The Null Result Penalty

Felix Chopra[1] Ingar Haaland [2] Chris Roth[3] Andreas Stegmann[4]

[1]University of Bonn
[2]University of Bergen
[3]University of Cologne
[4]University of Warwick

June 2022

Scientific Discovery in the Presence of a Null Result Penalty

- The scientific method is characterized by researchers testing hypotheses with empirical evidence (Popper, 1934).
  - Evidence accumulates with the publication of studies in scientific journals.
- Scientific progress thus requires a well-functioning publication system that evaluates research studies without bias.
- Publication system may favor studies with large and statistically significant results over papers documenting small results that are not statistically significant.
- Such selectivity can lead to **biased estimates** and misleading confidence sets in published studies (Andrews and Kasy, 2019).

Research questions

- Is there a penalty for null results?
- What mechanisms drive such a penalty?
  - What is the role of the communication of statistical uncertainty?
  - Are surprising null results more publishable?
  - Is the null result penalty a bias potentially arising partly due to errors in statistical reasoning?

Identification challenge

- Studies that yield null results might be different from studies that have non-null results both in terms of observables and unobservables.
  - E.g. null result could reflect the unobserved quality of execution.
  - It may be rational to believe that a null result study is of lower quality.
- We circumvent this identification challenge in several ways:
  - ✓ Our hypothetical vignettes hold constant all details of the study other than the result (null result versus significant result).
  - ✓ We also fix beliefs about the standard error of the estimate.

What we do

- Large-scale surveys with academic economists and PhD students.
- **Hypothetical vignettes**: We vary whether a study has a **null result** or a **significant result**, holding all other study features constant.
- **Expert predictions**: We vary whether respondents receive an expert forecast (either high or low forecast).
- **Obfuscation treatments**: We cross-randomize other study features (rank of university, seniority of authors to **obfuscate** the study purpose.
- **Mechanism experiment**: Measure perceived statistical precision of main finding, while providing all respondents with information about the precision of the main finding.

What we find

- Studies with null results are perceived to be **less publishable** than studies with significant results.
- Null results studies are also perceived to be **worse in other dimensions** (e.g. quality, importance and statistical precision)
  - Our finding on perceived precision suggests some role for errors in statistical reasoning in explaining the null result penalty.
- The null result penalty is of **similar magnitude** among PhD students and journal editors.
- The penalty is larger when...
  - experts predict a large effect.
  - when statistical uncertainty is communicated with p-values rather than standard errors.

Related literature

- Literature on on **publication bias** (Brodeur et al., 2021, 2016, 2020) and correction methods (Andrews and Kasy, 2019).
  - We contribute to this literature by studying mechanisms underlying publication bias.

- Descriptive literature on the beliefs and reasoning of **experts** (Andre and Falk, 2021; Andre et al., 2022; DellaVigna and Pope, 2018a,b; DellaVigna et al., 2019)

Outline of talk

1. Design
2. Main Results
3. Mechanism Experiment
4. Conclusion and Implications

Overview of design

- Hypothetical vignettes on research studies, providing details on the research question, study design and findings.
- Vignettes cover studies in various field (labor, development, history, behavioral) and using different methods (RCT, RDD, online experiment).
  - This allows us to hold fixed all study features while varying whether the main result of the study is a **null result** or not.
  - **Expert variation**: We vary whether respondents receive an expert forecast (either high or low forecast).
  - **Obfuscation treatments**: We cross-randomize other study features (rank of university, seniority of authors to **obfuscate** the study purpose.

Treatment arms

- **Null-result treatment**: We state the coefficient estimate of the causal effect as well as the standard error. Vary only the number of coefficient.

- **Obfuscation treatments**:
  - **Elite university treatment**: We mention the affiliation of the researchers. For half respondents those are elite institutions (Harvard, MIT, Berkeley, etc), for others good but not elite universities (Arizona State, University of Florida, etc)
  - **Seniority treatment**: We vary whether the study was conducted by PhD students vs. professors

- **Expert forecasts**: 50% of respondents do not receive information about expert forecasts.
  - Among the remaining respondents, half get a large expert prediction, while the other half gets an expert prediction which is close to zero.

Vignette: Female Empowerment Program

Female empowerment program

Background and study design: In 2018, a team of 4 PhD students from Columbia University conducted an RCT in Sierra Leone. The purpose of the RCT was to examine whether access to a female empowerment program increased women's labor supply.

In the RCT, 360 women were evenly randomized into a treatment group and a control group. Respondents in the treatment group were offered a female empowerment program, combining both psychosocial therapy and vocational skills training. The program was very intensive: participants attended meetings for up to 5 hours every day during a 12-month period.

Main result of the study: Treated respondents were 1.7 percentage points (standard error 5.0) more likely to take up a job offer compared to a control mean of 37.0 percent.

Expert prediction: 34 experts in this literature received the control mean and the same background and study design information as shown to you above. The experts on average predicted a treatment effect of 0.6 percentage points. The standard deviation of the expert

Vignette: Poverty and Patience

Salience of poverty and patience

Background and study design: In 2021, a team of 2 PhD students from UC Berkeley conducted an experiment on an online survey platform. The purpose of the experiment was to examine whether financial anxieties increase people's inclination to make more impatient choices.

800 US respondents were evenly randomized into a treatment and control group. Respondents were asked to write a few sentences about how they would raise $5,000 (treatment group) or $50 (control group) to cover an unexpected expense. The main outcome of interest was whether respondents choose to receive $100 now or $110 in a week. The choices were implemented for 25% of respondents.

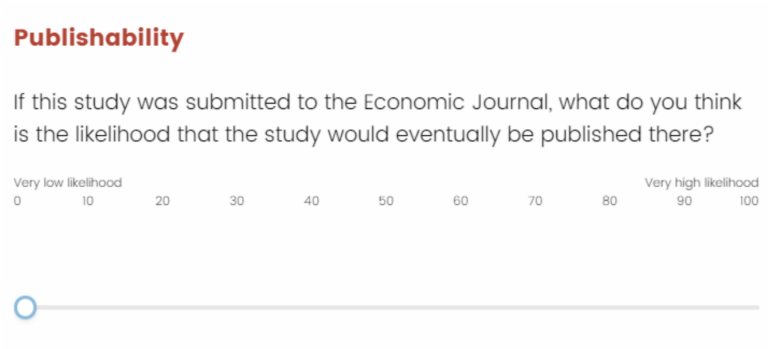
The treatment increased respondents' financial anxieties by 29.1 percent of a standard deviation.

Main result of the study: Treated respondents were 7.8 percentage points (standard error 3.5) more likely to choose money now

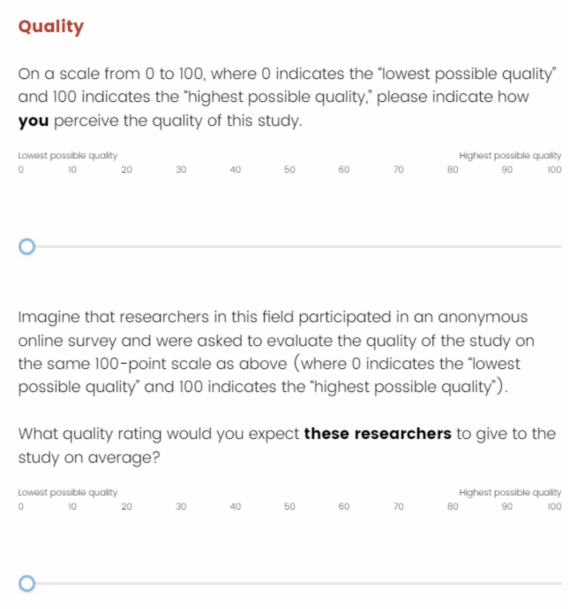
Overview of Outcomes

- We elicit a series of beliefs about the research paper.
  - Main outcome: Perceived likelihood of publication
  - Perceived quality of design (fob + sob)
  - Perceived importance of study (fob + sob)
- We randomize at the respondent-level whether we elicit perceptions of quality or importance to economize on survey time.

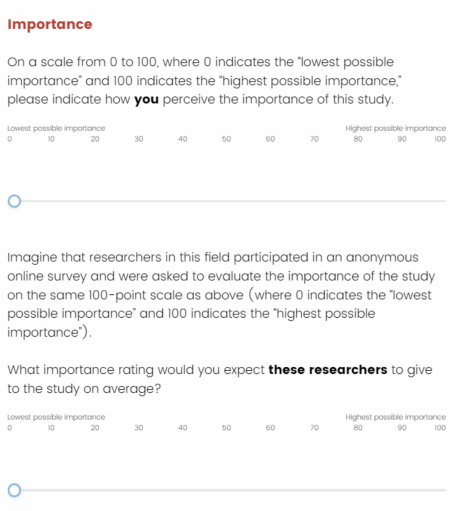
Main outcomes: Publishability



Main outcomes: Quality



Main Outcomes: Importance



Sample

In April and May 2022, we invited researchers in the field of economics affiliated with one of the top 200 institutions according to RePEc to participate in a 10-minute online survey:

- About **500** academic economists completed our online survey.
- Our sample is mostly comprised of **experienced researchers** with **substantial academic impact** in the field of economics.
  - On average, 1.3 research articles published in one of the "top five" economics journals.
  - Their work is also highly cited: average (median) h-index among respondents with a Google Scholar profile of 17.2 (11.5).
- Our respondents also have experience in different subfields of economics

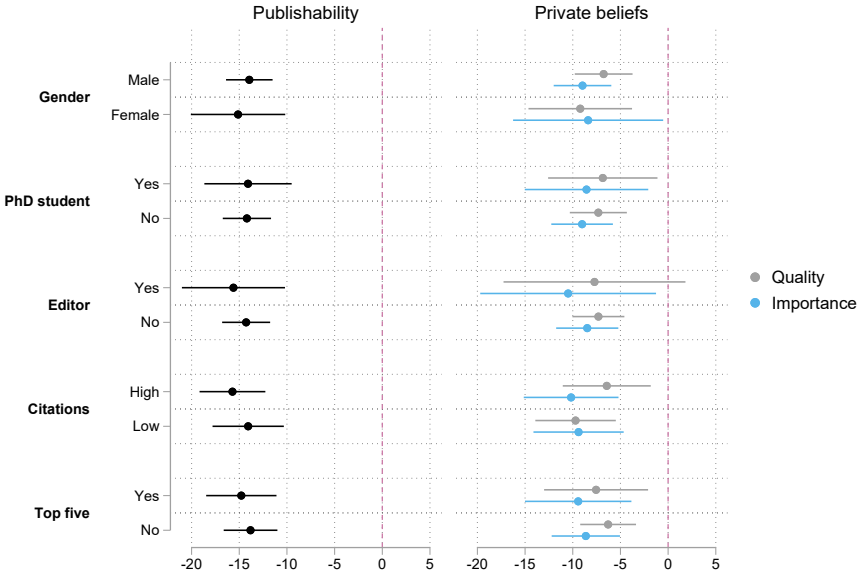
Outline of talk

1. Design
2. Main Results
3. Mechanism Experiment
4. Conclusion and Implications

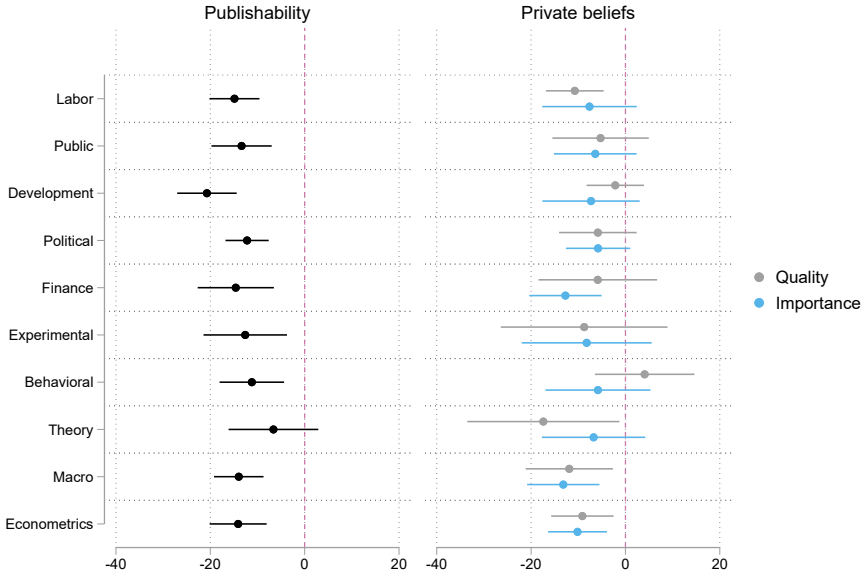
Main results

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Panel A: Fixed effects					
Null result treatment	-14.058***	-0.373***	-0.460***	-0.325***	-0.417***
	(1.090)	(0.062)	(0.062)	(0.054)	(0.056)
Panel B: No individual FE					
Null result treatment	-14.474***	-0.401***	-0.455***	-0.305***	-0.367***
	(1.224)	(0.069)	(0.072)	(0.062)	(0.069)
Observations	1,920	920	920	1,000	1,000

Heterogeneity across subgroups



Heterogeneity across fields



Heterogeneous treatment effects

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Panel A: Fixed effects					
Null result treatment	-11.072***	-0.029	-0.219	-0.330**	-0.390***
	(2.681)	(0.151)	(0.160)	(0.132)	(0.135)
Null result × Low expert forecast	-1.862	-0.169	0.130	0.030	0.058
	(2.470)	(0.162)	(0.159)	(0.120)	(0.117)
Null result × High expert forecast	-6.251**	-0.083	0.033	0.048	-0.025
	(2.632)	(0.165)	(0.152)	(0.124)	(0.127)
Null result × P-value framing	-3.652*	-0.344***	-0.362***	-0.021	0.049
	(2.164)	(0.122)	(0.120)	(0.109)	(0.112)
Observations	1,920	920	920	1,000	1,000

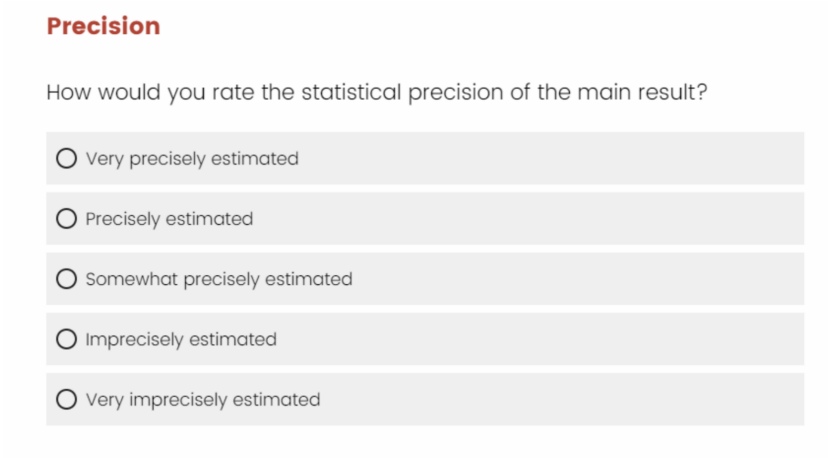
Outline of talk

1. Design
2. Main Results
3. Mechanism Experiment
4. Conclusion and Implications

Overview

- Do researchers perceive studies with null results to be **less precisely estimated**, even when they are provided with the standard error of the estimate?
- **Sample**: Graduate students and early career researchers.
- **Design**: Identical to our main experiment except for two differences.
  - Respondents rate the **statistical precision** of the main result instead of perceived quality and importance of the study.
  - Respondents are shown all five vignettes.

Perceived Prevision



Results

	(1) Publishability (in percent)	(2) Precision (z-scored)
Panel A: Fixed effects		
Null result treatment	-19.755*** (2.269)	-1.267*** (0.144)
Panel B: No individual FE		
Null result treatment	-18.134*** (2.605)	-1.086*** (0.148)
Observations	475	475

- Beliefs about the precision are influenced by the coefficient's statistical significance, even though standard errors are identical.

→ This suggest some role for **errors in statistical reasoning**.

Outline of talk

1. Design
2. Main Results
3. Mechanism Experiment
4. Conclusion and Implications

Conclusion

- We show that research studies with null results are perceived to be less publishable, of lower quality, of lower importance, and less precisely estimated.
  - holding constant all other study features, including the statistical precision of estimates.
  - The finding on perceived precision suggests a role for errors in statistical reasoning.
- The null result penalty is even larger when experts predict a non-null result.
- Communicating the statistical uncertainty of study results with p-values rather than standard errors further increases the null result penalty

Implications

- Our findings highlight the **potential value of pre-results review** in which the decision on publication is taken before the empirical results are known (Kasy, 2021; Miguel, 2021).

- Our results also suggest that journals should **provide referees with additional guidelines** on the evaluation of research by highlighting the informativeness of null results (Abadie, 2020).

- Communicating statistical uncertainty of estimates in terms of **standard errors rather than p-values** might help to counteract potential errors in statistical reasoning.

References

Abadie, Alberto, "Statistical nonsignificance in empirical economics," *American Economic Review: Insights*, 2020, *2* (2), 193–208.

Andre, Peter and Armin Falk, "What's worth knowing? Economists' opinions about economics," Technical Report, ECONtribute Discussion Paper 2021.

_ **, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart**, "Subjective Models of the Macroeconomy: Evidence from Experts and Representative samples," *The Review of Economic Studies*, 2022.

Andrews, Isaiah and Maximilian Kasy, "Identification of and correction for publication bias," *American Economic Review*, 2019, *109* (8), 2766–94.

Brodeur, A, S Carrell, D Figlio, and L Lusher, "Unpacking P-hacking and Publication Bias," Technical Report, Tech. rep 2021.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, 2016, *8* (1), 1–32.

_ **, Nikolai Cook, and Anthony Heyes**, "Methods matter: P-hacking and publication bias in causal analysis in economics," *American Economic Review*, 2020, *110* (11), 3634–60.

References (cont.)

DellaVigna, Stefano and Devin Pope, "Predicting experimental results: who knows what?," *Journal of Political Economy*, 2018, *126* (6), 2410–2456.

_ and _ , "What motivates effort? Evidence and expert forecasts," *Review of Economic Studies*, 2018, *85* (2), 1029–1069.

_ , _ , and Eva Vivalt, "Predict science to improve science," *Science*, 2019, *366* (6464), 428–429.

Kasy, Maximilian, "Of forking paths and tied hands: Selective publication of findings, and what economists should do about it," *Journal of Economic Perspectives*, 2021, *35* (3), 175–92.

Miguel, Edward, "Evidence on research transparency in economics," *Journal of Economic Perspectives*, 2021, *35* (3), 193–214.

Popper, Karl, *The logic of scientific discovery*, Routledge, 1934.