The Null Result Penalty

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

[1]University of Copenhagen
[2]Norwegian School of Economics
[3]University of Cologne
[4]University of Warwick

August 2023

Scientific discovery in the presence of a null result penalty

- The scientific method is characterized by researchers testing hypotheses with empirical evidence (Popper, 1934)
- Scientific progress requires a publication system that evaluates research w/o 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 bias meta-analytic estimates and CIs based on published studies
  - Selectivity could affect incentives to start, continue and submit research studies

Research questions

- Is there a perceived penalty against null results?
  - If so, does it differ across fields?
  - How editors of leading journals evaluate 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 arising due to errors in statistical reasoning?

Identification challenge: No two studies are really alike

- **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.
  - Studies with null results might have lower power to detect effects
- **Our approach**: Hypothetical vignettes

What we do

- Large-scale survey experiment with academic economists
- Hypothetical vignettes
  - Exogenously vary the statistical significance of the main result
  - Fix all other study characteristics, including the standard error
  - Measure perceived publishability prospects
- Study potential remedies for result-dependent evaluations
  - Expert forecasts
  - Communication of statistical uncertainty
- Mechanism experiment to study the role of errors in statistical reasoning

Related literature

- **Publication bias and correction methods**
  (Andrews and Kasy, 2019; Brodeur et al., 2021, 2016, 2020)

  → We study mechanisms underlying publication bias.

- **Editorial policies to promote research transparency and reduce publication bias**
  (Christensen and Miguel, 2018; Dufwenberg et al., 2014; Miguel et al., 2014)

  → We study the role of expert forecasts and the communication statistical uncertainty

- **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)

  → We study result-dependent perceptions of research studies

Outline of talk

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

Sample and logistics

- Pre-registered on AsPredicted (#95235 and #96599)
- **Sampling frame**: Economists at the top 200 institutions according to RePEc
- **Data collection**: April/May 2022
  - E-mail invitation to participate in a 10-minute survey
  - No reminder to reduce burden on respondents
- **Final sample**: 480 academic economists
  - Highly experienced and influential researchers
  - Diverse sample in terms of subfields of economics

Sample is more experienced than the overall researcher population

	Survey sample			Sampling population	
	Mean	Median	Obs.	Mean	Median
Demographics:					
Female	0.22		477	0.24	0
Years since PhD	14.81	11	308	16.09	13
PhD student	0.24		467		
Region of institution:					
Europe	0.54		478	0.36	0
North America	0.41		478	0.53	1
Australia	0.03		478	0.08	0
Asia	0.02		478	0.03	0
Academic output:					
H-index	17.22	11.5	328	8.83	5
Citations	4,348.34	846	328		
Number of top 5 publications	1.27		462	0.34	0
Number of top 5s refereed for	1.17		397		
Repeated top 5 referee	0.30		397	0.12	0
Research evaluation:					
Current editor	0.07		443	0.03	0
Current associate editor	0.13		441		
Ever editor	0.15		444		
Ever associate editor	0.19		441		

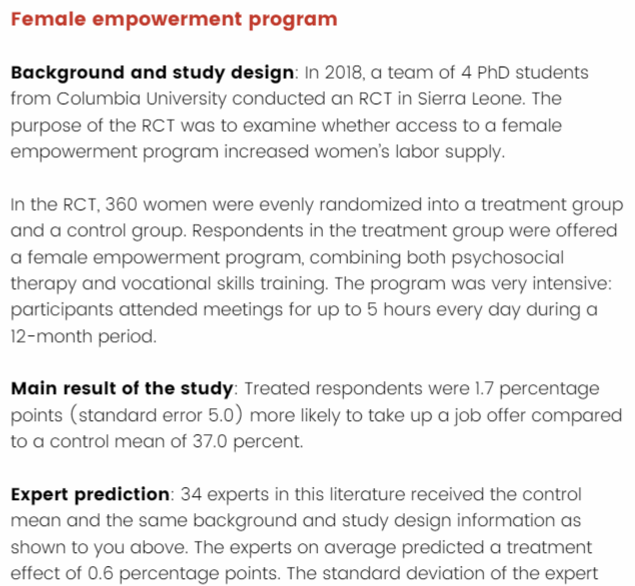
Overview of design

- **Hypothetical vignettes** on research studies
  - Details on the research question, study design and findings
  - Fix all study features while exogenously varying the statistical significance
  - Respondents evaluate 4 out of 5 vignettes
- **Within-subject variation** (across vignettes)
  - **Main treatment**: Vary effect size holding the standard error fixed
  - **Expert forecasts**: Vary whether respondents receive expert forecasts
  - **Obfuscation treatments**: Vary seniority and affiliation of authors
- **Between-subject variation**
  - Communicate statistical uncertainty via standard errors or p-values

Within-subject variation

- **Null-result treatment**: Vary only the coefficient estimate (high vs low)
- **Obfuscation treatments**
  - Elite university treatment: Author team is either affiliated with a top 5 institution (Harvard, MIT, Berkeley, etc) or not (Arizona State, University of Florida, etc)
  - Seniority treatment: Vary whether authors are PhD students or professors
- **Expert forecasts**
  - 50% of respondents: No expert forecast about the study's results
  - 25% of respondents: Experts predict large effect
  - 25% of respondents: Experts predict effect close to zero

Exaple 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

Example of a vignette

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 treatment

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 forecasts was 7.6.

Example of a vignette

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.

Obfuscation treatments

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 treatment

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 forecasts was 7.6.

Example of a vignette

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.

Obfuscation treatments

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 treatment

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 treatment

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 forecasts was 7.6.

Example of a vignette

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.

Obfuscation treatments

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 treatment

Between-subject: p-val vs SE

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 treatment

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 forecasts was 7.6.

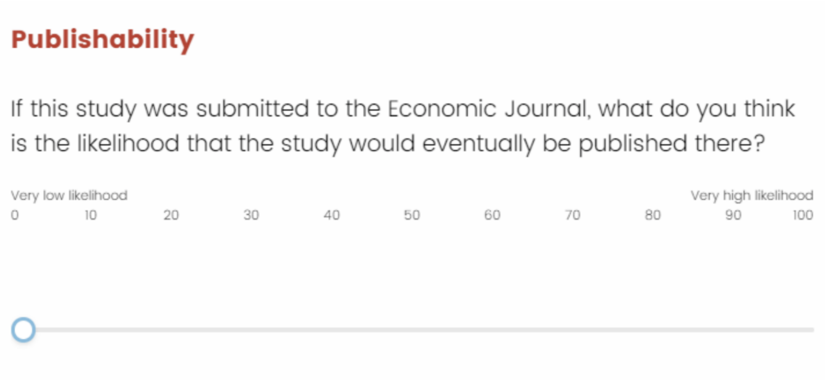
Other vignettes

- Marginal effects of merit aid for low-income students (RCT)
- Long-term effects of equal land sharing (RDD)
- Financial literacy program (RCT)
- Salience of poverty and patience (online experiment)

Overview of outcomes

- **Main outcome**: Perceived publishability
- **Perceived quality of the study**
  - First-order beliefs
  - Second-order beliefs
- **Perceived importance of the study**
  - First-order beliefs
  - Second-order beliefs
- **Cross-randomization**: 50% of respondents are asked about quality, while the other 50% are asked about importance

Perceived publishability



Quality of the study

Quality

On a scale from 0 to 100, where 0 indicates the "lowest possible quality" and 100 indicates the "highest possible quality," please indicate how **you** perceive the quality of this study.

Lowest possible quality — Highest possible quality

0 10 20 30 40 50 60 70 80 90 100

Imagine that researchers in this field participated in an anonymous online survey and were asked to evaluate the quality of the study on the same 100-point scale as above (where 0 indicates the "lowest possible quality" and 100 indicates the "highest possible quality").

What quality rating would you expect **these researchers** to give to the study on average?

Lowest possible quality — Highest possible quality

0 10 20 30 40 50 60 70 80 90 100

Importance of the study

Importance

On a scale from 0 to 100, where 0 indicates the "lowest possible importance" and 100 indicates the "highest possible importance," please indicate how **you** perceive the importance of this study.

Lowest possible importance — Highest possible importance
0 10 20 30 40 50 60 70 80 90 100

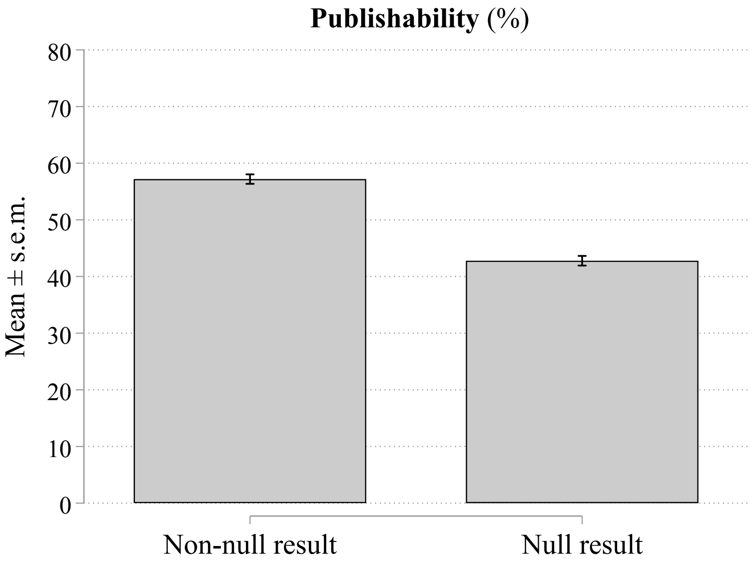
Imagine that researchers in this field participated in an anonymous online survey and were asked to evaluate the importance of the study on the same 100-point scale as above (where 0 indicates the "lowest possible importance" and 100 indicates the "highest possible importance").

What importance rating would you expect **these researchers** to give to the study on average?

Lowest possible importance — Highest possible importance
0 10 20 30 40 50 60 70 80 90 100

Outline of talk

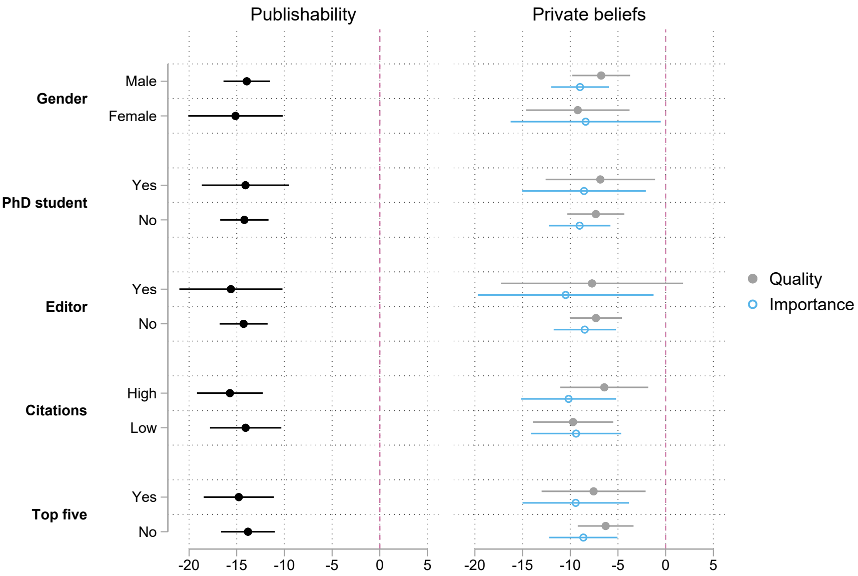
The null result penalty



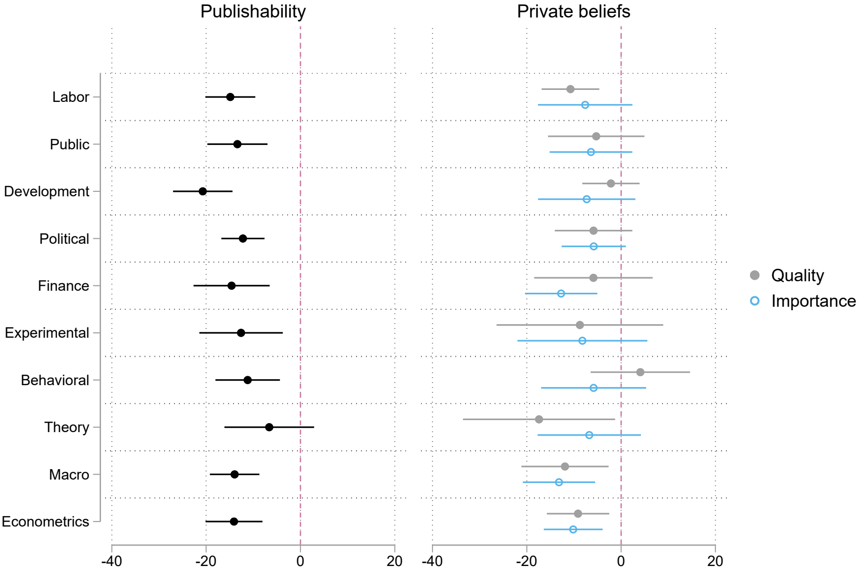
Negative perceptions of null results

	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: Individual 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
Respondents	480	230	230	250	250

Homogeneous null result penalty across groups



Little heterogeneity across academic fields



Robustness of the null result penalty

- ✓ Quantitatively similar effects using only the first vignette (between-subject)
- ✓ Null penalty robust across vignettes
- ✓ Post-stratification weights addressing selection concerns
- ✓ Robust to using only vignettes with high statistical power

Context matters: Expert forecasts and communication of uncertainty

	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

Mechanism experiment on **perceived statistical precision**

- **Question**: 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.
  - Elicit perceived **statistical precision** of the main result (instead of perceived quality and importance of the study)
  - Respondents are shown all five vignettes.

Perceived statistical precision

Precision

How would you rate the statistical precision of the main result?

- ○ Very precisely estimated
- ○ Precisely estimated
- ○ Somewhat precisely estimated
- ○ Imprecisely estimated
- ○ Very imprecisely estimated

Precision beliefs are affected by the statistical significance

	(1) Publishability (in percent)	(2) Precision (z-scored)
Panel A: Individual fixed effects		
Null result treatment	-19.755***	-1.267***
	(2.269)	(0.144)
Panel B: No individual FE		
Null result treatment	-18.134***	-1.086***
	(2.605)	(0.148)
Observations	475	475
Respondents	95	95

→ 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

Conclusion

1. Research studies with null results are perceived to be less publishable, of lower quality, of lower importance, and less precisely estimated
2. The null result penalty is larger when experts predict a non-null result
3. Communicating the statistical uncertainty of study results with p-values rather than standard errors further increases the null result penalty

Implications

- Potential value of pre-results review in which the decision on publication is taken before the empirical results are known (Kasy, 2021; Miguel, 2021)
- 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 counteract potential errors in statistical reasoning

Appendix Material

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, *89* (6), 2958–2991.

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.)

Christensen, Garret and Edward Miguel, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 2018, *56* (3), 920–80.

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.

Dufwenberg, Martin, Peter Martinsson et al., "Keeping researchers honest: The case for sealed-envelope-submissions," *IGIER (Innocenzo Gasparini Institute for Economic Research)*, 2014, *533.*

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.

References (cont.)

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan, "Promoting Transparency in Social Science Research," *Science*, 2014, *343* (6166), 30–31.

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.