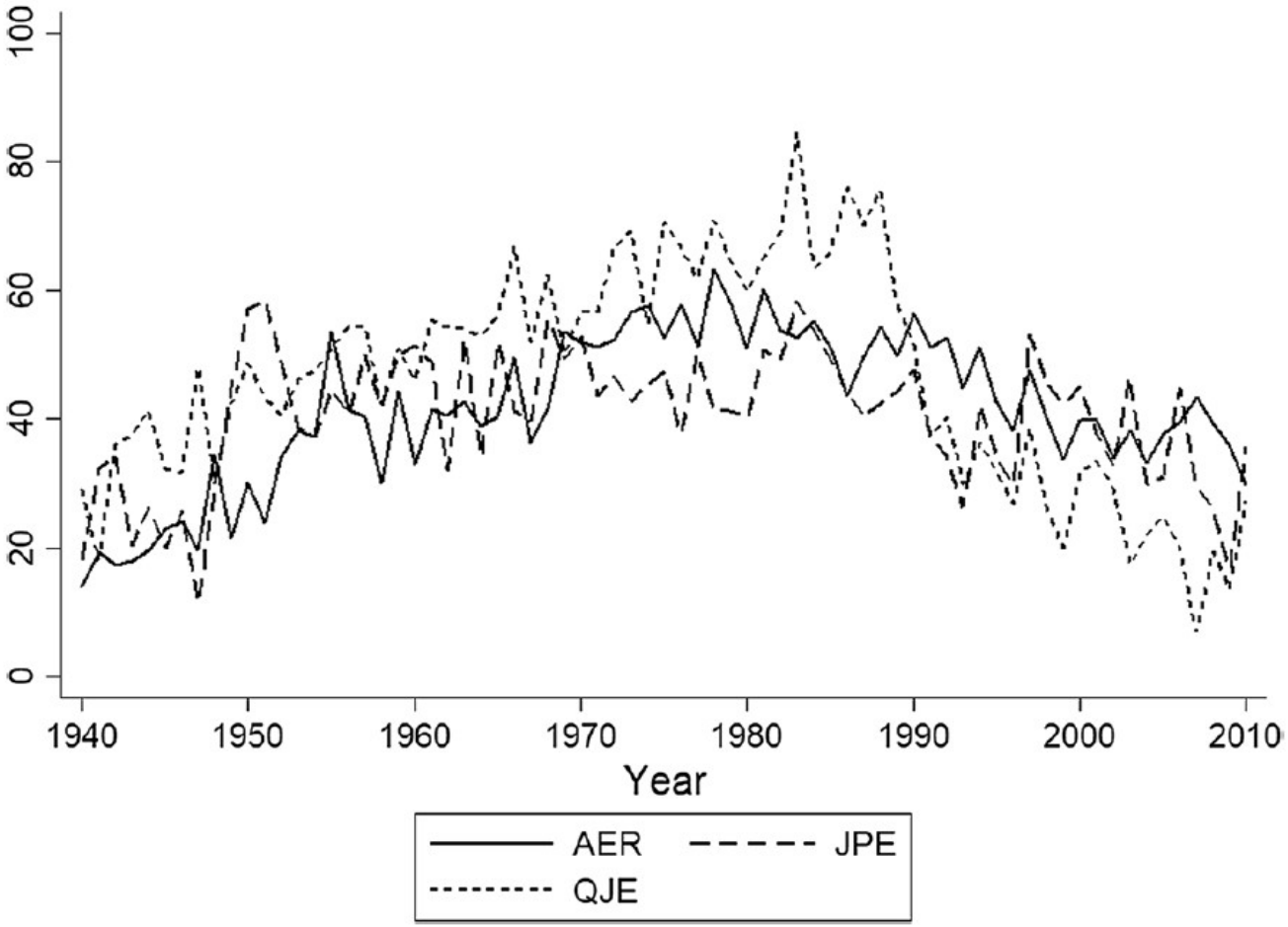
Trends Toward Empirical Research



Source: Angrist et al. (2017).

Trends Toward Empirical Research

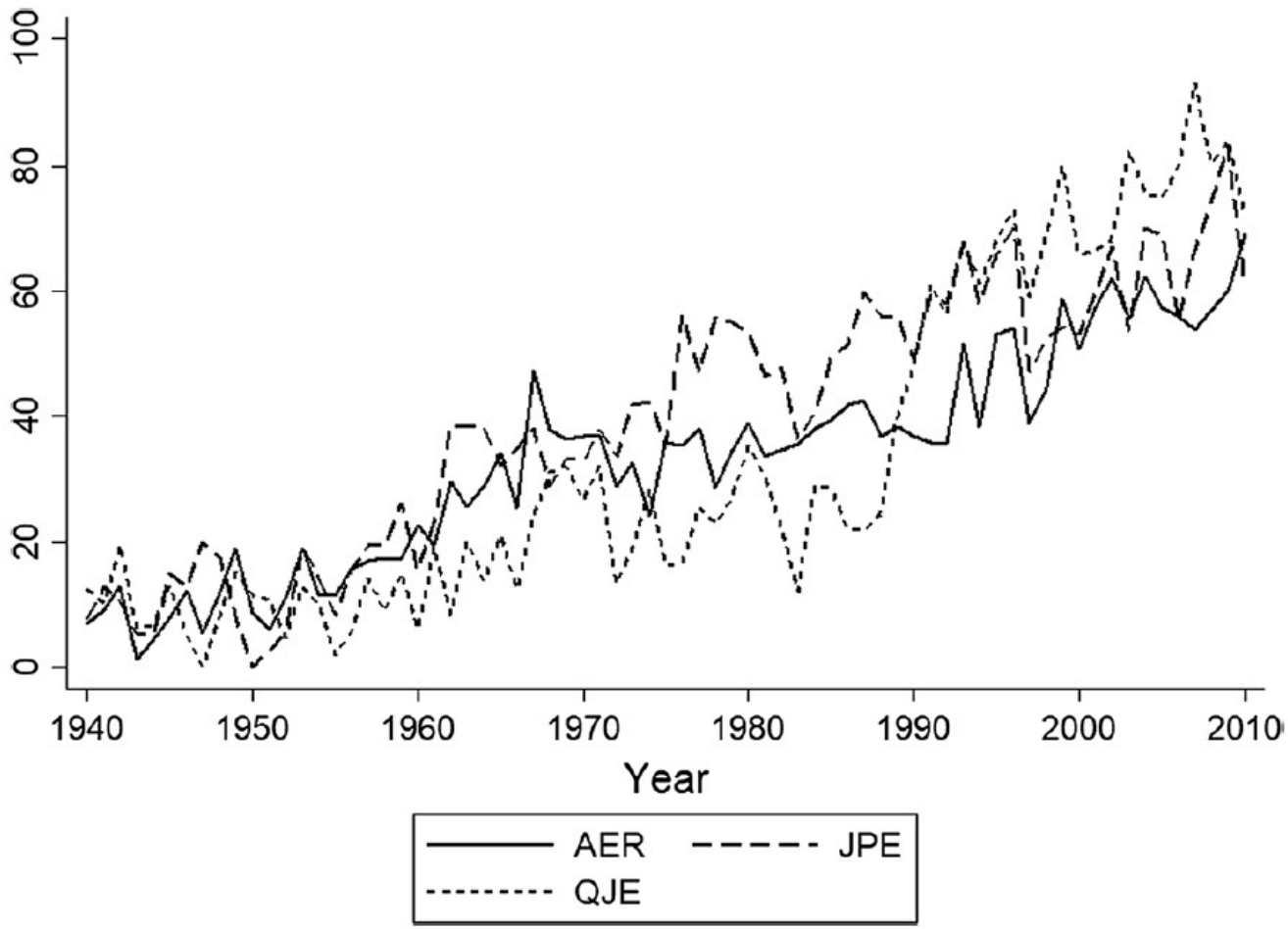
Theoretical Articles in "Top-3"



Source: Oliveira and Davila-Fernandez (2020).

Trends Toward Empirical Research

Applied Articles in "Top-3"



Source: Oliveira and Davila-Fernandez (2020).

Trends toward design-based methods

Trends Toward Design-Based Methods

Angrist and Pischke (2010)

- Until 1980s, regression analysis with causal claims.
- No attention to endogeneity (OVB) or reverse causality.

Trends Toward Design-Based Methods

Angrist and Pischke (2010)

- Until 1980s, regression analysis with causal claims.
- No attention to endogeneity (OVB) or reverse causality.
- Example: education production function literature:
  - Causal effect of class size on student performance?

Trends Toward Design-Based Methods

Angrist and Pischke (2010)

- Until 1980s, regression analysis with causal claims.
- No attention to endogeneity (OVB) or reverse causality.
- Example: education production function literature:
  - Causal effect of class size on student performance?
  - Regressions of performance on size yields negative estimates.
  - But selection: struggling students often grouped in smaller classes.

Trends Toward Design-Based Methods

- In the 1980s, increasing doubts about validity. Leamer (1983)
- In the 1990s, trend toward design-based methods to make causal claims. E.g. Card and Krueger (1994)
  - Gold standard: randomization studies.
  - Often not possible $\Longrightarrow$ search for quasi-natural experiments.
  - Main methods: matching, DiD, IV, RDDs.
  - Focus on exogenous sources of identifying variation (LATE).

Trends Toward Design-Based Methods

- In the 1980s, increasing doubts about validity. Leamer (1983)
- In the 1990s, trend toward design-based methods to make causal claims. E.g. Card and Krueger (1994)
  - Gold standard: randomization studies.
  - Often not possible $\Longrightarrow$ search for quasi-natural experiments.
  - Main methods: matching, DiD, IV, RDDs.
  - Focus on exogenous sources of identifying variation (LATE).
- Much more transparent causal claims.
- Focus on exogeneity and design, less on statistical aspects.

Criticisms of Design-Based Methods

- External validity. Deaton (2010) Heckman and Urzua (2009)

Criticisms of Design-Based Methods

- External validity. Deaton (2010) Heckman and Urzua (2009)
  - High internal validity but weak external validity.
  - Sources of identifying variations very local $\Longrightarrow$ little extrapolation.
  - But . . . "better LATE than nothing." Imbens (2010)
  - Empirical evidence is always local.
  - Replications helps understand heterogeneity due to context

Criticisms of Design-Based Methods

- Tackle unimportant questions ("playing small ball").

Criticisms of Design-Based Methods

- Tackle unimportant questions ("playing small ball").
  - Research driven by search for natural experiments.
  - Ends up giving trivial answers.
  - True: think about questions first, and design second.
  - But confounds triviality with narrow context.
  - Small balls win big games.

The replicability crisis

The Replicability Crisis

- Replicability issues:
  - Little replicability of published research findings.
  - Issues with *p*-values (publication bias, p-hacking. . .).
  - HARKing: hypothesizing ex-post.
  - Low statistical power.
  - Researcher degrees of freedom.

The Replicability Crisis

- Replicability issues:
  - Little replicability of published research findings.
  - Issues with *p*-values (publication bias, p-hacking. . .).
  - HARKing: hypothesizing ex-post.
  - Low statistical power.
  - Researcher degrees of freedom.
- Consequences for knowledge accumulation: a Bayesian approach.

The Replicability Crisis

- Replicability issues:
  - Little replicability of published research findings.
  - Issues with *p*-values (publication bias, p-hacking. . .).
  - HARKing: hypothesizing ex-post.
  - Low statistical power.
  - Researcher degrees of freedom.
- Consequences for knowledge accumulation: a Bayesian approach.
- Replication as a solution.

On these issues, see also Doucouliagos and Stanley (2011), Franco et al. (2014), Mervis (2014), Camerer et al. (2016), Duyx et al. (2017), Christensen and Miguel (2018), Andrews and Kasy (2019), Gordon et al. (2020), Ferraro and Shukla (2020)

Replicability issues

Little Replicability of Published Findings

Chang and Li (2017, 2022)

- Attempt to replicate 67 macro papers in 13 top journals.
- Use author-provided data and code.
- Successful replication: reproduce results qualitatively.

Little Replicability of Published Findings

Chang and Li (2017, 2022)

- Attempt to replicate 67 macro papers in 13 top journals.
- Use author-provided data and code.
- Successful replication: reproduce results qualitatively.
- Result:
  - Replicate 22 of 67 papers (33%).
  - Excluding papers with confidential data: 49% success rate.
  - Main reason: authors don't provide data and code.
  - When all files provided, 1/3 do not produce same qualitative results.

Little Replicability of Published Findings

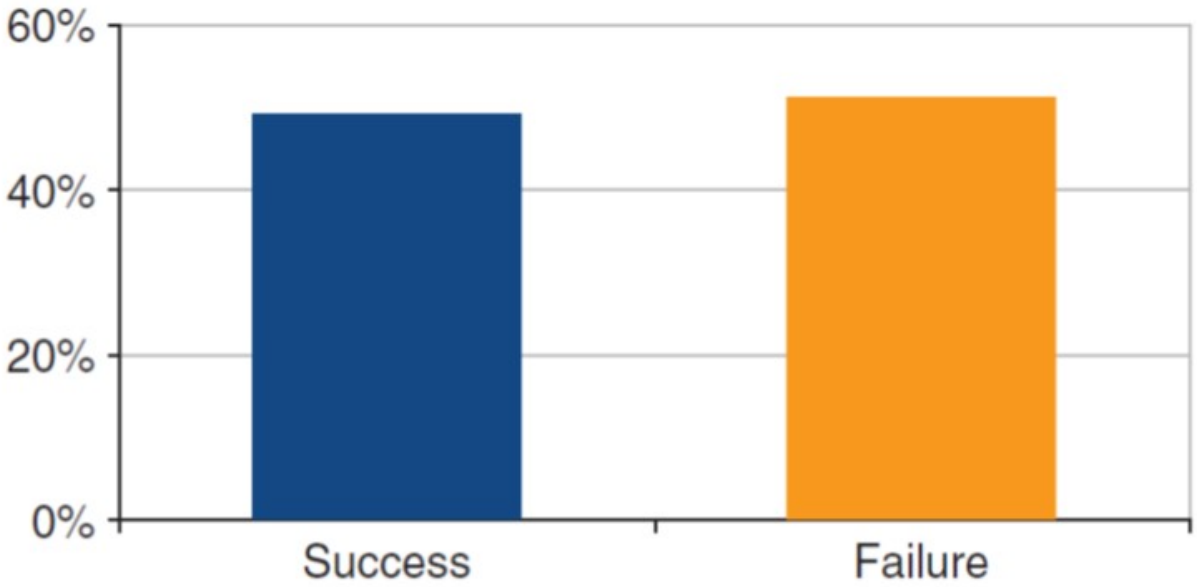


FIGURE 1. REPLICATION SUCCESS AND FAILURE RATES

Little Replicability of Published Findings

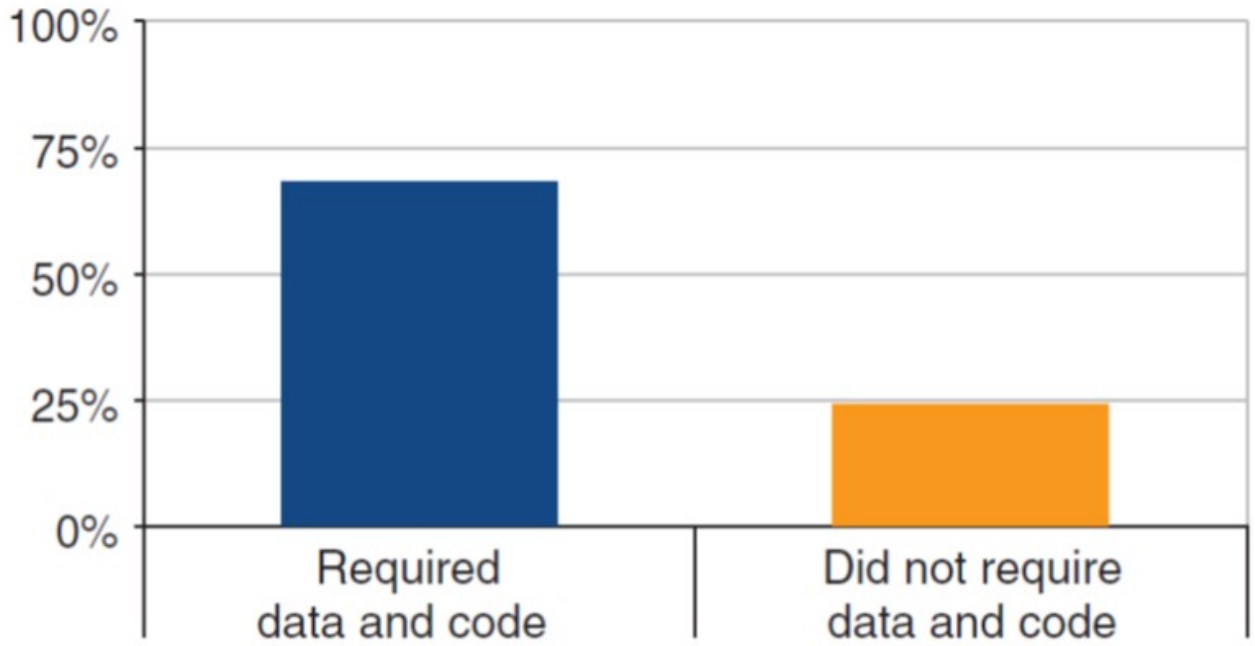


FIGURE 2. REPLICATION SUCCESS BY JOURNAL TYPE

Issues with P-Values

- Statistical inference:
  - To test hypothesis H_1, suppose null H_0 is true.
  - E.g. $H_0 : \overline{X} = 0$, $H_1 : \overline{X} \neq 0$.
  - Show data improbable under the null and reject H_0.

Issues with P-Values

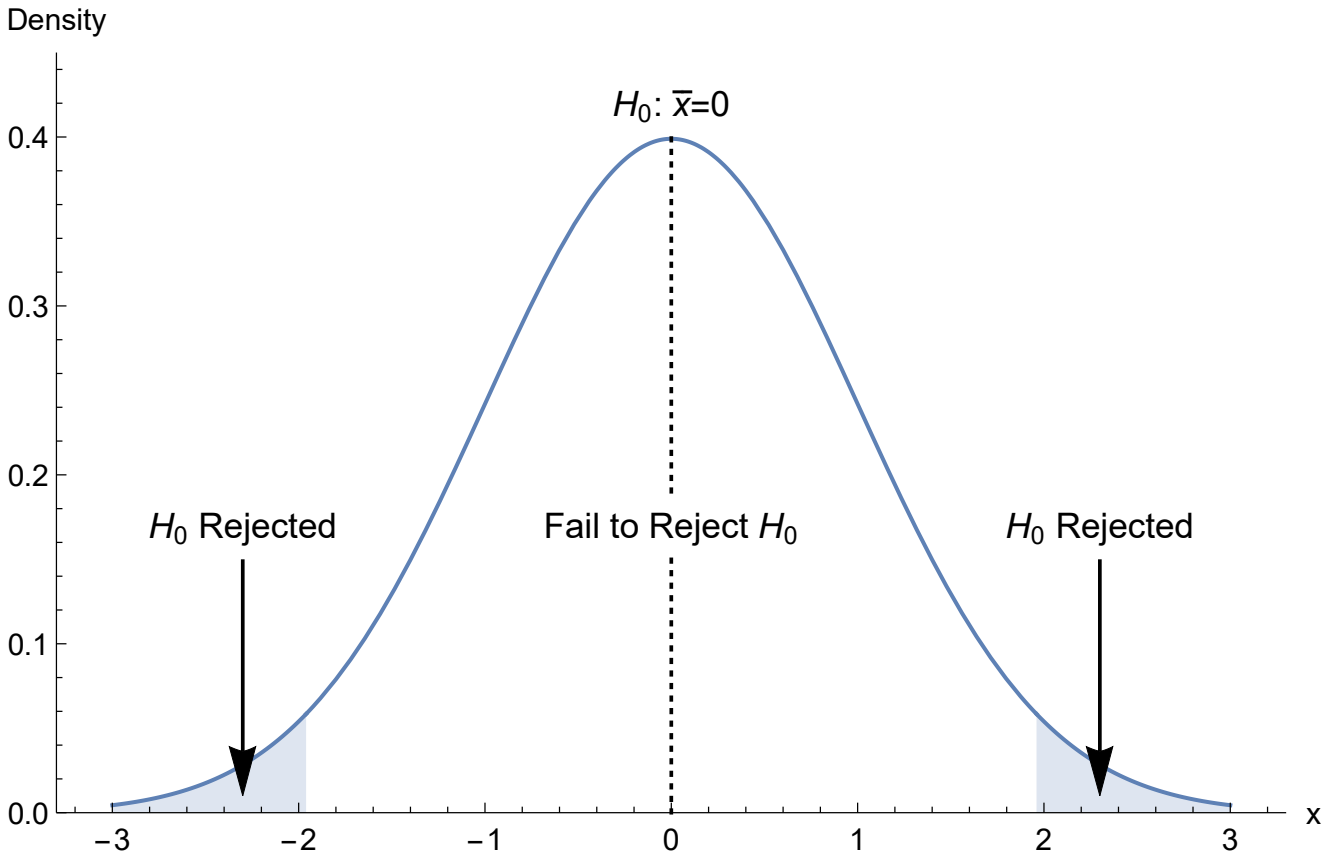
- Statistical inference:
  - To test hypothesis H_1, suppose null H_0 is true.
  - E.g. $H_0 : \overline{X} = 0$, $H_1 : \overline{X} \neq 0$.
  - Show data improbable under the null and reject H_0.
- p-value:
  - Probability that statistics more extreme than observed under H_0.
  - Statistics is "statistically insignificant" if unprobable under H_0.

Issues with P-Values

- Statistical inference:
  - To test hypothesis H_1, suppose null H_0 is true.
  - E.g. $H_0 : \overline{X} = 0$, $H_1 : \overline{X} \neq 0$.
  - Show data improbable under the null and reject H_0.
- p-value:
  - Probability that statistics more extreme than observed under H_0.
  - Statistics is "statistically insignificant" if unprobable under H_0.
- Significance level α:
  - Threshold such that reject H_0 if $p \leq \alpha$.
  - Type I error (false positive): $\alpha = \Pr(p \leq \alpha | H_0)$.
  - Usually $\alpha = 0.05$, but heated debates. Benjamin et al. (2017)

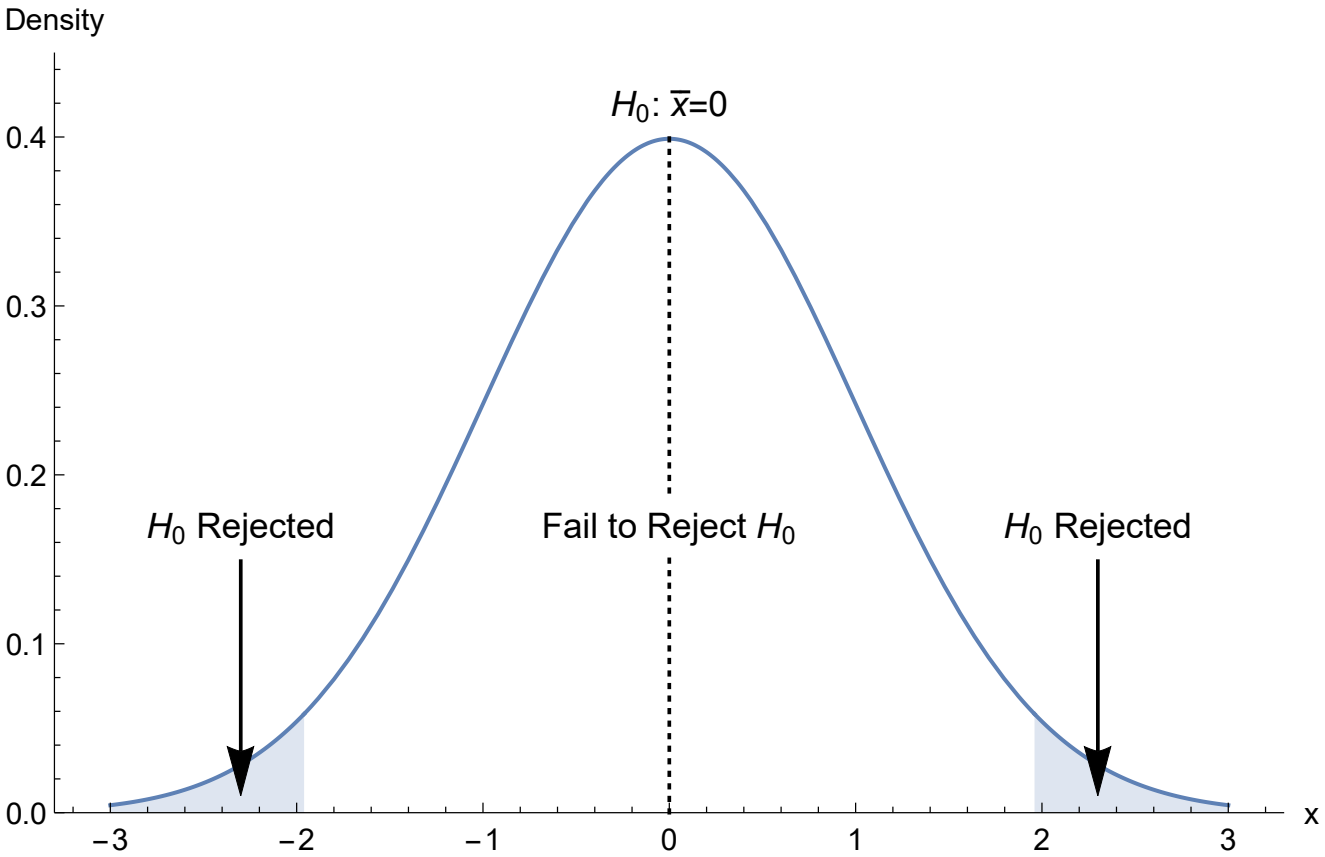
Issues with P-Values

- Random variable X follows standard normal: $X \sim \mathcal{N}(0, 1)$.



Issues with P-Values

- Random variable X follows standard normal: $X \sim \mathcal{N}(0, 1)$.



- p-value $= 0.05 \implies z$-score $= \pm 1.96$ s.d.

Issues with P-Values

- Common misunderstanding and misuse of p-values.
- The "cult of statistical significance." Ziliak and McCloskey (2008)
- American Statistical Association issued 6-point statement. Wasserstein and Lazar (2016)

$\Longrightarrow$ Everyone should read "Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations." Greenland et al. (2016)

Issues with P-Values

1. P-values can indicate how incompatible the data are with a specified statistical model.

Issues with P-Values

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the null hypothesis is true (it *assumes* the null is true).

Issues with *P*-Values

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the null hypothesis is true (it *assumes* the null is true).
3. Scientific conclusions should not be based only on whether a *p*-value passes a specific threshold.

Issues with *P*-Values

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the null hypothesis is true (it *assumes* the null is true).
3. Scientific conclusions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.

Issues with *P*-Values

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the null hypothesis is true (it *assumes* the null is true).
3. Scientific conclusions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Issues with *P*-Values

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the null hypothesis is true (it *assumes* the null is true).
3. Scientific conclusions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Issues with *P*-Values

Brodeur et al. (2016)

- Collect 50k test statistics from 641 articles in top-3 economics journals 2005–2011.
- Assume selection by journals is monotonically increasing with value of test statistics (standardized z-statistics).
- Extract residual from *selection* process to identify *inflation*.

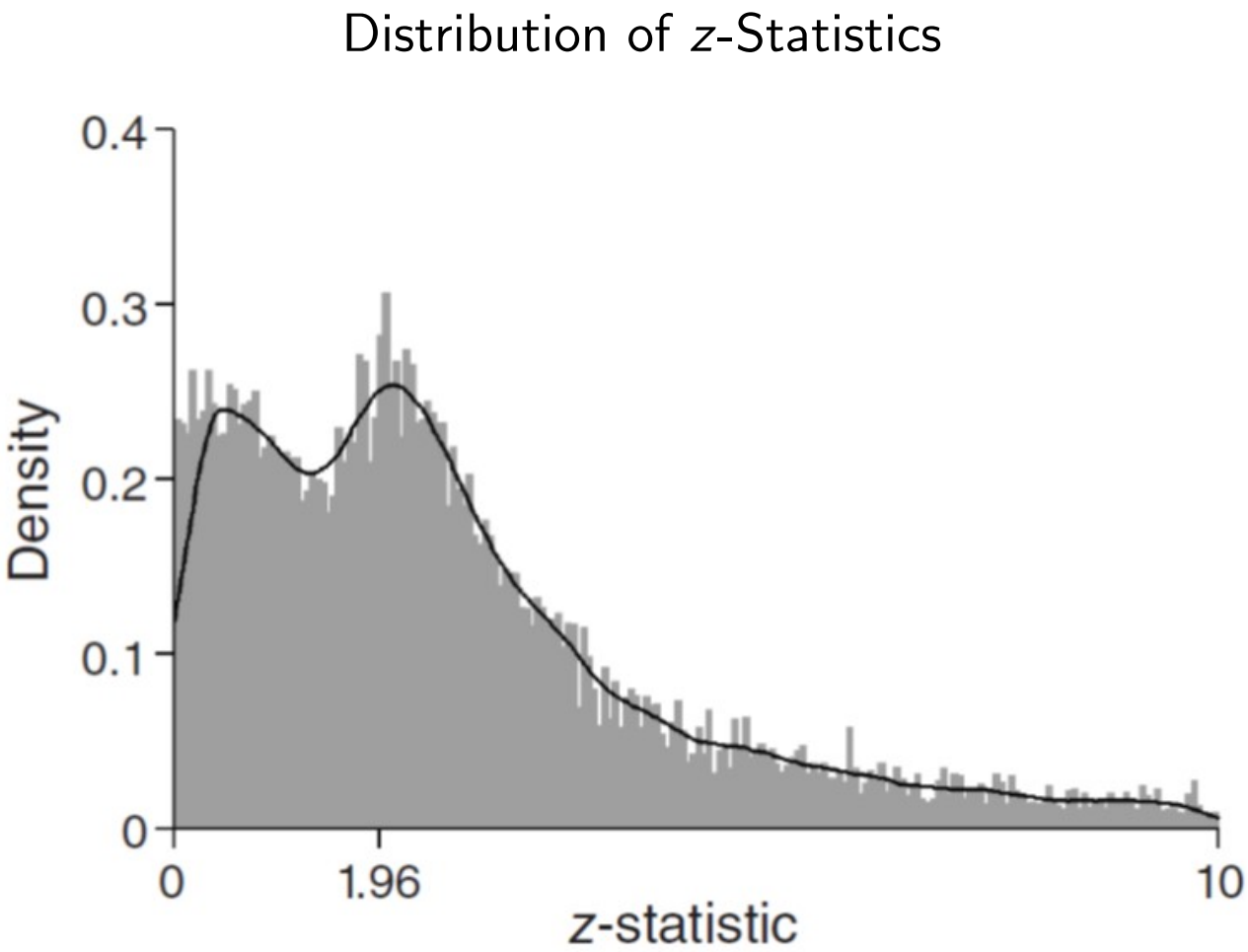
Issues with *P*-Values

Brodeur et al. (2016)

- Collect 50k test statistics from 641 articles in top-3 economics journals 2005–2011.
- Assume selection by journals is monotonically increasing with value of test statistics (standardized z-statistics).
- Extract residual from *selection* process to identify *inflation*.
- Find a two-humped camel shape with
  - Too little marginally insignificant results (p-value of 0.10–0.25).
  - Too many marginally significant results (p-value of 0.05–0.10)

$\Longrightarrow$ Evidence of p-hacking and specification searching.

Issues with P-Values



Issues with P-Values

Brodeur et al. (2020)

- Assess how sensitive this issue is to empirical design.
- Degrees of freedom differ by method.
- Collect 11k test statistics for 300 articles in 25 top journals in 2015.

Issues with *P*-Values

Brodeur et al. (2020)

- Assess how sensitive this issue is to empirical design.
- Degrees of freedom differ by method.
- Collect 11k test statistics for 300 articles in 25 top journals in 2015.
- Results relative to RCT:
  - IV and DID are 15% more likely to be statistically significant.
  - RDD are not.
- Due to bunching around 5% significance thresholds.

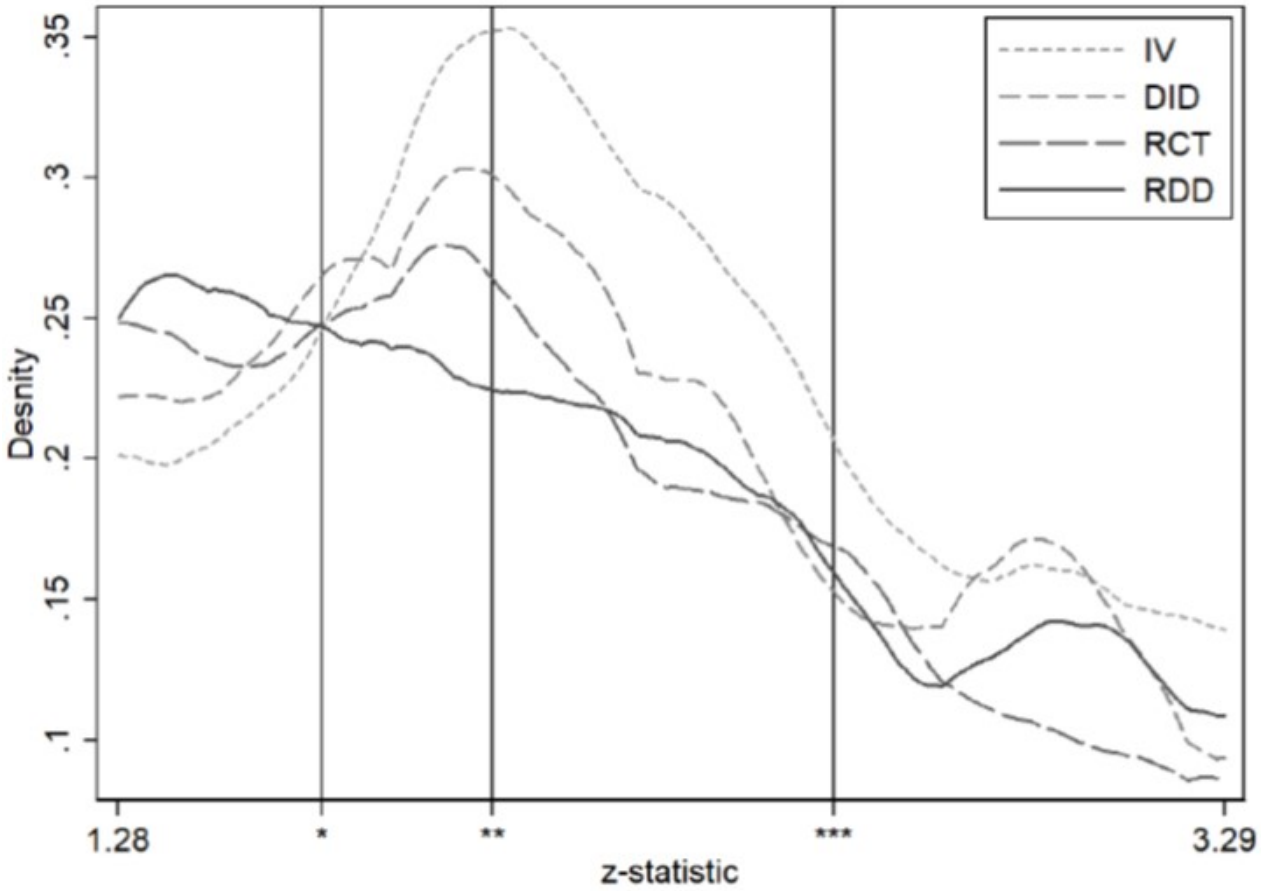
Issues with *P*-Values

Brodeur et al. (2020)

- Assess how sensitive this issue is to empirical design.
- Degrees of freedom differ by method.
- Collect 11k test statistics for 300 articles in 25 top journals in 2015.
- Results relative to RCT:
  - IV and DID are 15% more likely to be statistically significant.
  - RDD are not.
- Due to bunching around 5% significance thresholds.
- For IV: same result for F-statistic at 10-threshold.

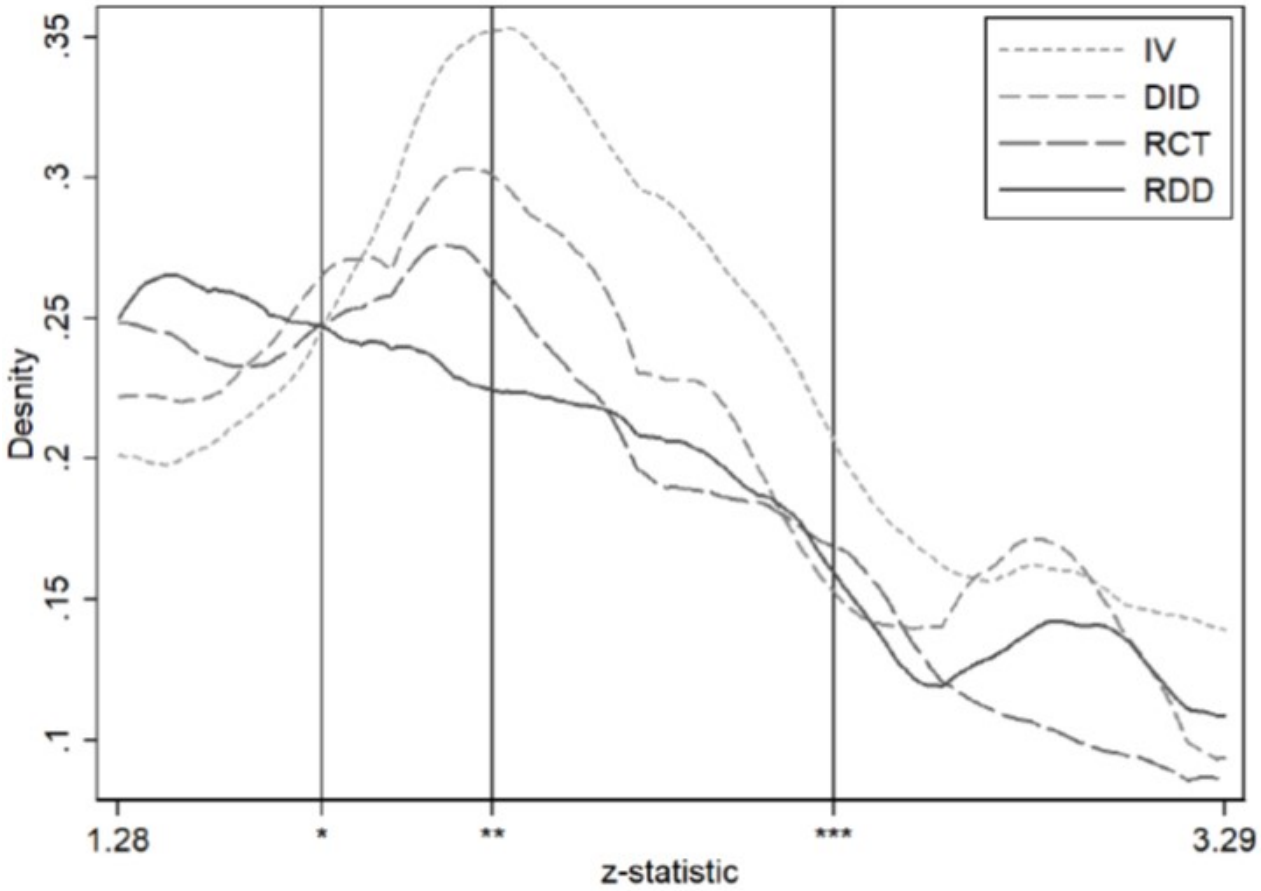
Issues with P-Values

z-Curves by Method



Issues with *P*-Values

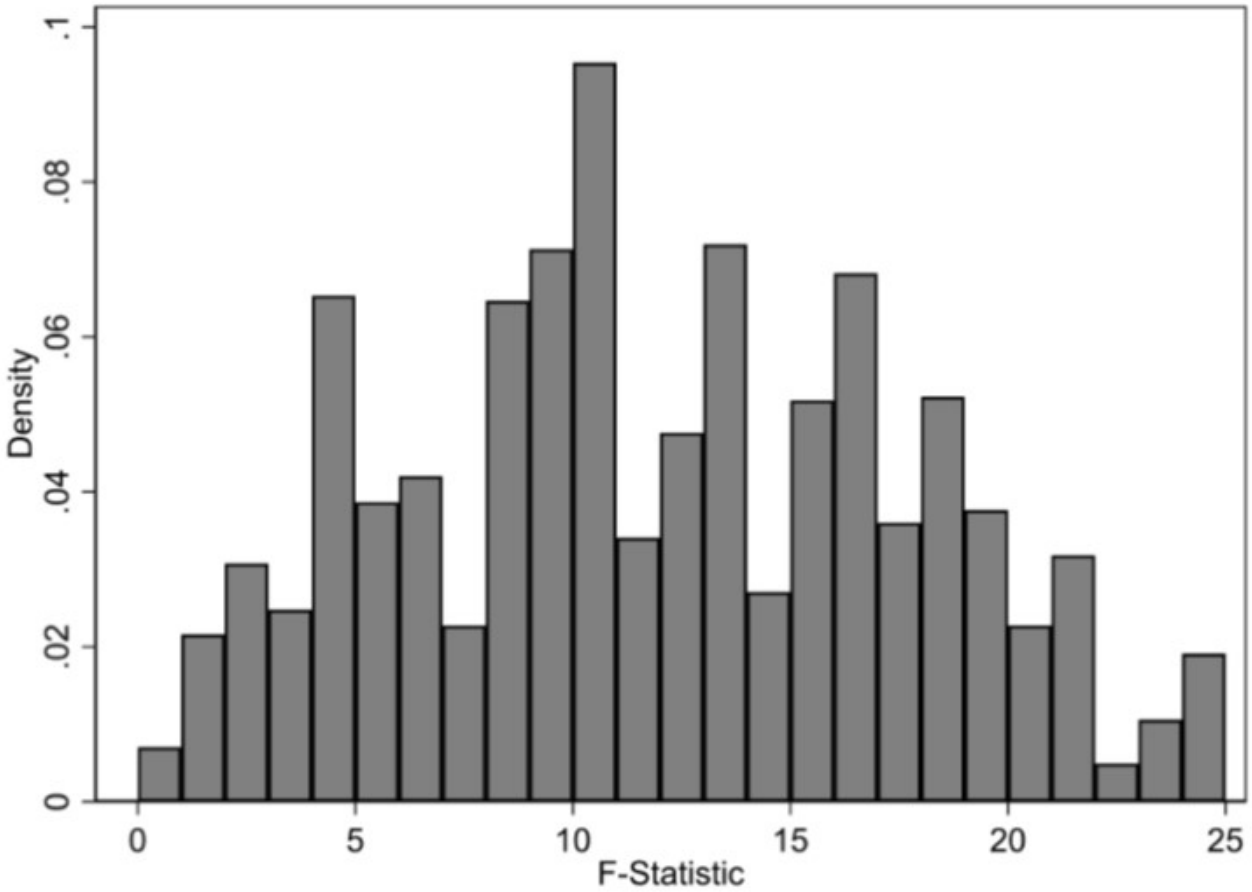
z-Curves by Method



- But "speed" of z-changes vary by method.

Issues with P-Values

First Stage F-Statistics



HARKing: Hypothesizing Ex-Post

- HARKing: hypothesizing after the results are known. Kerr (1998)
- Converse to Popperian approach to science.

HARKing: Hypothesizing Ex-Post

- HARKing: hypothesizing after the results are known. Kerr (1998)
- Converse to Popperian approach to science.
- Worse case (SHARKing): secretly HARKing, i.e. reporting as a ex-ante theorizing a hypothesis formed after results are known.

HARKing: Hypothesizing Ex-Post

- HARKing: hypothesizing after the results are known. Kerr (1998)
- Converse to Popperian approach to science.
- Worse case (SHARKing): secretly HARKing, i.e. reporting as a ex-ante theorizing a hypothesis formed after results are known.
- Incentives from editors and readers preferences of simple stories.
- Result of cognitive biases (confirmation and hindsight).

HARKing: Hypothesizing Ex-Post

- HARKing: hypothesizing after the results are known. Kerr (1998)
- Converse to Popperian approach to science.
- Worse case (SHARKing): secretly HARKing, i.e. reporting as a ex-ante theorizing a hypothesis formed after results are known.
- Incentives from editors and readers preferences of simple stories.
- Result of cognitive biases (confirmation and hindsight).
- More informative to contradict than confirm theory (if enough power). Abadie (2020)

HARKing: Hypothesizing Ex-Post

- HARKing: hypothesizing after the results are known. Kerr (1998)
- Converse to Popperian approach to science.
- Worse case (SHARKing): secretly HARKing, i.e. reporting as a ex-ante theorizing a hypothesis formed after results are known.
- Incentives from editors and readers preferences of simple stories.
- Result of cognitive biases (confirmation and hindsight).
- More informative to contradict than confirm theory (if enough power). Abadie (2020)
- Damaging to science: translate type I error into theory.

Low Statistical Power

Ioannidis et al. (2017)

- Adequate statistical power (e.g. 80%): empirical methods and data should be able to detect an effect, should it be there.
- Low power $\Longrightarrow$ high rate of false negatives.

Low Statistical Power

Ioannidis et al. (2017)

- Adequate statistical power (e.g. 80%): empirical methods and data should be able to detect an effect, should it be there.
- Low power $\Longrightarrow$ high rate of false negatives.
- Examine 159 meta-analyzes, providing 64k estimates from 7k studies.

Low Statistical Power

Ioannidis et al. (2017)

- Adequate statistical power (e.g. 80%): empirical methods and data should be able to detect an effect, should it be there.
- Low power $\Longrightarrow$ high rate of false negatives.
- Examine 159 meta-analyzes, providing 64k estimates from 7k studies.
- Half of economic literatures have 90% of results under-powered.
- Median statistical power is 18%.

Researcher Degrees of Freedom

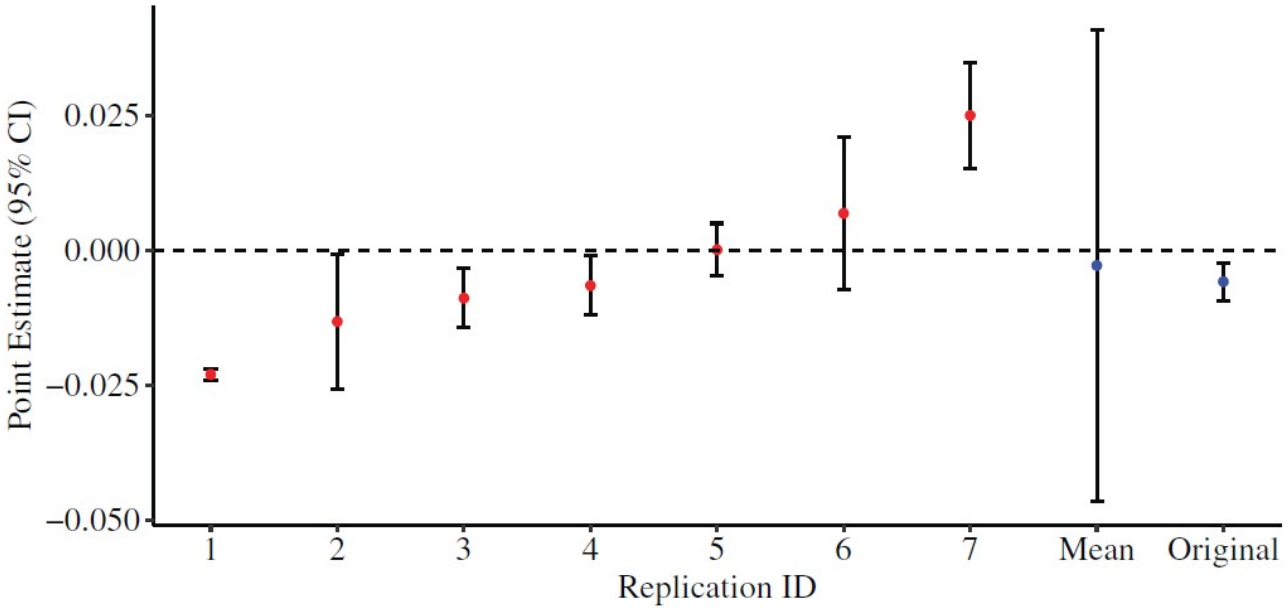
- Hundreds of hidden decisions made from data cleaning to variable definition, even if empirical approach constant.
- How much variation or noise induced by unreported decisions?

Researcher Degrees of Freedom

- Hundreds of hidden decisions made from data cleaning to variable definition, even if empirical approach constant.
- How much variation or noise induced by unreported decisions?
- Many-analyst approach independently on same raw data and same question. Hungtington-Klein et al. (2020)
  - Allow researcher freedom in construction and cleaning dataset.
  - Focus on one estimate in 2 studies, each replicated 7 times:
    - Black et al. (2008) Effect of compulsory schooling on teenage pregnancy.
    - Fairlie et al. (2011) Effect of employer-based health insurance on entrepreneurship.
  - Main reason for differences: data cleaning.

Researcher Degrees of Freedom

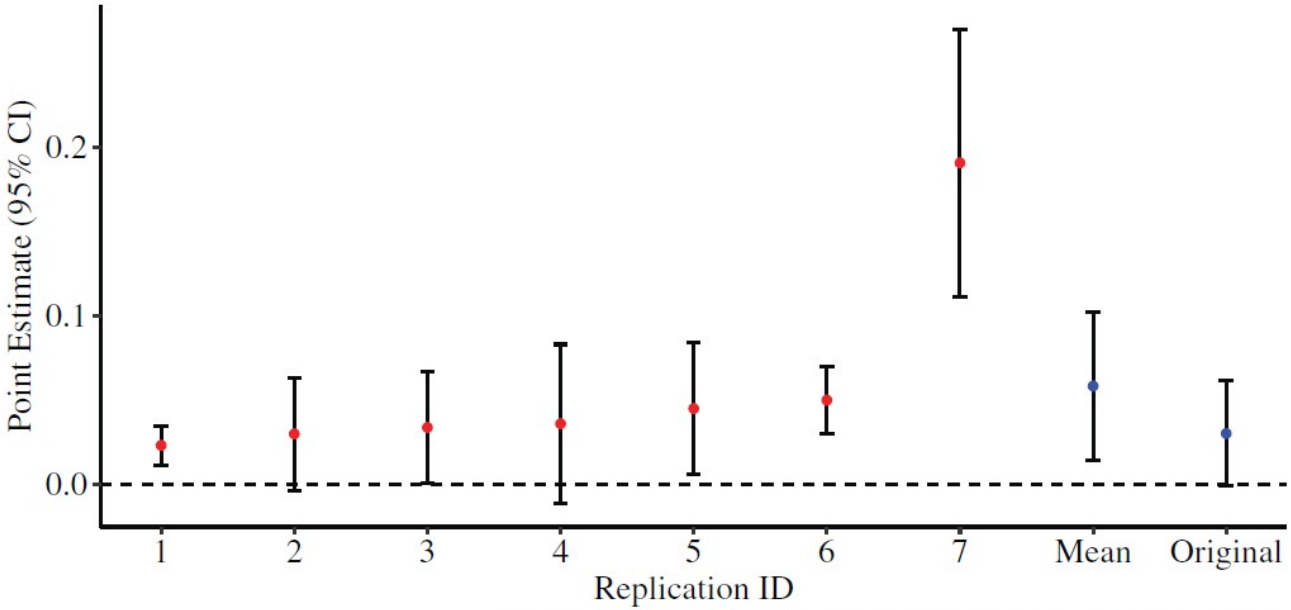
Results on Black et al. (2008)



"Mean" shows mean of replications and confidence interval for that mean.
"Original" is from the original publication.

Researcher Degrees of Freedom

Results on Fairlie et al. (2011)



"Mean" shows mean of replications and confidence interval for that mean.
"Original" is from the original publication.

A Bayesian approach to knowledge accumulation

Statistics Reminder

- Statistical inference:
  - To test hypothesis H_1, suppose null H_0 is true.
  - E.g. $H_0 : \overline{X} = 0$, $H_1 : \overline{X} \neq 0$.
  - Show data improbable under the null and reject H_0.

Statistics Reminder

- Statistical inference:
  - To test hypothesis H_1, suppose null H_0 is true.
  - E.g. $H_0 : \overline{X} = 0$, $H_1 : \overline{X} \neq 0$.
  - Show data improbable under the null and reject H_0.
- Significance level α:
  - Threshold such that reject H_0 if $p \leq \alpha$.
  - Type I error (false positive): $\alpha = \Pr(p \leq \alpha | H_0)$.
  - Usually $\alpha = 0.05$, but heated debates.

Statistics Reminder

- Statistical inference:
  - To test hypothesis H_1, suppose null H_0 is true.
  - E.g. $H_0 : \overline{X} = 0$, $H_1 : \overline{X} \neq 0$.
  - Show data improbable under the null and reject H_0.
- Significance level α:
  - Threshold such that reject H_0 if $p \leq \alpha$.
  - Type I error (false positive): $\alpha = \Pr(p \leq \alpha | H_0)$.
  - Usually $\alpha = 0.05$, but heated debates.
- Statistical power $1 - \beta$:
  - Type II error (false negative): $\beta = \Pr(\texttt{fail reject } H_0 | H_0 \texttt{ false})$.
  - Effect size and sample size affect power.

Statistics Reminder

	Null hypothesis H_0	
	True	False
Reject H_0	Type I error *false positive* α	Correct *true positive* $1-\beta$
Fail to reject H_0	Correct *true negative* $1-\alpha$	Type II error *false negative* β

A Bayesian Approach to Knowledge

Ioannidis (2005) Maniadis, Tufano and List (2014, 2017)

- Notations:
  - n: number of scientific associations to be investigated.
  - π: fraction of true associations (or prior).
  - $1-\beta$: power of test.
  - α: statistical significance, or false positive probability.

A Bayesian Approach to Knowledge

Ioannidis (2005) Maniadis, Tufano and List (2014, 2017)

- Notations:
  - n: number of scientific associations to be investigated.
  - π: fraction of true associations (or prior).
  - $1-\beta$: power of test.
  - α: statistical significance, or false positive probability.
- Quantities of interest:
  - True associations: $\pi \cdot n$
  - False associations: $(1-\pi) \cdot n$

A Bayesian Approach to Knowledge

Ioannidis (2005) Maniadis, Tufano and List (2014, 2017)

- Notations:
  - n: number of scientific associations to be investigated.
  - π: fraction of true associations (or prior).
  - $1-\beta$: power of test.
  - α: statistical significance, or false positive probability.
- Quantities of interest:
  - True associations: $\pi \cdot n$
  - False associations: $(1-\pi) \cdot n$
  - **True associations declared true:** $(1-\beta) \cdot \pi \cdot n$

A Bayesian Approach to Knowledge

Ioannidis (2005) Maniadis, Tufano and List (2014, 2017)

- Notations:
  - n: number of scientific associations to be investigated.
  - π: fraction of true associations (or prior).
  - $1-\beta$: power of test.
  - α: statistical significance, or false positive probability.
- Quantities of interest:
  - True associations: $\pi \cdot n$
  - False associations: $(1-\pi) \cdot n$
  - **True associations declared true:** $(1-\beta) \cdot \pi \cdot n$
  - **False associations declared true:** $\alpha \cdot (1-\pi) \cdot n$

A Bayesian Approach to Knowledge

- Post-study probability (PSP):
  - PSP: probability that an association declared true is actually true.
  - PSP$=\Pr\left(H_0 \texttt{ false} | \texttt{reject } H_0\right)$.

A Bayesian Approach to Knowledge

- Post-study probability (PSP):
  - PSP: probability that an association declared true is actually true.
  - PSP$=\Pr\left(H_0 \text{ false} | \text{reject } H_0\right)$.
- Using Baysian updating (Bayes rule).

$$\text{PSP} = \frac{\text{number of true associations declared true}}{\text{number of associations declared true}}$$

A Bayesian Approach to Knowledge

- Post-study probability (PSP):
  - PSP: probability that an association declared true is actually true.
  - PSP$=\Pr\left(H_0 \text{ false} | \text{reject } H_0\right)$.
- Using Baysian updating (Bayes rule).

$$\text{PSP} = \frac{\text{number of true associations declared true}}{\text{number of associations declared true}}$$

$$\text{PSP} = \frac{(1-\beta)\,\pi}{\underbrace{(1-\beta)\,\pi}_{\text{true positives}} + \underbrace{\alpha\,(1-\pi)}_{\text{false positives}}}$$

A Bayesian Approach to Knowledge

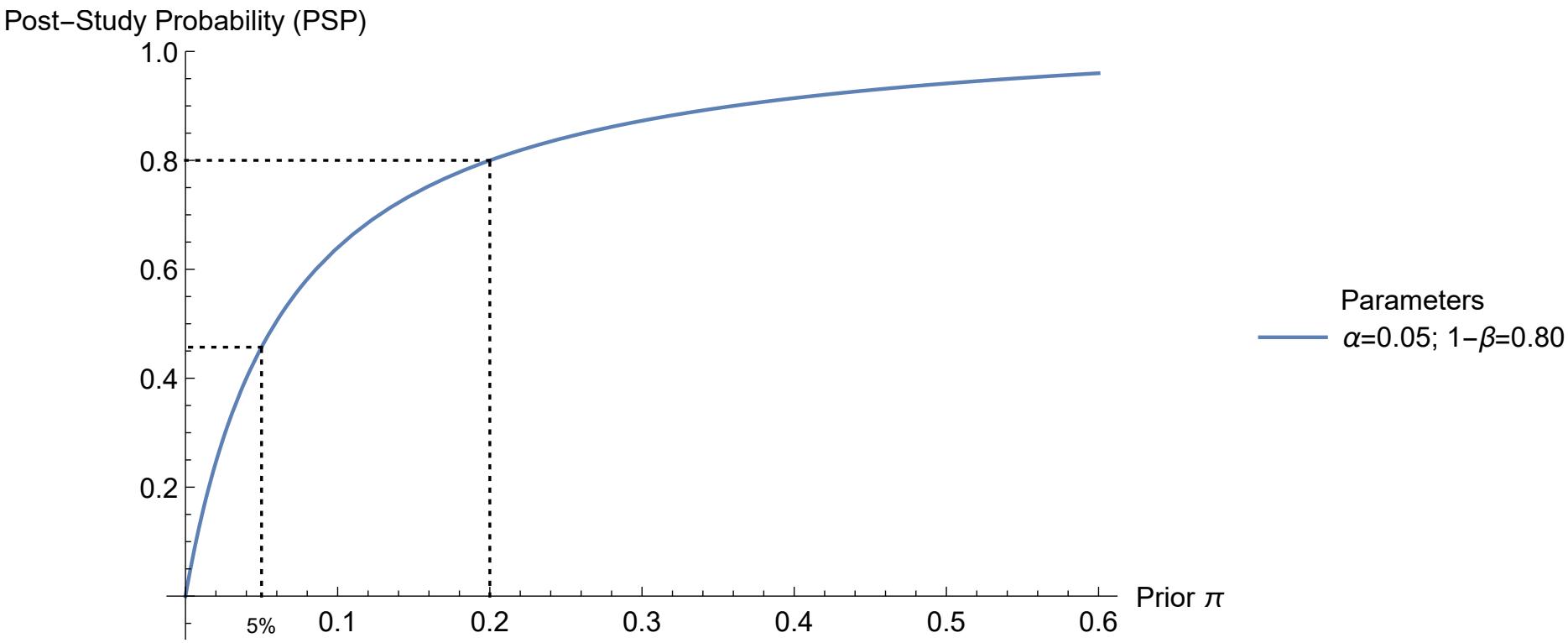
PSP: probability that an association declared true is actually true.

$$\texttt{Post-Study Probability} = \frac{(1-\beta)\,\pi}{\underbrace{(1-\beta)\,\pi}_{\texttt{true positives}} + \underbrace{\alpha\,(1-\pi)}_{\texttt{false positives}}}$$

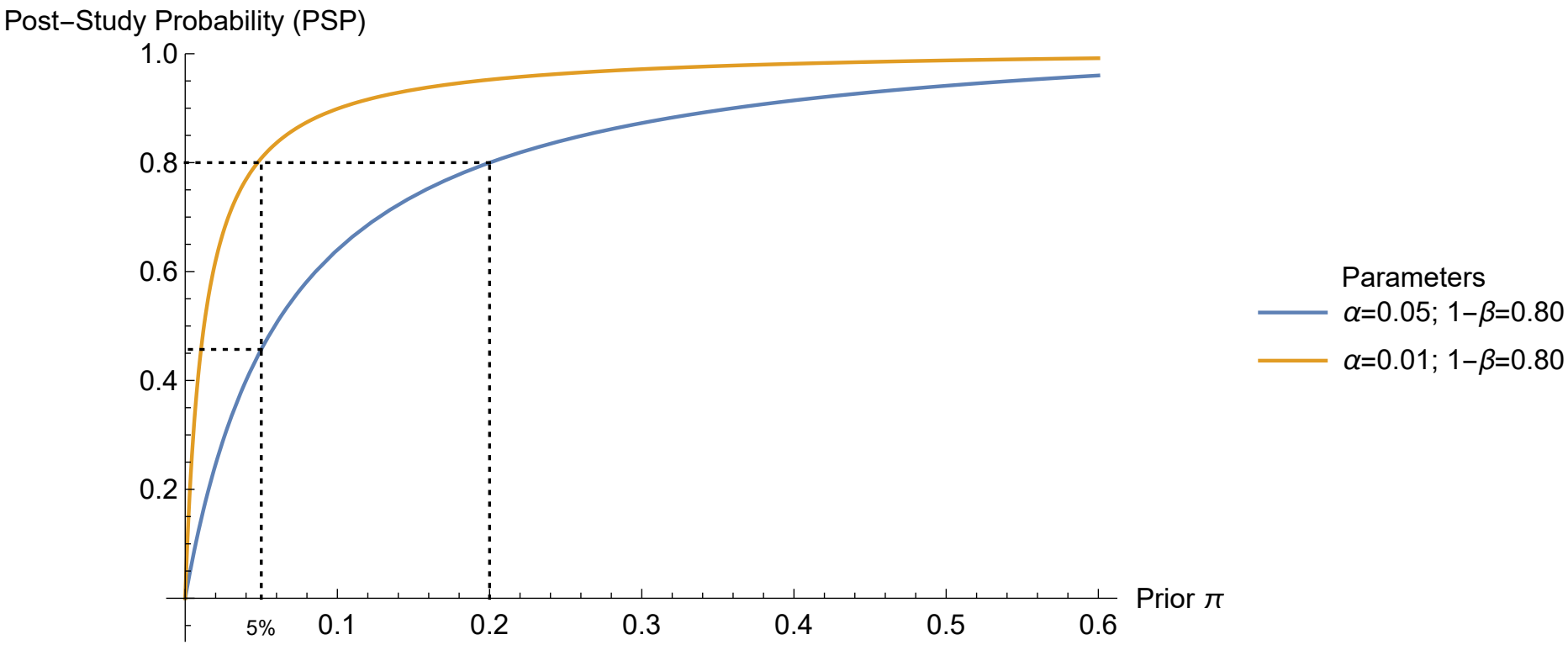
Comparative statics:

- Prior $\pi \uparrow \Longrightarrow$ PSP $\uparrow$.
- Statistical significance $\alpha \downarrow \Longrightarrow$ PSP $\uparrow$.
- Power $1-\beta \uparrow \Longrightarrow$ PSP $\uparrow$.

What Can We Learn From One Study?

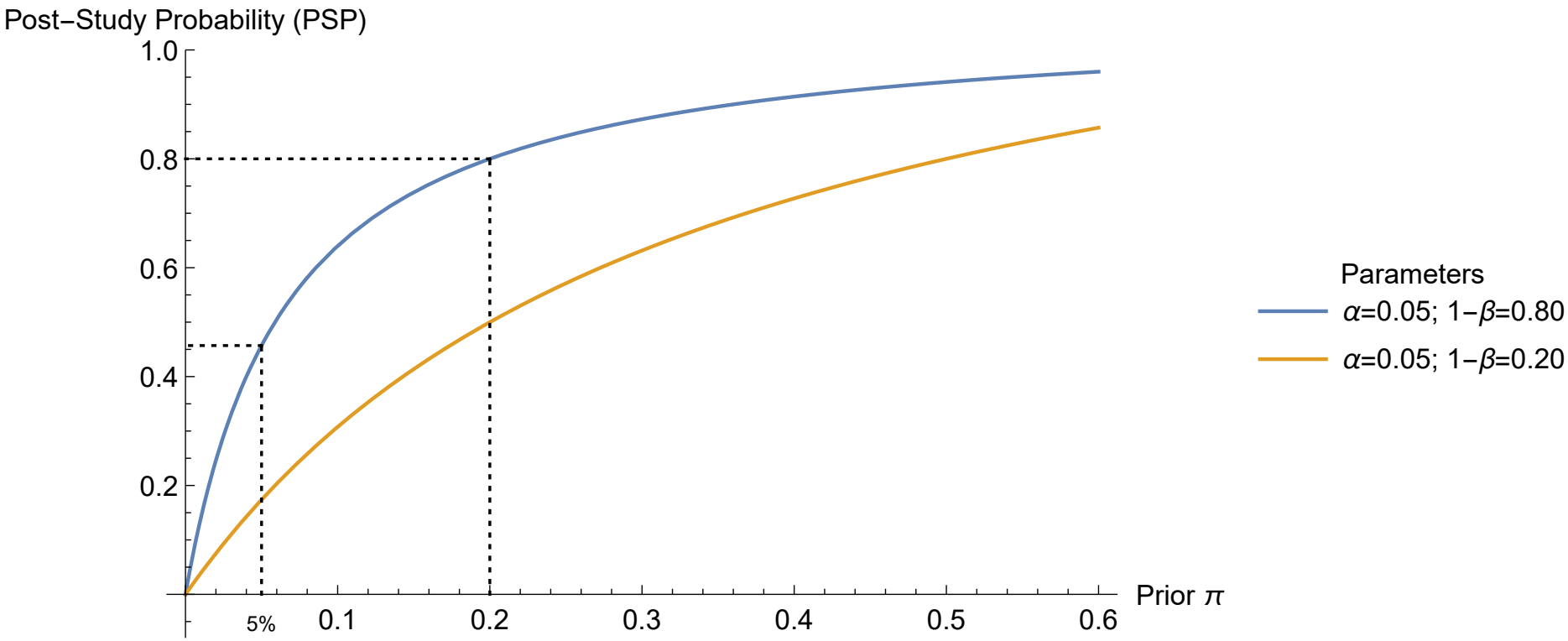


What Can We Learn From One Study?



Statistical significance $\alpha \downarrow \Longrightarrow$ PSP $\uparrow$

What Can We Learn From One Study?



Statistical power $1 - \beta \uparrow \Longrightarrow$ PSP $\uparrow$

What Can We Learn From Many Studies?

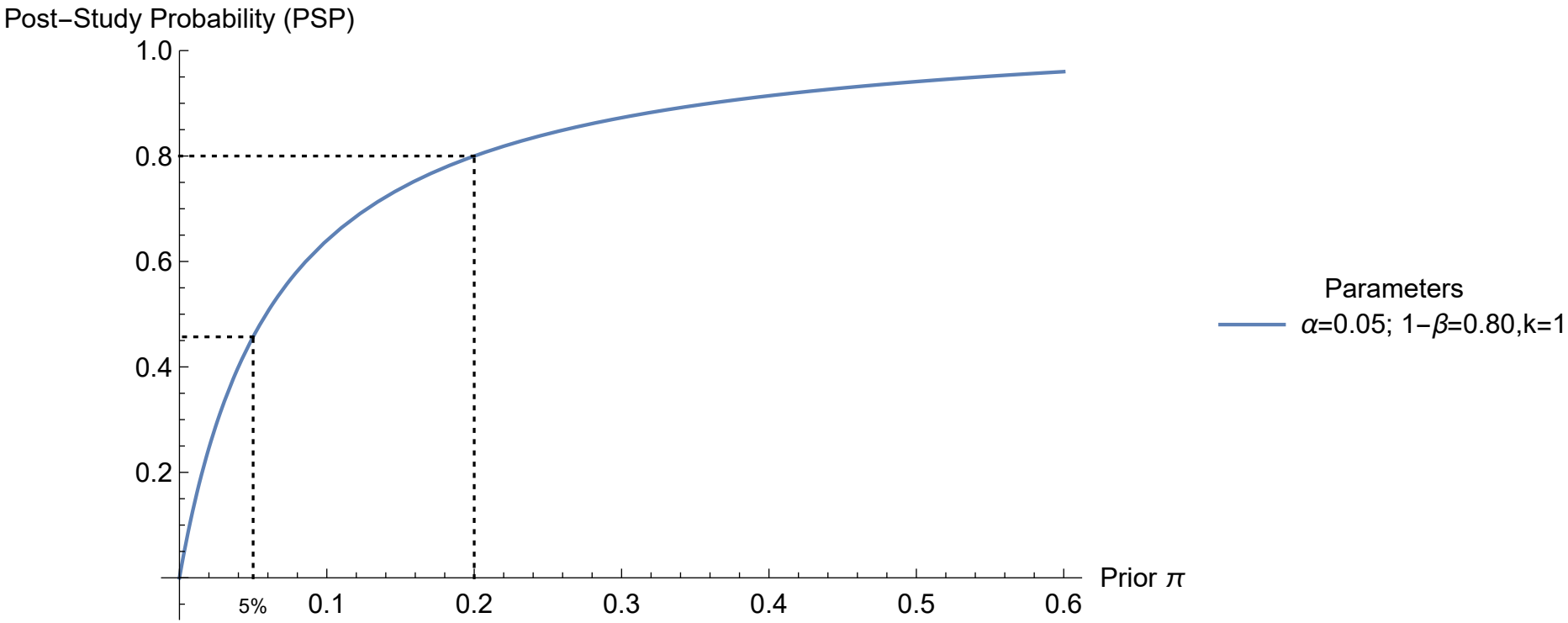
- Usually number of researchers simultaneously working $k > 1$.
- Two implications:
  - Number of true positives $\uparrow$.
  - Number of false positives $\uparrow$.
- But false positives $\uparrow$ faster than true positives.

What Can We Learn From Many Studies?

- Usually number of researchers simultaneously working $k > 1$.
- Two implications:
  - Number of true positives $\uparrow$.
  - Number of false positives $\uparrow$.
- But false positives $\uparrow$ faster than true positives.

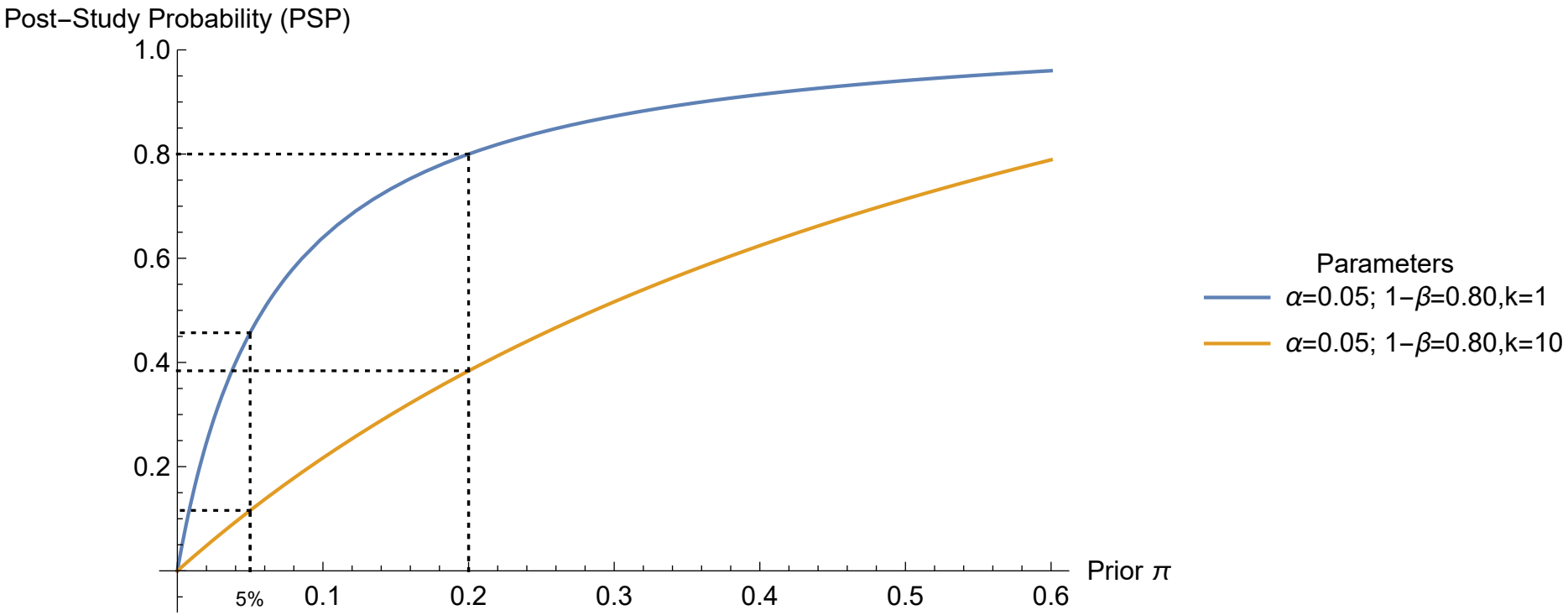
$\Longrightarrow$ Post-study probability $\downarrow$ when number of researchers $\uparrow$

What Can We Learn From Many Studies?



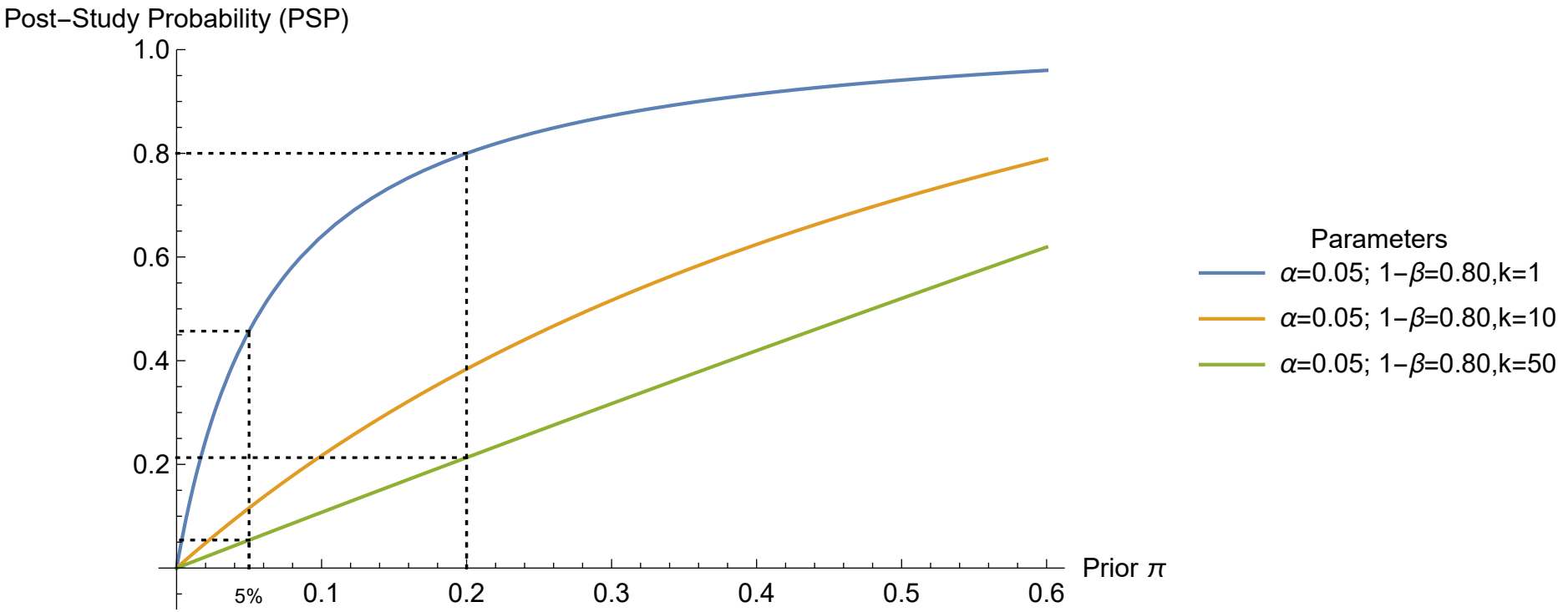
$k = 1$ researcher

What Can We Learn From Many Studies?



$k = 10$ researchers

What Can We Learn From Many Studies?



$k = 50$ researchers

Replication as a Solution

Replication as a Solution

Clemens (2017)

- Replication tests:
  - Verification tests:
    - Original data, same sample, identical method.
    - Solves measurement errors, coding errors, data construction errors.

Replication as a Solution

Clemens (2017)

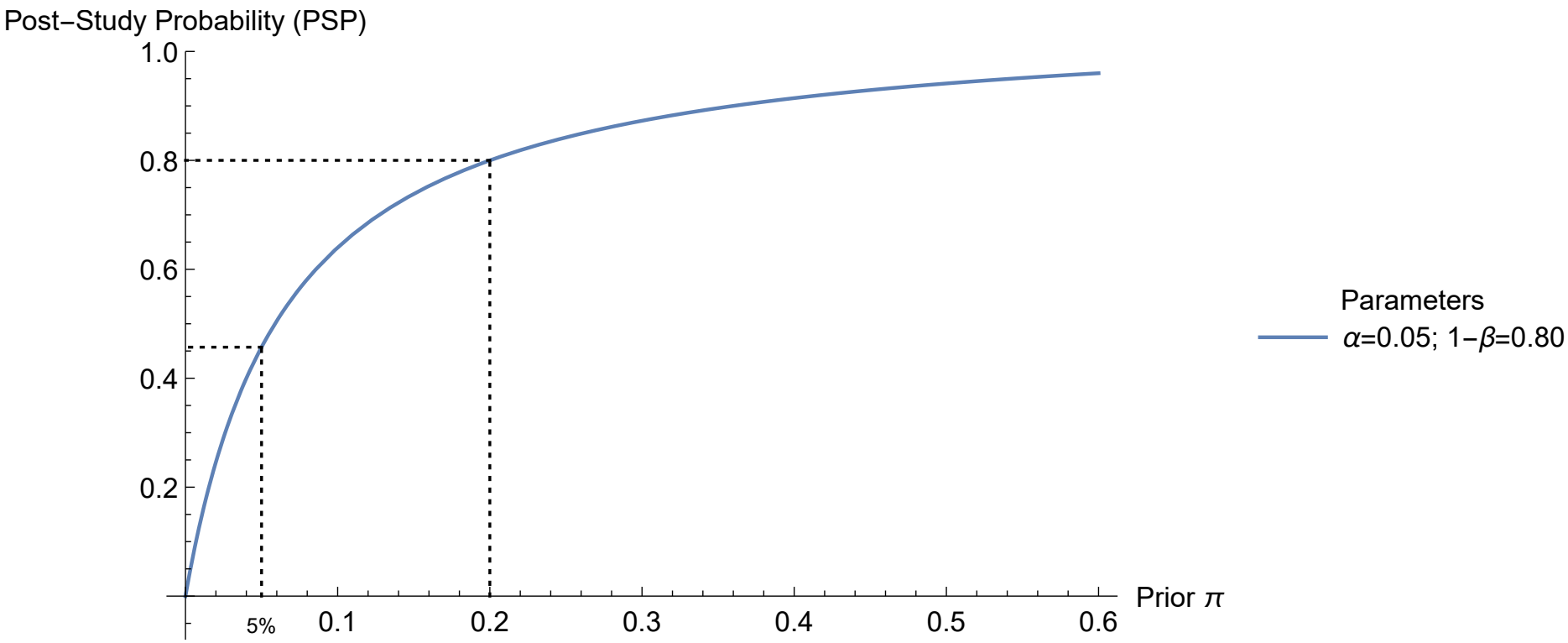
- Replication tests:
  - Verification tests:
    - Original data, same sample, identical method.
    - Solves measurement errors, coding errors, data construction errors.
  - Reproduction tests:
    - Resampling same population, identical method.
    - Solves sampling errors, low power.

Replication as a Solution

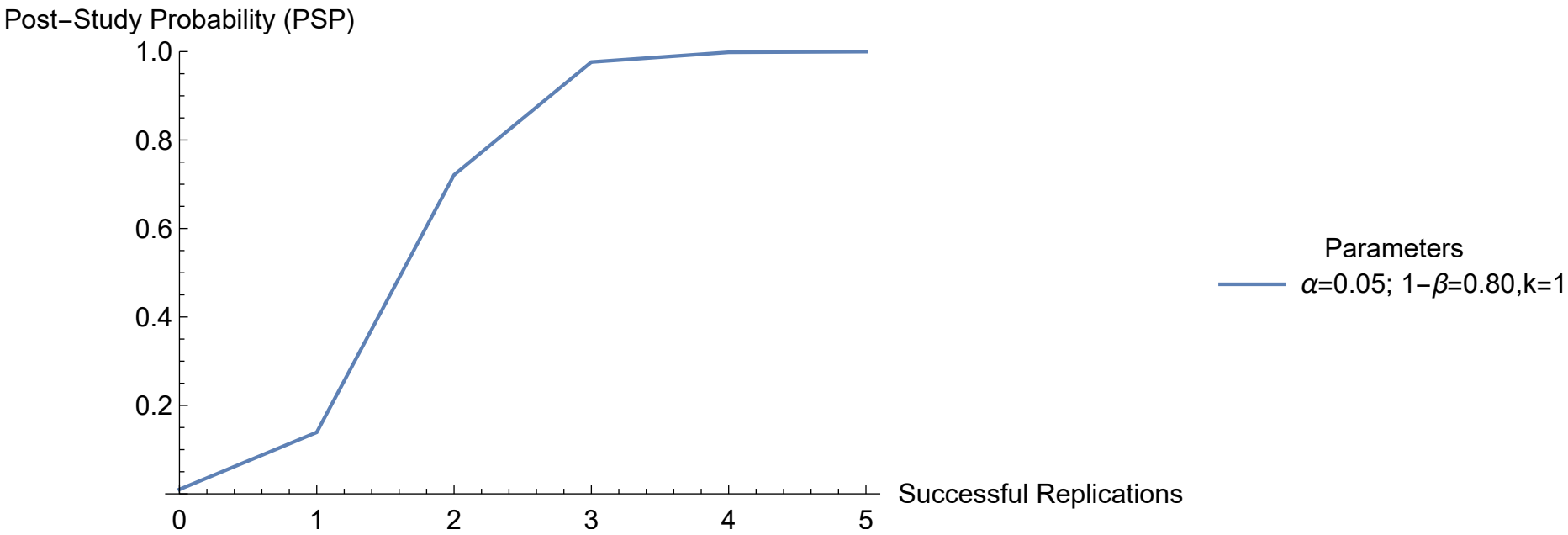
Clemens (2017)

- Replication tests:
  - Verification tests:
    - Original data, same sample, identical method.
    - Solves measurement errors, coding errors, data construction errors.
  - Reproduction tests:
    - Resampling same population, identical method.
    - Solves sampling errors, low power.
- Robustness tests:
  - Reanalysis: same population, different method.
  - Extension: resampling, different method.

Replication as a Solution

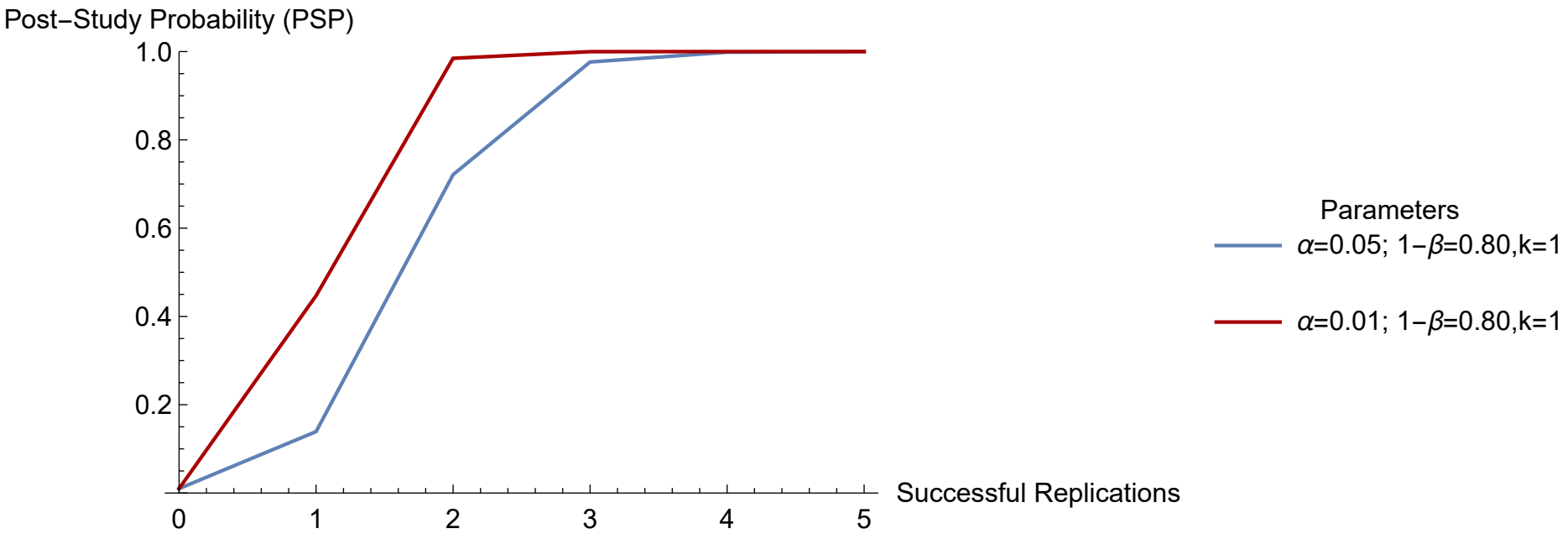


Replication as a Solution



Starting prior: $\pi = 1\%$.

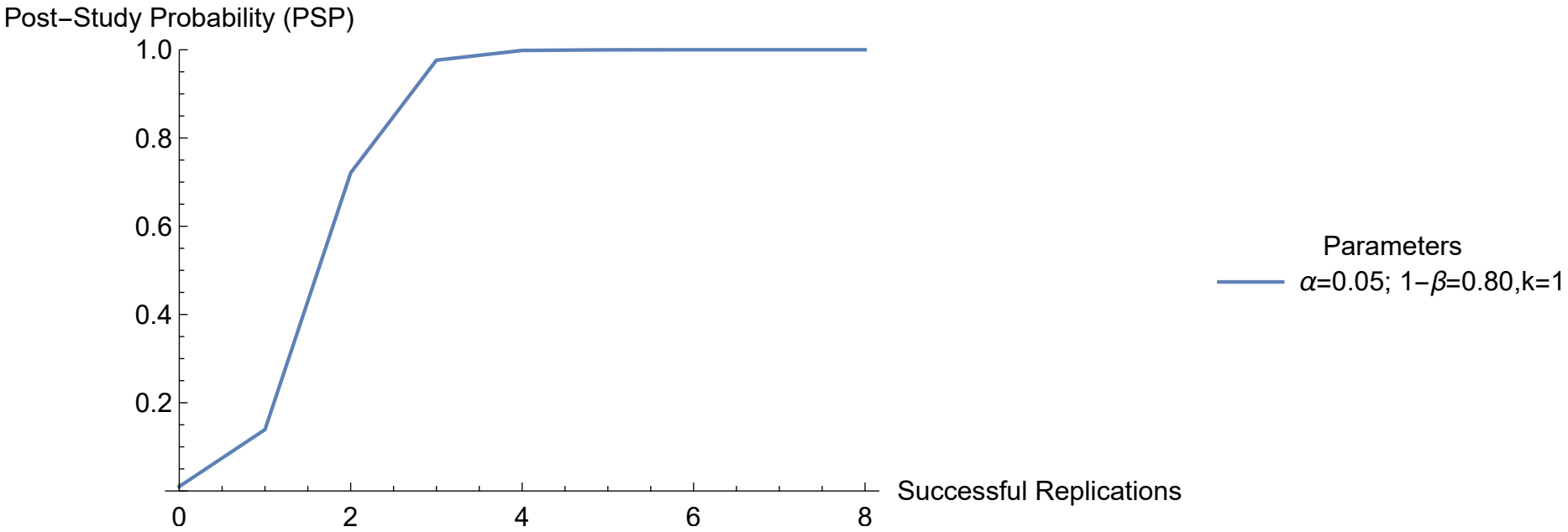
Replication as a Solution



Starting prior: $\pi = 1\%$. $\downarrow$ statistical significance $\alpha \Longrightarrow \uparrow$ PSP

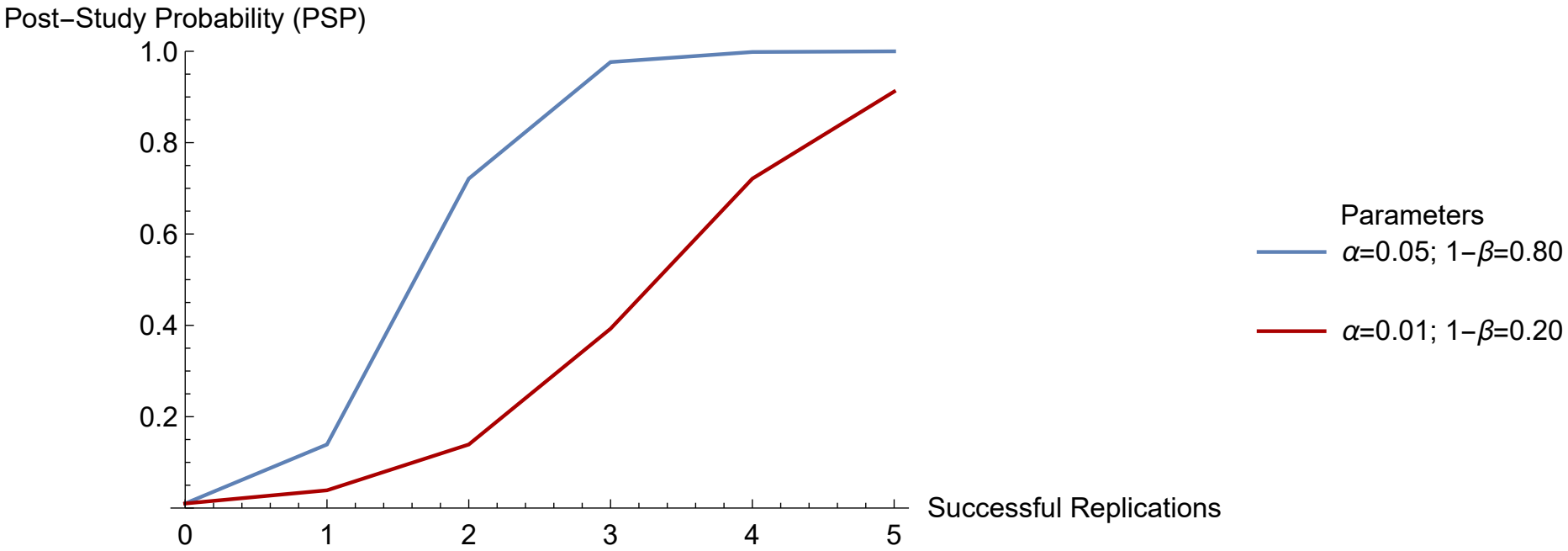
But perverse incentives for achieving low α.

Replication as a Solution



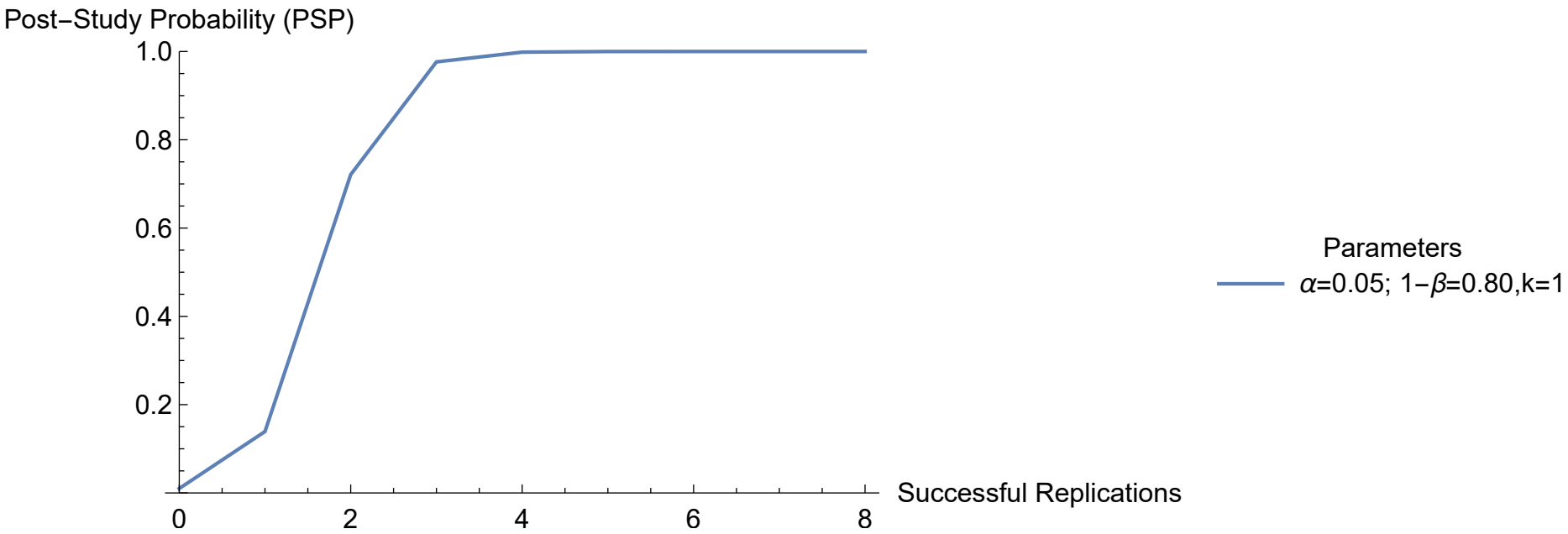
Starting prior: $\pi = 1\%$.

Replication as a Solution



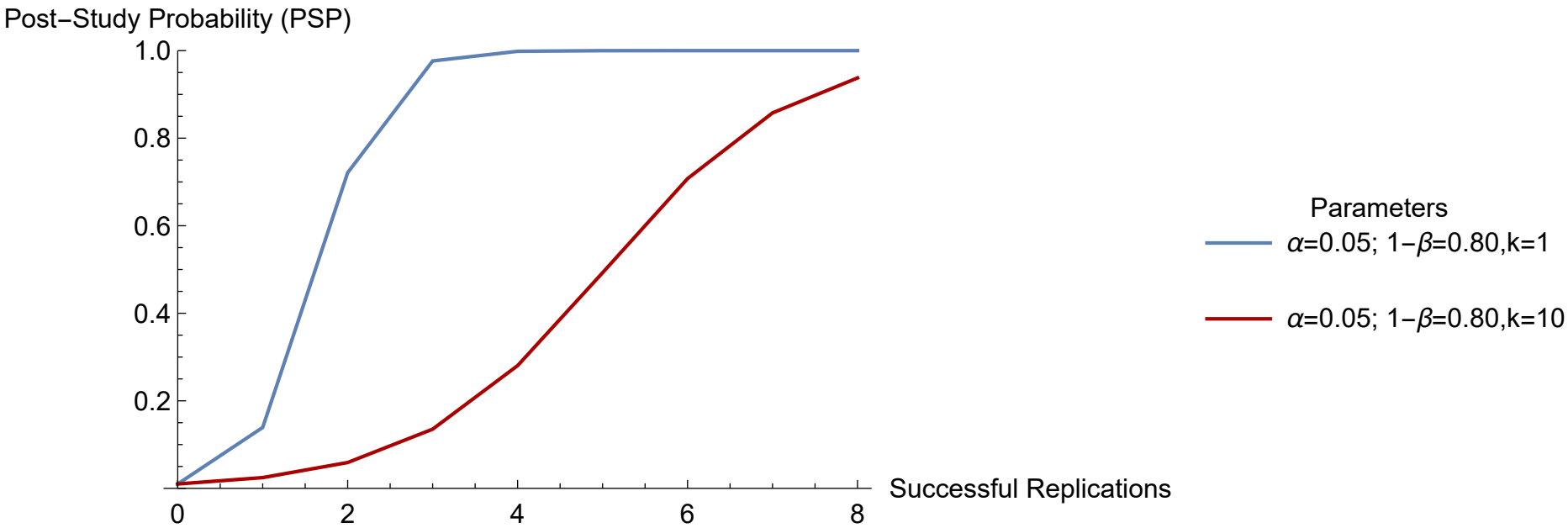
Starting prior: $\pi = 1\%$. $\downarrow$ statistical power $1 - \beta \Longrightarrow \downarrow$ PSP

Replication as a Solution



Starting prior: $\pi = 1\%$.

Replication as a Solution



Starting prior: $\pi = 1\%$. $\uparrow$ researchers $k \Longrightarrow \downarrow$ PSP

Replication as a Solution

- Reproductions are relatively rare in economics.
  Berry et al. (2017) Hamermesh (2017)

- Incentives are not well-aligned.
  Duvendack et al. (2017) Gertler et al. (2018)

Replication as a Solution

- Changing journal policies. Christensen and Miguel (2018)
  - Many require code and data online.
  - Some replicate results before publication e.g. *JAE*, *AJPS*.
  - *JDE* allows for registered reports approved before analysis.
  - Some have explicit sections for replication e.g. *JAE*.
  - Some have special replication issues e.g. *AEPP* 2022.

Replication as a Solution

- Changing journal policies. Christensen and Miguel (2018)
  - Many require code and data online.
  - Some replicate results before publication e.g. *JAE*, *AJPS*.
  - *JDE* allows for registered reports approved before analysis.
  - Some have explicit sections for replication e.g. *JAE*.
  - Some have special replication issues e.g. *AEPP* 2022.
- Replication wiki platform: https://replication.uni-goettingen.de/
- Replication journal with specific guidelines: *International Journal for Re-Views in Empirical Economics*.
- BITTS' Accelerating Computational Reproductibility in the Social Sciences initiative.

Averting the Replication Crisis

Ferraro and Shukla (2020)

- Emphasize questions rather than results.
- Emphasize designs rather than results.
- Specify research plan through preregistration.
- Run replications.
- Report results (and data) with transparency.

Replicability in Practice

Replicability in Practice

- Replication recipes.
- Workflow.
- Coding.
- Documenting.
- Reporting.

Replication recipes

Replication Recipes

A few guides to conducting replications.

- Define replication. Clemens (2015)
- Make a precise preanalysis plan. Chang (2018)
- Define replication paradigm and space. Anderson (2017)
- The problem of statistical adequacy. Owen (2018)
- Diagnostic approach for replications. Brown and Wood (2018)
- General guidelines in *International Journal for Re-Views in Empirical Economics*.

Workflow

Workflow

Orozco et al. (2020) Gentzkow and Shapiro (2014)

- A lot of research process happens in background.
- Systematic practices are key for reproducible research.
- Three general principles:
  - Organize your work.
  - Code for others.
  - Automate as much as possible.

Organize Your Work: Tasks

- Very helpful when working in teams.
- Organize and track progress and tasks.

Organize Your Work: Tasks

- Very helpful when working in teams.
- Organize and track progress and tasks.
- Simple text file, or electronic laboratory notebook (ELN) to synchronize across collaborators e.g. Evernote, OneNote, Etherpad.
- Task management systems (TMS) for more general collaborative and centralized task organization e.g. Trello, Asana, MS Project, Wrike.

Organize Your Work: Files

- Important to be consistent with:
  - Directory structure.
  - Naming convention.

Organize Your Work: Files

- Important to be consistent with:
  - Directory structure.
  - Naming convention.
- Directory structure:
  - No perfect solution, but project should have distinct folders.
  - Ideas and notes, Data, Analysis, Paper, Literature.
- My usual structure:
  - notes
  - literature
  - source_data
  - source_code (data.do, analysis.do, temp_files)
  - output
  - paper

Organize Your Work: Files

- Shared working spaces:
  - Necessary for any collaborative work.
  - Dropbox is most convenient (synchronization on computer).
  - Other options: Google drive, Joomla, Agora, SharePoint.
  - To share files: We Transfer, JustBeamIt, FileSender (RENATER).

Organize Your Work: Writing

- Collaborative writing can be tricky.
- Depends on writing software:
  - WYSIWYG (What You See Is What You Get) e.g. MS Word. Use Google doc.
  - WYSIWYM (What You See Is What You Mean) e.g. Latex. Use OverLeaf.
- References:
  - Reference managers: Mendeley, Zotero.
  - Use BibTex for Latex.

Coding

Coding Principles

- Code for others:
  - Your future self.
  - Your TAs and professors.
  - Other researchers.

Coding Principles

- Code for others:
  - Your future self.
  - Your TAs and professors.
  - Other researchers.
- Your code should be usable by anyone.

Coding Principles

- Code for others:
  - Your future self.
  - Your TAs and professors.
  - Other researchers.
- Your code should be usable by anyone.
- It should compile *entirely* with a push of a button (though you can divide data construction and analysis).
- **Only thing to change: directory and globals.**

Coding Principles

- Code for others:
  - Your future self.
  - Your TAs and professors.
  - Other researchers.
- Your code should be usable by anyone.
- It should compile *entirely* with a push of a button (though you can divide data construction and analysis).
- **Only thing to change: directory and globals.**
- Make it as automated, generic, and clear as possible.

Coding Principles

- Some guides:
  - Pep 8 Style Guide
  - Julian Reif's Stata Coding Guide.
  - Social Science Data Editors guide.

General Coding Conventions

- Limit the length to 80 characters (there is a vertical line).
- To jump line, use `///` at length 80.
- For comments, use `/*...*/`
- Have a space around mathematical operators ($=$, $+$, $-$, $>$, $<\ldots$)
- Make layout structure as clear as possible.
- Document code with sections using e.g. `* I. CLEAN DATASET`
- Use consistent variables naming convention.
- DRY: Don't Repeat Yourself.

Example

- Illustrate coding principles through example (on Moodle).
- Data based on Boehnke and Gay (2020)
  - Study in Open Access: 10.3368/jhr.57.4.0419-10151R1.
  - Replication data: 10.7910/DVN/AP1HZ8.
  - Impact of WWI military fatalities on FLFP in France using DiD.
  - Panel of 90 départements from 1901 to 1936.

Example

- Illustrate coding principles through example (on Moodle).
- Data based on Boehnke and Gay (2020)
  - Study in Open Access: 10.3368/jhr.57.4.0419-10151R1.
  - Replication data: 10.7910/DVN/AP1HZ8.
  - Impact of WWI military fatalities on FLFP in France using DiD.
  - Panel of 90 départements from 1901 to 1936.
- Steps:
  - Setting up the `.do` file.
  - Cleaning the (fake) raw dataset.
  - Run the analysis.

Document Your Preamble

- Many different ways, this is the spirit.

Document Your Preamble

- Many different ways, this is the spirit.
- File information: title, coder, date of creation, date and author of last edits. . .

Document Your Preamble

- Many different ways, this is the spirit.
- File information: title, coder, date of creation, date and author of last edits...
- Program setup:
  - Stata version: `version 16`. Avoid compatibility issues.
  - Clear memory: `clear all`.
  - No stop when output window full: `set more off`.
  - Line size if create TeX tables in Stata: `set linesize 82`
  - System diagnostic (optional): version, variant

Document Your Preamble

- Relative paths (very important!).
  - Directory.
  - Input and output globals.
  - You can add a permanent global (see Julian Reif's guide).
- For reproducibility, create relevant folders in replication file.

Coding Principles: Document Your Preamble

```
*************************************************************************
* M1 APPLIED ECONOMETRICS, FALL 2021                                    *
* AUTHOR: VICTOR GAY                                                    *
* CREATED: OCTOBER 5 2021                                               *
* LAST MODIFIED: OCTOBER 5 2021                                         *
* LAST MODIFIED BY: VICTOR GAY                                          *
*************************************************************************

*****************
* PROGRAM SETUP *
*****************

version 16
clear all
set more off
set linesize 82

*************
* DIRECTORY *
*************

global PATH "ENTER YOUR DIRECTORY HERE"
cd "$PATH/source_code"

***********
* GLOBALS *
***********

global DATA "$PATH/source_data"
global OUTPUT "$PATH/output"
global TEMP "$PATH/source_code/temp_files"
```

Data Cleaning

- Check the format of your data.
  - Use `codebook`.
  - Directly see the data using `edit` (strings are red, numerical are black).
  - Transform to appropriate format e.g. `destring year, replace`.

Data Cleaning

- Check the format of your data.
  - Use `codebook`.
  - Directly see the data using `edit` (strings are red, numerical are black).
  - Transform to appropriate format e.g. `destring year, replace`.
- Label your data.
  - Name and label appropriately your variables.

Data Cleaning

- Check the format of your data.
  - Use `codebook`.
  - Directly see the data using `edit` (strings are red, numerical are black).
  - Transform to appropriate format e.g. `destring year, replace`.
- Label your data.
  - Name and label appropriately your variables.
  - Use value labels for categorical variables, then tabulate with `fre`.
  - You might need to install it.
  - Data with value labels are in blue in the dataset.
  - Generally, avoid strings. But sometimes no choice.

Data Cleaning

```
*******************************************************************************
* I. CLEAN DATASET
*******************************************************************************

** I.1. LOAD DATA

use "$DATA/dataset_raw", clear

** I.2. FORMAT: STRINGS AND NUMERICAL VARIABLES

codebook year
*edit
destring year, replace

** I.3. VARIABLE NAMES AND LABELS

codebook

rename flfp_temp flfp
label variable flfp "Female labor force participation rate"
codebook flfp
```

Data Cleaning

- Some départements in Germany before WWI.

```
** I.4. VALUE LABELS

codebook sample
ssc install fre
fre sample

/* generate new variable */
generate dep_type = .
replace  dep_type = 1 if sample == "new"
replace  dep_type = 2 if sample == "non-war" | sample == "war"

/* use value labels */
label define dep_type 1 "New département" 2 "Old département"
label values dep_type dep_type
label values dep_type dep_type
fre dep_type
order dep_type, after(sample)
drop sample
```

Data Cleaning

- Document missing values!
- There should be no undocumented missing values.
- Use extended missing values rather than ad hoc values.
- Assign value labels to document missings.
- Here: no military death rates in "old" départements.

```
** I.5. EXTENDED MISSING VALUES

codebook deathrate
replace deathrate = .a if deathrate == .
label define deathrate .a "New département"
label values deathrate deathrate
codebook deathrate
fre deathrate if deathrate >= .
```

Data Cleaning

- Summarize all your data and check units.
- In particular, check all are expressed in same unit.
- Usually: careful with shares (0–1) and percent (0–100).
- Matters (a lot!) for the interpretation.
- Adjust your variable labels.

```
** I.6. CHECK UNITS

summarize deathrate flfp rural bindep
replace flfp = flfp * 100
label variable flfp "Female labor force participation (%)"
codebook deathrate flfp bindep

save "$TEMP/dataset_temp", replace
```

Data Cleaning

- Sometimes the data is in wide form, but you want long form.
- Learn how to `reshape`.
- Don't hesitate to use `help reshape`!

```
** I.7. RESHAPE

use "$DATA/dataset_wide", clear
help reshape
reshape long deathrate, i(dep_id) j(year)
```

Data Cleaning

- Sometimes, you need to merge datasets (here: add `rural`).
- Learn how to use `merge`
- Never use `m:m`, always `1:m` or `1:1` or `1:m`.
- Always use `assert` for no surprise.
- Use `keep` as needed after having checked `_merge`.

  $\Longrightarrow$ Btw: use `assert` extensively for integrity tests.

```
use "$TEMP/dataset_temp", clear
merge 1:1 dep_id year using "$DATA/dataset_rural", assert(1 3)
fre dep_type if _merge == 1
drop if dep_type == 1
assert(_merge == 3)
drop _merge

erase "$TEMP/dataset_temp.dta"
* alternative: merge 1:1 dep_id year using "$DATA/dataset_rural", assert(1 3) keep(3)
```

Coding Principles: Data Cleaning

- Looping over variables is useful.
- Use a `foreach` loop.
- In this example, I also `preserve` and `restore` the data.

```
** I.9. LOOPS

/* loop over variables and transform percent into shares */
preserve

foreach v of varlist flfp deathrate bindep rural {
summarize ‘v’
replace ‘v’ = ‘v’ / 100
summarize ‘v’
display "Now ‘v’ is a share"
}
restore
```

Coding Principles: Data Cleaning

- Looping over variables is useful.
- Use a `foreach` loop.
- In this example, I also `preserve` and `restore` the data.

```
** I.9. LOOPS

/* loop over variables and transform percent into shares */
preserve

foreach v of varlist flfp deathrate bindep rural {
summarize ‘v’
replace ‘v’ = ‘v’ / 100
summarize ‘v’
display "Now ‘v’ is a share"
}
restore
```

- You can also loop over values: `forvalues i = 1 / 10 ...`

Coding Principles: Data Cleaning

- Looping over variables is useful.
- Use a `foreach` loop.
- In this example, I also `preserve` and `restore` the data.

```
** I.9. LOOPS

/* loop over variables and transform percent into shares */
preserve

foreach v of varlist flfp deathrate bindep rural {
summarize ‘v’
replace ‘v’ = ‘v’ / 100
summarize ‘v’
display "Now ‘v’ is a share"
}
restore
```

- You can also loop over values: `forvalues i = 1 / 10 ...`.
- Sometimes it is very useful to loop of all values of a variable.

```
levelsof year, local(years)
foreach y of local years {
display ‘y’
}
```

Coding Principles: Analysis

- Declare your panel structure using `xtset unit time`.
- Unit should be numeric. If string: `encode`.
- This will help with fixed effects regressions.

```
encode dep_id, generate(dep_id1)
order dep_id1, after(dep_id)
xtset dep_id1 year
```

Coding Principles: Analysis

- Declare your panel structure using `xtset unit time`.
- Unit should be numeric. If string: `encode`.
- This will help with fixed effects regressions.

```
encode dep_id, generate(dep_id1)
order dep_id1, after(dep_id)
xtset dep_id1 year
```

- Use locals to make code readable, e.g. for controls.

Coding Principles: Analysis

- Declare your panel structure using `xtset unit time`.
- Unit should be numeric. If string: `encode`.
- This will help with fixed effects regressions.

```
encode dep_id, generate(dep_id1)
order dep_id1, after(dep_id)
xtset dep_id1 year
```

- Use locals to make code readable, e.g. for controls.
- No need for `xi:...` anymore: use operator `i.`.
- Use `i.categorical#c.continuous` for interactions.
- If no need for FE estimates, absorb with `reghdfe`.

Coding Principles: Analysis

- Example from Boehnke and Gay (2020).
- Impact of WWI military fatalities on FLFP in France (1901–1936).
- Empirical strategy: difference-in-differences.

$$\texttt{FLFP}_{d,t} = \beta\, \texttt{death\_rate}_d \times \texttt{post}_t + \gamma_d + \delta_t + \varepsilon_{d,t}$$

  - $\texttt{post}_t$ = indicator for $t > 1918$.
  - γ_d = department fixed effects.
  - δ_t = year fixed effects.

- Standard errors clustered at département level.
- Censuses: 1901, 1906, 1911, 1921, 1926, 1931, 1936.

Coding Principles: Analysis

- Many ways to run fixed-effects regressions:
  - `regress`
  - `xtreg`
  - `areg`
  - `reghdfe`
- All provide same point estimates.
- `xtreg` and `reghdfe` adjust s.e. when FE nested within clusters.

```
global controls "rural bindep"

regress flfp deathrate_post $controls i.year i.dep_id, cluster(dep_id)
areg flfp deathrate_post $controls i.year, a(dep_id) cluster(dep_id)
xtreg flfp deathrate_post $controls i.year, fe cluster(dep_id)
reghdfe flfp deathrate_post $controls i.year, absorb(dep_id) cluster(dep_id)
```

Coding Principles: Analysis

- Many ways to run fixed-effects regressions:
  - `regress`
  - `xtreg`
  - `areg`
  - `reghdfe`
- All provide same point estimates.
- `xtreg` and `reghdfe` adjust s.e. when FE nested within clusters.

```
global controls "rural bindep"

regress flfp deathrate_post $controls i.year i.dep_id, cluster(dep_id)
areg flfp deathrate_post $controls i.year, a(dep_id) cluster(dep_id)
xtreg flfp deathrate_post $controls i.year, fe cluster(dep_id)
reghdfe flfp deathrate_post $controls i.year, absorb(dep_id) cluster(dep_id)
```

- Generally: use `xtreg`.
- Prefer `reghdfe` if multi-way clustering and/or IV and/or many FE.

```
ssc install reghdfe
ssc install ftools
```

Coding Principles: Analysis

- Always cluster at some appropriate level.
- `reghdfe` allows easy adjustments.
- For a difference-in-differences, cluster over time and space.

```
reghdfe flfp deathrate_post ‘controls’ i.year, absorb(dep_id) cluster(year)
reghdfe flfp deathrate_post ‘controls’ i.year, absorb(dep_id) cluster(dep_id year)
```

Coding Principles: Analysis

- Always cluster at some appropriate level.
- `reghdfe` allows easy adjustments.
- For a difference-in-differences, cluster over time and space.

```
reghdfe flfp deathrate_post ‘controls’ i.year, absorb(dep_id) cluster(year)
reghdfe flfp deathrate_post ‘controls’ i.year, absorb(dep_id) cluster(dep_id year)
```

- Export results using `estab` or `estout`.
- Alternative: use `parmest` to create datasets of estimates, then construct your personalized Latex table in your `.do` file.

Documenting

Documenting

- Document so that your work can be reproduced by others.
  - Write enough details in the main text.
  - Put exact details in appendix.
  - Exact procedures are in the code.
- Some necessary elements:
  - Data sources (be exact).
  - Sample selection procedures.
  - Data transformations.

Reporting

Reporting

- What is the hypothesis you want to test?
- If not transparent from text, write it down e.g. $H_1 : \overline{X} \neq 0$.

Reporting

- What is the hypothesis you want to test?
- If not transparent from text, write it down e.g. $H_1 : \overline{X} \neq 0$.
- What is your design?
- What is your exogenous source of identifying variation?
- What is the identifying assumption?

Reporting

- What is the hypothesis you want to test?
- If not transparent from text, write it down e.g. $H_1 : \overline{X} \neq 0$.
- What is your design?
- What is your exogenous source of identifying variation?
- What is the identifying assumption?
- If proper design, no need for controls except those determinant for assignment to treatment.
- Therefore: no specification search or *p*-hacking!
- Only show coefficients of interest.

Reporting

- Always interpret the estimates.
- What do they say exactly?
- What about their magnitude? Think in terms of means and/or standard deviations of your outcome and treatment.
- Is it large? Useful to use Cohen's d.
  - Relative size of coefficient relative to sample standard deviation in treatment, normalize by sample standard deviation in outcome.
  - $\widehat{\beta} \times \frac{\widehat{\sigma}_X}{\widehat{\sigma}_Y}$

A Reproducible Project

- Final product should contain one folder with various elements.

A Reproducible Project

- Final product should contain one folder with various elements.
- Readme file (`.txt`) containing
  - List of all subfolders and their content.
  - System-software combination used in analysis, processing time.
  - Order in which compile (e.g. data then analysis).

A Reproducible Project

- Final product should contain one folder with various elements.
- Readme file (`.txt`) containing
  - List of all subfolders and their content.
  - System-software combination used in analysis, processing time.
  - Order in which compile (e.g. data then analysis).
- Paper

A Reproducible Project

- Final product should contain one folder with various elements.
- Readme file (`.txt`) containing
  - List of all subfolders and their content.
  - System-software combination used in analysis, processing time.
  - Order in which compile (e.g. data then analysis).
- Paper
- Appendix with details on robustness and data:
  - Data sources.
  - Transformation of raw data.
  - Variables definitions.
  - Example: see data appendix of Boehnke and Gay (2020)

A Reproducible Project

- Raw data.
- Code to go from raw to final data (in 1 click!).
- Code to run the analysis (in 1 click!).
- The only variable element is the directory and globals.