THE ECONOMIC JOURNAL

The Economic Journal, **134** (*January*), 193–219 https://doi.org/10.1093/ej/uead060 © The Author(s) 2023. Published by Oxford University Press on behalf of Royal Economic Society. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Advance Access Publication Date: 7 August 2023

THE NULL RESULT PENALTY*

Felix Chopra, Ingar Haaland, Christopher Roth and Andreas Stegmann

We examine how the evaluation of research studies in economics depends on whether a study yielded a null result. Studies with null results are perceived to be less publishable, of lower quality, less important and less precisely estimated than studies with large and statistically significant results, even when holding constant all other study features, including the sample size and the precision of the estimates. The null result penalty is of similar magnitude among PhD students and journal editors. The penalty is larger when experts predict a large effect and when statistical uncertainty is communicated with *p*-values rather than standard errors. Our findings highlight the value of a pre-result review.

The scientific method is characterised 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. However, the publication system may favour research studies reporting large and statistically significant results over research papers documenting small results that are not statistically significant (Greenwald, 1975; Simonsohn *et al.*, 2014a; Camerer *et al.*, 2016). Selection of this type can lead to biased estimates and misleading confidence sets in published studies (Andrews and Kasy, 2019) and has led to a call for changes to the publication system (Nosek *et al.*, 2012; Camerer *et al.*, 2019; Kasy, 2021; Miguel, 2021).

In this paper, we investigate whether researchers penalise studies with null results, what mechanisms can explain the presence of a null result penalty and potential ways to mitigate the extent of a null result penalty. To address these questions, we conduct an experiment with about 500 researchers recruited from the leading top 200 economics departments in the world, of which about 20% have editorial experience at scientific journals.

* Corresponding author: Christopher Roth, WiSo Faculty, Albertus-Magnus-Platz, University of Cologne, ECONtribute, CEPR, briq, Max-Planck Institute for Collective Goods Bonn, 50923 Köln. Email: roth@wiso.uni-koeln.de

This paper was accepted on 3 August 2023. The Editor was Sule Alan.

The data and codes for this paper are available on the Journal repository. They were checked for their ability to reproduce the results presented in the paper. The authors were granted an exemption to publish parts of their data because access to these data is restricted. However, the authors provided the Journal with temporary access to the data, which enabled the Journal to run their codes. The codes for the parts subject to exemption are also available on the Journal repository. The restricted access data and these codes were also checked for their ability to reproduce the results presented in the paper. The replication package for this paper is available at the following address: https://doi.org/10.5281/zenodo.8168773.

We thank all participants of this study for generously sharing their time. We thank the editor (Sule Alan) and four anonymous referees for very helpful and highly constructive feedback. We also thank Peter Andre, Isaiah Andrews, Lukas Hensel, Johannes Hermle, Alex Imas, Max Kasy, Matt Lowe, Erzo Luttmer, Andrew Oswald, Nick Otis, David Schindler, Jesse Shapiro, Abhijeet Singh, Dmitry Taubinsky and seminar audiences at DIW Berlin, CEBI (University of Copenhagen) and the Norwegian School of Economics for excellent suggestions. We thank Shruti Agarwal, Pietro Ducco and Apoorv Kanongo for excellent research assistance. We thank Peter Andre and Armin Falk for sharing data. We received ethics approval from the ethics committee of the University of Cologne. The experiments were pre-registered in the AsPredicted registry (#95235 and #96599).

Roth acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2126/1-390838866. The project received financial support from the Research Council of Norway through its Centres of Excellence Scheme, FAIR project no 262675. The activities of the Center for Economic Behavior and Inequality (CEBI) are financed by the Danish National Research Foundation, Grant DNRF134.

Identifying whether there is a penalty for null results is challenging as studies that obtain or do not obtain statistically significant results might differ systematically in important dimensions. For instance, studies that obtain statistically non-significant results might have lower statistical power and less precise estimates, leading to large confidence intervals. To study whether there is a *penalty* for null results, we, therefore, rely on an experimental approach that holds all other study characteristics constant, including the statistical precision of the estimates. We present participants with four hypothetical vignettes that are based on actual research studies, but modified for the purposes of the experiment. For each of the vignettes, we randomise whether the point estimate of the main treatment effect was sizeable and statistically significant or close to zero and not statistically significant. To fix the statistical precision of estimates, we keep the standard error of the main finding constant across treatments.

To examine whether a potential penalty for null results depends on the communication of the statistical uncertainty of the main result, we cross-randomise at the respondent level whether the statistical precision of the main finding is communicated in terms of p-values or the standard error of the estimate. To study how the evaluation of research studies depends on expert priors, we cross-randomise whether the vignette includes expert forecasts of the treatment effect. For vignettes including expert forecasts, we further randomise whether the experts predict a large and statistically significant effect or a small and not statistically significant effect. Finally, to obfuscate the purpose of the experiment, we further cross-randomise a series of other salient study characteristics, including the seniority of the research team and their university affiliations.

Our main outcome of interest are beliefs about the publishability of the research studies. We elicit these beliefs by asking respondents how likely they think it is that the study in question would be published in a specific journal. We cross-randomise at the vignette level whether the journal in question is a general-interest journal or a field journal. To examine mechanisms, we further elicit personal beliefs about the quality and importance of the study as well as beliefs about other researchers' evaluation of the quality and importance of the study. We measure first-order beliefs to understand participants' private assessments of the research studies—irrespective of how other researchers and editors may evaluate such studies. The data on second-order beliefs furthermore allow us to examine whether there is a wedge between personal beliefs and the perceived beliefs of other researchers in the field.

We document that studies with a null result are perceived to be less publishable, of lower quality and of lower importance than studies with statistically significant results even when holding constant all other study features, including the statistical precision of estimates. Specifically, our respondents associate null result studies with a 14.1 percentage point (or 24.9%) lower chance of being published (95% confidence interval (CI) $[-16.2, -11.9]$; $p < 0.001$) as well as 37.3% of an SD lower quality (95% CI $[-49.6, -25]$; $p < 0.001$) and 32.5% of an SD lower importance (95% CI $[-43, -21.9]$; $p < 0.001$). Our respondents also think that other researchers would associate studies that yield a null result with 46% of an SD lower quality (95% CI $[-58.1, -33.8]$; $p < 0.001$) and 41.7% of an SD (95% CI $[-52.6, -30.7]$; $p < 0.001$) lower importance. The effect size on beliefs about others' assessments of null results in terms of quality and importance is thus slightly larger than the effect on their personal assessments, suggesting some degree of pluralistic ignorance—that is, the phenomenon of people incorrectly believing that their peers hold different beliefs than themselves—about how negatively other researchers evaluate null results.

The null result penalty is of similar magnitude for various subgroups of researchers, from PhD students to journal editors.[1] This suggests that the null result penalty is not an artefact of inexperience with the publication process itself. Rather, we find that even highly cited researchers and editors of scientific journals perceive studies with null results to be less publishable, of lower quality and less important. The fact that we find a null result penalty among journal editors is particularly noteworthy since editors observe the referee recommendations and editorial decisions for a substantial number of papers and should thus hold more accurate beliefs about the existence and magnitude of a null result penalty. Finally, the null result penalty robustly emerges in each of the vignettes presented to participants and is of similar magnitude for vignettes describing research studies with varying degrees of statistical power.

A long-standing concern in the academic community is that an excessive focus on *p*-values could amplify problems related to the replicability of scientific findings (Camerer *et al.*, 2016; Wasserstein and Lazar, 2016). To examine the potential role of how we communicate statistical uncertainty in research studies, we examine heterogeneity in treatment effects by whether respondents were given information about the *p*-value or the standard error of the main estimate presented in the vignettes. We find that the negative effect on the perceived probability of being published is somewhat larger when the main results are reported with *p*-values (95% CI [−10, −0.4]; $p = 0.032$). Moreover, reporting results with *p*-values instead of the standard error further leads our respondents to associate a null result study with 29.6% of an SD lower quality (95% CI [−56.5, −2.8]; $p = 0.03$) and also makes them think that other researchers will associate the study with 28.6% of an SD lower quality (95% CI [−56.4, −0.8]; $p = 0.044$).

The null result penalty can lead to biased estimates and misleading confidence sets in published studies (Andrews and Kasy, 2019). However, deciding on which findings should be published is a normative question. A null result penalty might be optimal depending on the social objective function underlying the publication process. For example, researchers might think that the publication process favours a criterion based on policy relevance over the objective of unbiased inference from published results. To the best of our knowledge, Frankel and Kasy (2022) are the only ones to propose an operationalisation of policy relevance grounded in economic theory. In their model, policy relevance is maximised by selectively publishing studies with surprising findings relative to the prior in the literature.[2]

To test whether this can explain our results, we examine heterogeneity in treatment effects by whether the null result is in line with expert forecasts. First, we find that the null result penalty is unchanged when respondents additionally receive an expert forecast predicting a null result rather than not receiving an expert forecast. Second, we find that the negative effect of null results on publishability is even aggravated when a null result is at odds with expert forecasts: respondents evaluate a study with a null result as having a further 6.4 percentage point lower chance of being published (95% CI [−11.6, −1.2]; $p = 0.016$). These patterns are inconsistent with the conjecture that respondents believe that the publication process favours research findings with surprising results.

Finally, we conduct an additional experiment with early career researchers in which we test whether individuals perceive studies with null results as less precisely estimated even when these studies have the same objective statistical precision as studies with larger and statistically significant effects. We employ the same vignettes as in the main experiment, but replace the

[1] The results remain virtually identical when re-weighting our sample to be representative of the underlying population of researchers.

[2] Abadie (2020) showed that failure to reject a null hypothesis is very informative in many settings.

questions about quality and importance with a question about the perceived precision of the main result. PhD students and early career researchers associate studies with null results with a 19.8 percentage point (or 32.5%) lower chance of being published (95% CI [−24.3, −15.2]; $p < 0.001$). Furthermore, they associate studies with null results with a 126.7% of an SD lower precision (95% CI [−155.2, −98.2]; $p < 0.001$). Given that we fixed respondents' beliefs about the standard error of the treatment effect, this finding is inconsistent with Bayesian explanations of learning about unobservables and instead suggests that researchers may use simple heuristics to assess the statistical precision of findings. Indeed, some researchers may erroneously equate statistical significance with statistical precision.

Our study relates to a growing literature on the publication process (Card and DellaVigna, 2013; Kasy, 2019; Card *et al.*, 2020; Card and DellaVigna, 2020; Ersoy and Pate, 2021; Frankel and Kasy, 2022) and in particular publication bias (Ioannidis, 2005; Dwan *et al.*, 2008; Gerber and Malhotra, 2008; Franco *et al.*, 2014; Brodeur *et al.*, 2016; 2020; Blanco-Perez and Brodeur, 2020).[3] This literature has examined the extent to which null results are less likely to be published (Simonsohn *et al.*, 2014a). Brodeur *et al.* (2016) studied the distribution of *p*-values in published papers. Their accounting exercise showcases a missing mass of *p*-values between 0.25 and 0.10 and an excess mass just below the 0.05 significance threshold, consistent with either researchers selectively reporting research findings or studies with marginally significant results being favoured in the peer review system. Brodeur *et al.* (2021) showed that initial submissions display significant bunching in *p*-values, suggesting that the abnormal distribution among published statistics is at least in part a result of researchers being selective in terms of which findings to write up and submit for publication. Yet, Brodeur *et al.* (2021) also showed that reviewer recommendations are significantly affected by statistical thresholds, consistent with marginally significant results being favoured in the peer review system.[4]

We contribute to this literature by studying mechanisms underlying publication bias in tightly controlled, large-scale experiments with hypothetical vignettes, circumventing the potential confound that studies that obtain large and statistically significant results might be systematically different from studies with small and not statistically significant results. This approach allows us to flexibly control for a variety of study features, specifically to hold constant issues related to the selection of papers up until the submission stage and to identify a null result penalty conditional on papers being submitted for publication. We also examine mechanisms underlying the null result penalty with rich data on how null results shape perceptions of the quality, importance and precision of the studies. Our finding that studies with null results are perceived to be more noisily estimated suggests some role for errors in statistical reasoning in explaining the null result penalty.

Our work also relates to a literature on the adoption of editorial policies aiming to promote research transparency and to reduce publication bias (Dufwenberg and Martinsson, 2014; Miguel *et al.*, 2014; Nosek *et al.*, 2015; Christensen and Miguel, 2018), such as the effect of editorial statements emphasising the potential merit of scientific studies irrespective of the statistical significance of their main empirical estimates (Blanco-Perez and Brodeur, 2020). Our experimental

[3] A related literature has examined the replicability of research findings (Klein *et al.*, 2014; Simonsohn *et al.*, 2014a,b; Open Science Collaboration, 2015; Camerer *et al.*, 2016; 2018; Klein *et al.*, 2018) and has discussed research transparency efforts (Christensen *et al.*, 2019).

[4] The influence of null results on the publishability of studies has also been examined in medicine and other social sciences, though not with a focus on uncovering the mechanisms behind a null result penalty. Emerson *et al.* (2010) and Elson *et al.* (2020) examined publication bias using audit studies with medical scientists and psychologists, respectively. Berinsky *et al.* (2021) employed conjoint experiments with a sample of political scientists, focusing on publication biases in the context of replication studies.

approach allows us to study additional, potential measures to mitigate the null result penalty, such as providing expert forecasts and expressing statistical uncertainty in terms of standard errors rather than *p*-values.

Finally, our paper relates to a descriptive literature on the beliefs and reasoning of academic experts (Casey *et al.*, 2012; Dreber *et al.*, 2015; DellaVigna and Pope, 2018; Andre and Falk, 2021; Andre *et al.*, 2022a,b) and policymakers (Vivalt and Coville, 2020; 2023; Hjort *et al.*, 2021). We assess how academic economists' perceptions of the publishability, quality, importance and precision of research studies hinge on the results of the study.

Our paper proceeds as follows. Section 1 describes the sample and the experimental design. In Section 2, we present the main results, and heterogeneous treatment effects. We present evidence on mechanisms in Section 3. Section 4 examines the robustness of our results. Finally, Section 5 discusses the implications of our findings for the publication system and the production of research.

1. Experimental Design and Data

1.1. *Sample*

In April and May 2022, we invited 14,087 academic researchers in the field of economics affiliated with one of the top 200 institutions according to RePEc (as of March 2022) to participate in a 10-min online survey. We chose to send out only one invitation email without any subsequent reminder emails. While reminder emails could have increased the overall response rate, we decided not to send more than one invitation email in light of the increasing popularity of expert surveys and the overall burden imposed on researchers by receiving invitation emails. In total, 480 researchers follow our invitation and complete the online survey, implying an overall response rate of 3.4%.

Table 1 provides relevant summary statistics for this sample of academic experts. Reflecting imbalances in the wider profession, our sample is not gender balanced with a male share of 78.0%. Of our respondents, 24.4% are PhD students. Respondents with a PhD in our sample graduated 14.8 years ago on average (as of 2022). In line with most top 200 economics departments being located in Europe and North America, the large majority of our respondents are based at institutions in Europe (54.4%) and North America (40.6%). Many of our respondents have substantial experience as both producers and evaluators of academic research. Our respondents have on average 1.3 research articles published in one of the 'top five' economics journals. Their work is also highly cited. The average (median) h-index among our respondents with a Google Scholar profile is 17.2 (11.5). Furthermore, their average (median) total citations are 4,348.3 (845.5). Our respondents have on average refereed for 1.2 of the top five economics journals. Furthermore, sizeable fractions of our respondents are currently an editor (7.2%) or an associate editor (12.7%) of a scientific journal. Our respondents also have experience in different subfields of economics, including labour economics (21.1%), econometrics (14.1%), development economics (17.9%), political economy (16.7%), finance (10.5%), behavioural economics (9.1%), macroeconomics (14.1%) and theory (6.7%). These summary statistics underscore that our sample is diverse and contains a large fraction of highly experienced researchers with substantial academic impact in the field of economics. The large diversity in terms of both research fields, academic output and experience with research evaluation mitigates concerns about external validity.

Table 1. *Descriptive Statistics.*

	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 five publications	1.27		462	0.34	0
Number of top fives refereed for	1.17		397		
Repeated top five 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		
Professional memberships:					
NBER affiliate	0.08		454		
CEPR affiliate	0.17		451		
Academic fields:					
Labour	0.21		418		
Public	0.13		418		
Development	0.18		418		
Political	0.17		418		
Finance	0.11		418		
Experimental	0.06		418		
Behavioural	0.09		418		
Theory	0.07		418		
Macro	0.14		418		
Econometrics	0.14		418		

Note: This table displays characteristics of the participants in the main experiment. These data are not matched with responses, but instead are externally collected from publicly available CVs (i.e., not self-reported). Online Appendix B.1 contains a description of each variable. Data on the underlying sampling population were shared by Peter Andre and Armin Falk (see Andre and Falk, 2021). Online Appendix B.2 describes how our measures differ from those obtained from Andre and Falk (2021), in particular, 'Years since PhD' and 'h-index'.

For a subset of characteristics, we are also able to present averages for the underlying sampling population of all researchers affiliated with one of the top 200 institutions according to RePEc (as of March 2022) thanks to data by Andre and Falk (2021). As shown in Table 1, the comparison to the population averages indicates that respondents in our sample are relatively more senior and experienced with the publication process (as evidenced by holding an editorial position or acting as a repeated top five referee) than the average researcher in the overall sampling frame. Moreover, the share of respondents who are based in Europe exceeds the population average (mostly at the expense of researchers based in North America). While our sample is thus not perfectly representative of the sample frame, the availability of population averages for the

underlying sampling frame allows us to examine the robustness of our analysis to re-weighting our sample to match key moments of the sampling population.

1.1.1. *Pre-specification*

The data collections were pre-registered in the AsPredicted registry (#95235 and #96599). We pre-specified the sampling procedure, the main outcomes of interest, the main right-hand-side variable of interest, as well as the baseline specifications. The pre-analysis plans can be found in Section F of the Online Appendix. All of our analyses follow the pre-analysis plans unless otherwise noted.

1.2. *Design*

1.2.1. *Baseline design*

We created five hypothetical vignettes describing different research studies. Each vignette is loosely based on an actual research paper in economics. We inform respondents about the hypothetical nature of the vignettes after obtaining their consent to avoid deception. The vignettes draw on a variety of different fields (labour, education, economic history, behavioural economics, development economics and household finance) and methods (randomised controlled trials (RCTs), regression discontinuity design and online experiments). The vignette approach gives us a lot of flexibility in varying study characteristics while fixing all other observable characteristics. A potential concern about vignette designs, however, is that respondents might make inferences about unobservable characteristics that we do not strictly control (Haaland *et al.*, 2023). Table 2 provides a summary of the characteristics of the studies used for the different vignettes.

All of the vignettes follow the same structure. We first describe some background information about the study and introduce the research question. We next outline the key features of the research design, including details about the main treatment variation and the primary outcome of interest. For studies without a reduced-form effect of direct interest, a relevant first stage is necessary to judge the quantitative importance of the main finding. In such cases, we provide information about the size of the first stage before presenting the main result to respondents. Furthermore, in the context of natural research designs such as regression discontinuity design, we also provide information about the validity of the identifying assumptions before presenting the main result. The baseline instructions for one of the vignettes are as follows.

> **Background and study design:** Three professors from Brown University conducted an RCT in Texas in the years 2015–9. The purpose of the RCT was to examine the effects of a randomly assigned $8,000 merit aid program for low-income students on the likelihood of completing a bachelor's degree.
> The researchers worked with a sample of 1,188 high school graduates from low-income, minority and first-generation college households. Of those students 594 were randomly assigned to receive $8,000 in merit aid for one year, while the remainder of the students did not receive any additional aid.

1.2.2. *Null result treatment*

We next provide respondents with information on the main result of the study (randomised at the respondent-vignette level): half of the respondents are informed that the study had a main effect close to zero (*null result* treatment), while the other half of respondents are informed that the study had a sizeable main effect (*significant result* treatment). Importantly, while we vary the point estimate between treatments, we keep the sample size and the standard error of the estimates constant across treatments. Yet, we cannot differentiate between a penalty for

Table 2. *Overview of the Vignettes.*

	Marginal effects of merit aid for low-income students (1)	Long-term effects of equal land sharing (2)	Female empowerment program (3)	Financial literacy program (4)	Salience of poverty and patience (5)
Panel A: general information					
Fields	Labour, education	Economic history	Behavioural, labour, development	Development, household finance	Behavioural economics
Country	USA	Germany	Sierra Leone	India	USA
Type of study	RCT	Regression discontinuity	RCT	RCT	Online experiment
Outcome	Completion of a four-year bachelor's degree	County income	Take-up of job offer	Any savings	Choose money now over money later
Nature of outcome	Dummy	Continuous	Dummy	Dummy	Dummy
Panel B: numerical features					
Observations	1,188	400	360	780	800
Control group mean	17.0	–	37.0	42.0	45.0
Standard error	2.9	2.4	5.0	3.8	3.5
Main effect: high	6.6	6.2	13.1	8.4	7.8
Main effect: low	1.1	0.5	1.7	1.6	1.6
p-value: high main effect	0.02	0.01	0.01	0.03	0.03
p-value: small main effect	0.71	0.83	0.73	0.68	0.64
MDE (% of an SD, 80% power)	20%	30%	30%	20%	20%
Panel C: expert forecasts					
Number of experts	24	23	34	26	22
Prior: high mean	5.7	7.4	12.0	9.5	8.8
Prior: low mean	0.2	1.7	0.6	2.7	2.7
SD	3.2	4.7	7.6	5.8	6.9
Panel D: journals					
Field journal	JHR	JEG	JDE	JPubEc	EE
General-interest journal	EJ	ReStud	Science	ReStat	PNAS
Panel E: university					
Higher-ranked university	Brown University	Northwestern University	UC Berkeley	Columbia University	Harvard University
Lower-ranked university	University of Illinois	University of Arkansas	Boston College	University of Pittsburgh	Ohio State University

Note: This table provides an overview of the vignettes. The abbreviations of the field journals stand for the following journals: JHR, Journal of Human Resources; JEG, Journal of Economic Growth; JDE, Journal of Development Economics; JPubEc, Journal of Public Economics; EE, Experimental Economics. The abbreviations of the general-interest journals stand for the following journals: ReStud, The Review of Economic Studies; EJ, The Economic Journal; ReStat, Review of Economics and Statistics; PNAS, Proceedings of the National Academy of Sciences.

statistically non-significant results and a penalty for results with coefficients close to zero. We chose to keep the precision of the estimates constant across conditions because, all else equal, a less precise result is less informative about the effects being studied and thus less likely to be perceived as publishable.[5] We construct the standard errors such that the main effect is not

[5] The only way to vary significance while holding the precision of the estimates fixed is to vary the point estimate of the treatment effects. We thus have one condition with a small and not statistically significant effect and one condition with a large and statistically significant effect. An alternative approach to vary the significance of the findings would be to keep the point estimates constant, but to vary the precision of the estimates. However, this approach is confounded by varying quality on a dimension not related to the results of the study.

statistically significant in the *null result* treatment and statistically significant (at conventional significance thresholds) in the *significant result* treatment.[6]

For instance, in the vignette on merit aid for low-income students discussed above, respondents in the *null result* treatment receive the following instructions.

> **Main result of the study:** The treatment increased the completion rate of a 4-year bachelor's degree by 1.1 percentage points (standard error 2.9) compared to a control mean of 17%.

In contrast, respondents in the *significant result* treatment of this vignette receive the following instructions.

> **Main result of the study:** The treatment increased the completion rate of a 4-year bachelor's degree by 6.6 percentage points (standard error 2.9) compared to a control mean of 17%.

We randomly assign respondents to assess four of the five vignettes we designed, generating data at the vignette-respondent level and within-respondent variation in the statistical significance of the displayed treatment effect estimate. We also randomise the order of vignettes.

1.2.3. *Expert predictions*

We cross-randomise at the respondent-vignette level whether we provide respondents with expert predictions of the main treatment effect estimate, allowing us to examine whether the evaluation of studies with null results depends on whether the result is surprising to experts or in line with expert predictions.[7] Specifically, one-third of the vignettes do not include any expert forecast. For the remaining two-thirds of the vignettes, half include a high expert forecast and half include a low expert forecast. We construct the low and high expert forecasts such that they are close, but not identical, to the magnitude of the coefficient estimate in the *null result* and *significant result* treatments, respectively. We also provide the SD of the expert forecast to communicate the degree of disagreement among experts. To ensure that there is scope for substantial updating, we set the SD of the expert forecasts to be two to three times the standard error of the point estimate in each vignette.[8]

In the context of the vignette on merit aid for low-income students, respondents assigned to the high expert prediction receive the following instructions.

> **Expert prediction:** 24 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 5.7 percentage points. The standard deviation of the expert forecasts was 3.2.

In contrast, respondents assigned to the low expert prediction receive the following instructions.

> **Expert prediction:** 24 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 0.2 percentage points. The standard deviation of the expert forecasts was 3.2.

1.2.4. *Communication of statistical uncertainty*

A common view in the academic community is that an excessive focus on *p*-values might amplify problems related to the replicability of scientific findings (Camerer *et al.*, 2016; Wasserstein and

[6] Section C of the Online Appendix contains a full description of the data-generating process that we used to generate the numerical values for the vignette features that we vary experimentally.

[7] Expert priors have been argued to be a potential remedy against a null result penalty and are by now increasingly used in social science research (DellaVigna *et al.*, 2019).

[8] This ensures that the findings from the study in each vignette are valuable in that they either lead to movement of the posterior mean or a substantial reduction in the uncertainty of the posterior belief about the true effect size.

Lazar, 2016). To examine how the communication of statistical uncertainty affects the evaluation of studies with null results, we also cross-randomise whether the statistical uncertainty of the main finding is communicated in terms of the *p*-value or the standard error of the main treatment effect. To minimise the scope for experimenter demand effects (de Quidt *et al.*, 2018), we cross-randomised this feature between respondents.

In the context of the vignette on merit aid for low-income students, respondents assigned to the *null result* treatment and cross-randomised to the *p*-value treatment receive the following instructions.

> **Main result of the study:** The treatment increased the completion rate of a 4-year bachelor's degree by 1.1 percentage points (p-value = 0.73) compared to a control mean of 17%.

In contrast, respondents in the *null result* treatment cross-randomised to the standard error treatment receive the following instructions.

> **Main result of the study:** The treatment increased the completion rate of a 4-year bachelor's degree by 1.1 percentage points (standard error 2.9) compared to a control mean of 17%.

1.2.5. *Obfuscation treatments*

We further cross-randomise two additional features for each vignette. First, we vary the seniority of the researchers conducting the study. Respondents assessing a given vignette are either informed that the study was carried out by a group of professors or a team of PhD students. Second, we vary the rank of the institution to which the researchers conducting the study are affiliated (shown in panel E of Table 2). The obfuscation treatments are featured at the beginning of each vignette to increase their salience.

The main purpose of these cross-randomised conditions is to obfuscate the study purpose to reduce concerns about experimenter demand and social desirability bias (Haaland *et al.*, 2023). This additional within-respondent variation substantially increases the difficulty of correctly guessing the true purpose of our research study as respondents only observe variation in vignette features for a relatively small number of vignettes. For instance, by making the university affiliation and the seniority of the research team salient, respondents could have guessed that we wanted to study discrimination against younger researchers or researchers from lower-ranked institutions in the publication process.[9] Furthermore, on top of the benefits from obfuscation, both conditions provide us with an opportunity to investigate the extent to which there are heterogeneous treatment effects of the null result treatment and to test whether respondents paid attention to the instructions.

1.2.6. *Summary of the factorial design*

Figure 1 presents an overview of the factorial design. Our full factorial design consists of five different factors: the variation in the style in which statistical uncertainty is communicated (two levels), the variation in the magnitude of the estimated treatment effect (two levels), the variation in the availability and magnitude of the expert predictions (three levels), the variation in the study authors' seniority (two levels) and the rank of the institution with which the study authors are affiliated (two levels); hence we have $2 \times 2 \times 3 \times 2 \times 2 = 48$ factorial combinations. The 480 respondents in our main study complete four out of five vignettes, providing us with 1,920

[9] To keep the survey below 10 min, we did not include any open-ended questions about the study purpose at the end of the survey.

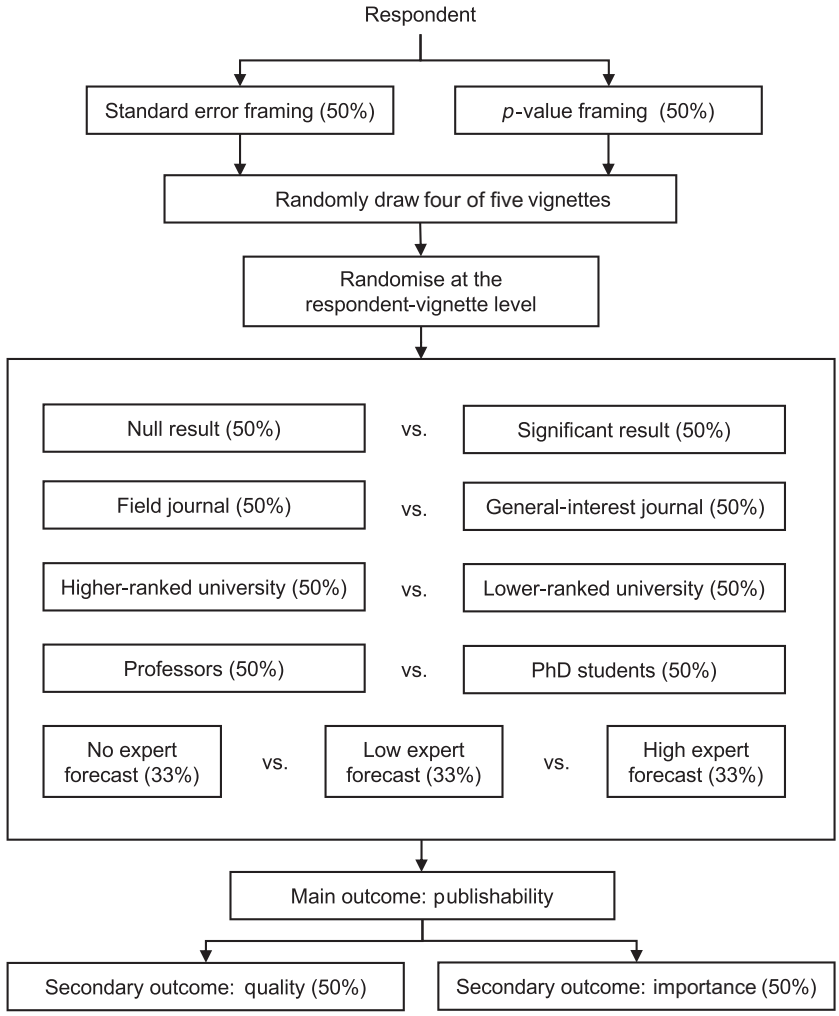

Fig. 1. *Overview of the Factorial Design.*
Note: This figure presents an overview of the factorial design, including the between-subject randomisation of the scientific communication treatment (standard errors versus *p*-value framing) and the between-subject randomisation of secondary outcomes (first- and second-order perceptions of quality versus importance). The figure also presents the randomisation of vignette features at the respondent-vignette level. The respondent-vignette level randomisation includes five factors: statistical significance (two levels), type of journal (two levels), rank of the university that the researcher team is affiliated with (two levels), seniority of the researcher team (two levels), and the presence and magnitude of expert forecasts (three levels). In the mechanism experiment (introduced in Section 3.3), respondents were shown all five hypothetical vignettes, but the cross-randomisation of features is otherwise identical.

observations in total. We, therefore, obtain 40 observations per factorial combination (median 40, min. 27, max. 54).

Using a clustered bootstrap procedure that resamples respondents with replacement, we estimate a minimum detectable effect size of 15.6% of an SD (corresponding to a 4 percentage

point difference in perceived publication chances) at 80% power in our main specification and a significance threshold of 5%.

1.3. *Main Outcomes*

After the presentation of each vignette, we ask our respondents three questions. Our main outcome of interest are researchers' perceptions of the likelihood that the study would eventually be published in a given journal. For each study, we cross-randomise whether the journal is a general-interest journal or a relevant field journal (shown in panel D of Table 2). For example, for the vignette on merit aid for low-income students, we cross-randomise whether respondents estimate the likelihood that the paper will eventually be published in the *Economic Journal* or the *Journal of Human Resources*.[10] The exact wording of this question is as follows: 'If this study was submitted to the Economic Journal, what do you think is the likelihood that the study would eventually be published there?' To answer this question, respondents move a slider between 0 and 100.

We then measure respondents' personal beliefs (first-order beliefs) and beliefs about the beliefs of others (second-order beliefs) about the quality of the research study (quality condition) or the importance of the research study (importance condition). We measure first-order beliefs to understand participants' private assessments of the research studies—irrespective of how other researchers and editors may evaluate such studies. The data on second-order beliefs furthermore allow us to shed light on a potential wedge between personal beliefs and the perceived beliefs of other researchers in the field.

To reduce concerns about survey fatigue, we cross-randomise at the respondent level whether respondents are asked about quality or importance. For respondents in the quality condition, we elicit respondents' perceptions of the quality of the study using a scale from 0 to 100, where 0 is 'lowest possible quality' and 100 is 'highest possible quality'. We measure second-order beliefs by asking each respondent to 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. We then ask respondents what quality rating they would expect these researchers to give to the study on average. For researchers in the importance condition, we elicit respondents' perceptions of the importance of the study using a scale from 0 to 100, where 0 is 'lowest possible importance' and 100 is 'highest possible importance'. We then measure beliefs about other researchers' assessments of the importance of the research study using a similar wording as in the quality condition.

2. Main Results

2.1. *Econometric Specification*

To estimate the effect of the *null result* treatment on researchers' evaluations of the research studies presented in our vignettes, we estimate the following pre-registered specification using OLS (Ordinary Least Squares):

$$y_{iv} = \alpha + \beta \,\text{null result}_{iv} + X'_{iv}\gamma + \delta_v + \tau_i + \varepsilon_{iv}. \quad (1)$$

[10] To avoid anchoring responses towards the acceptance rate of each journal, we did not include information about the journal acceptance rate in the vignettes.

Table 3. *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: 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 fixed effects					
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

Note: The table shows regression estimates of our treatment effects on our key outcomes of interest from (1). The data set is at the vignette-respondent level and contains four observations for each respondent. 'Null result treatment' is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. The regressions in panel A (panel B) include (do not include) individual-level fixed effects. All regressions in both panels include treatment indicators for the cross-randomised conditions in addition to vignette fixed effects. ***denotes statistical significance at the 0.01 threshold.

Here 'null result$_{iv}$' is a binary indicator taking value one if respondent i learns that the main treatment effect estimate in vignette v is small in magnitude and not statistically significant, and zero otherwise; X_{iv} is a vector of six binary indicators for all other cross-randomised treatment conditions; δ_v is a vignette fixed effect and τ_i is a respondent fixed effect. For inference, we use robust standard errors clustered at the respondent level.

When exploring heterogeneous treatment effects based on the other cross-randomised features, we follow the pre-analysis plan and run a set of separate regressions including interaction terms between the *null result* indicator and indicators z_{iv} for other cross-randomised features once at a time. Specifically, we estimate the following specification to examine heterogeneous treatment effects:

$$y_{iv} = \alpha + \beta \,\text{null result}_{iv} + \beta_z \, z_{iv} \,\text{null result}_{iv} + X'_{iv}\gamma + \delta_v + \tau_i + \varepsilon_{iv}. \tag{2}$$

Here β_z captures the effect of the interaction between the null result indicator and another cross-randomised feature, z_{iv}. In line with the pre-registration, we exclude respondent fixed effects (τ_i) when examining heterogeneity by the p-value framing, which is varied only between respondents.

2.2. *Main Treatment Effects*

Table 3 shows the effects of the *null result* treatment on our main outcomes of interest. Panel A shows estimates controlling for respondent fixed effects, while panel B shows estimates excluding respondent fixed effects. As shown in column (1), respondents assigned to the *null result* treatment indicate that the studies have a 14.1 percentage point lower probability of being published (95% CI [−16.2, −11.9]; $p < 0.001$). This effect size corresponds to a 24.9% reduction in perceived

publication chances. In other words, there is a substantial perceived penalty for studies with small and non-significant results in the publication system.[11]

Columns (2)–(5) examine some of the mechanisms behind this null result penalty. As shown in column (2), respondents in the *null result* treatment associate the studies with 37.3% of an SD lower quality (95% CI [−49.6, −25]; $p < 0.001$), consistent with a mechanism in which researchers broadly associate studies that yield null results with lower quality. Furthermore, as shown in column (3), they also think that other researchers would associate the studies with 46% of an SD lower quality (95% CI [−58.1, −33.8]; $p < 0.001$). The effect size on beliefs about others' assessments is thus slightly larger than for their personal beliefs ($p = 0.10$), suggesting some form of pluralistic ignorance about the perceived quality of studies with null results.

Columns (4) and (5) also show sizeable treatment effects on the perceived importance of the studies. Respondents in the *null result* treatment associate the studies with 32.5% of an SD lower importance (95% CI [−43, −21.9]; $p < 0.001$) and think other researchers would associate the studies with 41.7% of an SD lower importance (95% CI [−52.6, −30.7]; $p < 0.001$). We thus also see suggestive evidence consistent with some pluralistic ignorance for perceptions about importance ($p = 0.057$), though it is important to emphasise that the updating about quality and importance is large and negative both for personal beliefs and the beliefs of others.

2.3. *Heterogeneity*

2.3.1. *Heterogeneity by respondent characteristics*

As discussed in Section 1.1, there is substantial heterogeneity in experience with the production and evaluation of research studies among our respondents. While not part of the pre-analysis plan, we next analyse treatment heterogeneity by respondent characteristics. Figure 2 shows that the treatment effects are similar across different subgroups. For instance, the null result penalty is of similar magnitude among male and female respondents. It is also of similar magnitude among experienced researchers with top five publications and many citations and less experienced researchers with fewer citations and no top five publications as well as among professors and PhD students. Furthermore, as shown in Figure 3, we also see that the null result penalty is homogeneous across different fields of specialisation, including among respondents who specialise in econometrics. The lack of heterogeneous effects across different subgroups underscores that the null result penalty is applied broadly across the profession and is not driven by, for instance, a set of inexperienced researchers with less influence in the publication process or by researchers from a particular subfield of economics. These largely homogeneous treatment effects by respondent characteristics also mitigate concerns about external validity.[12]

As shown in Figure 2, we also see a very similar belief in a null result penalty—both in terms of perceived publishability and personal beliefs about quality and importance of the research studies—among researchers with experience as editors of scientific journals and those who have never held an editorial position (Online Appendix Figure A.1 shows similar patterns for beliefs about others). This result is particularly noteworthy for two reasons. First, journal editors observe a substantial number of editorial decisions and referee recommendations. Journal editors

[11] Panel B of Table 3 shows that we obtain virtually identical results when excluding respondent fixed effects. In contrast to the main specification, this specification also uses respondents who are always shown studies with null results and those always shown studies with significant results.

[12] We also see homogeneous effects across subgroups in the mechanism experiment introduced in Section 3.3 (results available upon request).

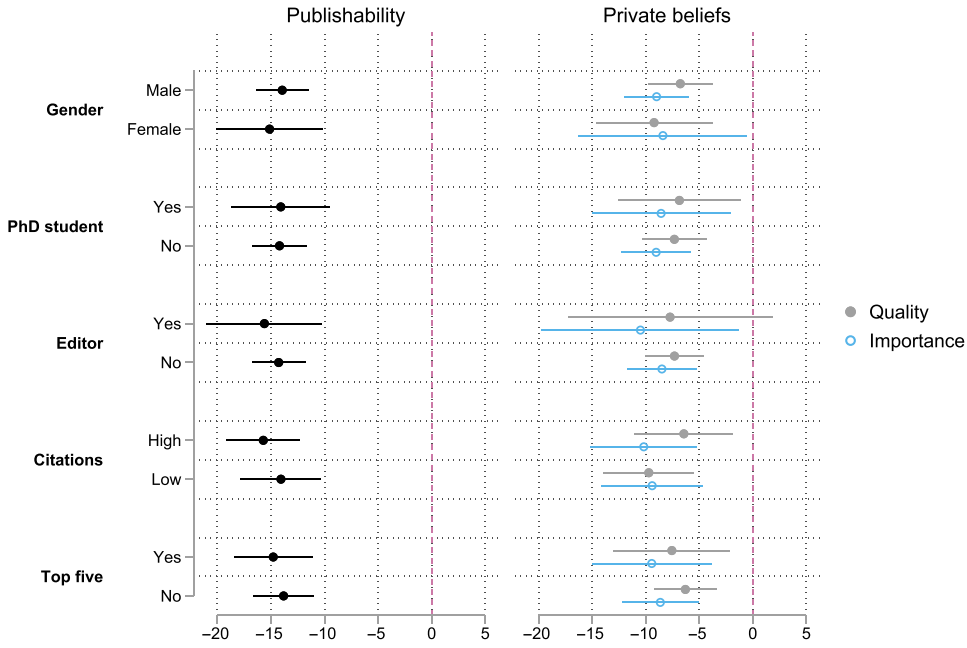

Fig. 2. *Heterogeneity in Treatment Effects by Respondent Characteristic.*
Note: This figure shows regression estimates in which beliefs about the percent chance of the study being published (measured on a scale from 0 to 100) as well as personal beliefs about importance and quality of the study (both measured on a scale from 0 to 100) are regressed on the 'null result treatment' indicator, separately for each subgroup indicated in the figure. Citations are measured using Google Scholar data as of May 2022 and 'low' and 'high', respectively, refer to below or above median citations in our sample. 'Editor' refers to whether the respondent ever has been an editor of a scientific journal. 'Top five' refers to whether the respondent has published a paper in any of the 'top five' economics journals. All regressions include controls for the other cross-randomised features at the vignette level as well as respondent fixed effects. Standard errors are clustered at the respondent level. We indicate 95% confidence intervals in the figure.

should thus hold relatively accurate beliefs about the existence and the magnitude of a null result penalty. Second, one of the main tasks of journal editors is to screen which papers to send out for peer review. These decisions are many times based on the abstract and an initial reading of the introduction. As such, our vignettes—which give respondents a summary of the key design elements as well as the main results—arguably provide information at a level similar to what is used in many editorial screening decisions, resulting in an especially high external validity of our results for editorial decision-making.[13]

2.3.2. *Heterogeneity by vignette characteristics*

While we included a set of cross-randomised conditions primarily to obfuscate the purpose of our study, the analysis of heterogeneous treatment effects of these cross-randomised 'obfuscation treatments' also provides valuable insights on the determinants of the null result penalty. As shown

[13] A similar screening mechanism likely operates when evaluating papers for conference submissions or when choosing which job market candidates to interview.

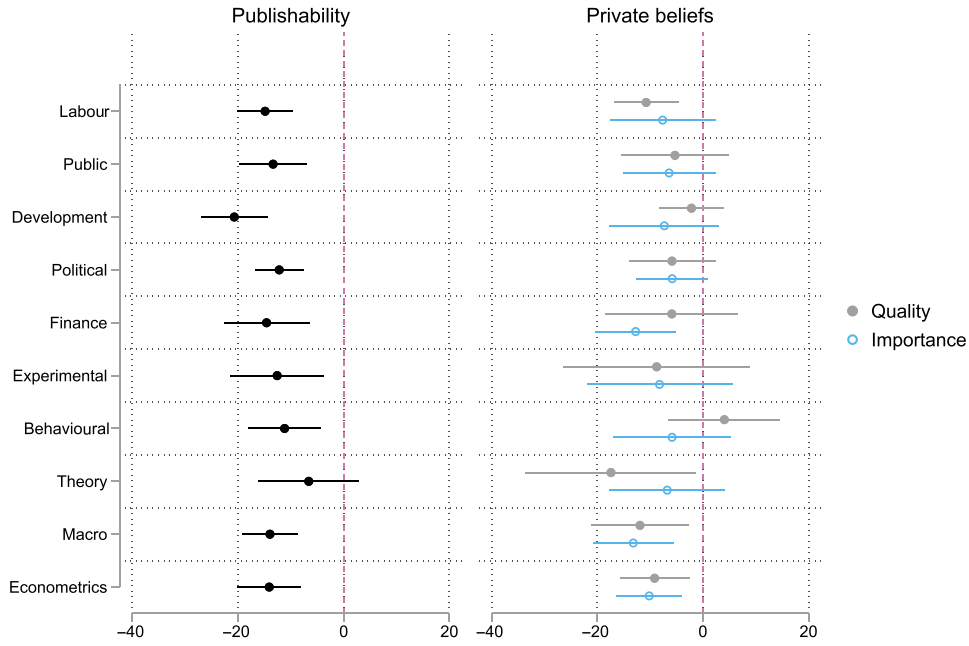

Fig. 3. *Heterogeneity by Field of Specialisation.*
Note: This figure shows regression estimates in which beliefs about the percent chance of the study being published (measured on a scale from 0 to 100) as well as personal beliefs about importance and quality of the study (both measured on a scale from 0 to 100) are regressed on the 'null result treatment' indicator, separately for each subgroup indicated in the figure. The regressions include controls for all cross-randomised features at the vignette level as well as respondent fixed effects. Standard errors are clustered at the respondent level. We indicate 95% confidence intervals in the figure.

in panel B of Table 4, we find a similar null result penalty for field journals and for general-interest journals. We also do not observe heterogeneous treatment effects of the null result treatment when the research team is composed of PhD students or when the researchers conducting the study are affiliated with a lower-ranked university (as shown in panels C and D of Table 4). This suggests that respondents believe that the null result penalty is rather universal and not specific to particular types of research teams.

Although our respondents do not think that the null result penalty interacts with any of the obfuscation treatments, we still observe large and precisely estimated main effects of the obfuscation treatments on the perceived publishability of the studies. Respondents indicate that they perceive studies to have a 12.2 percentage point higher probability of being published in field journals compared to general-interest journals (95% CI [9.5, 15]; $p < 0.001$). Moreover, they expect studies authored by PhD students rather than professors to have a 4.5 percentage point lower probability of being published (95% CI [−7.3, −1.8]; $p < 0.001$). Similarly, they expect studies conducted by researchers affiliated with lower-ranked universities to have 4.0 percentage point lower publication chances (95% CI [−6.7, −1.3]; $p = 0.004$). These results suggest that respondents read the vignettes attentively.

Table 4. *Heterogeneity by Vignette Characteristics.*

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1)	(2)	(3)	(4)	(5)
	Beliefs in percent	First-order beliefs	Second-order beliefs	First-order beliefs	Second-order beliefs
Panel A: expert forecast					
Null result treatment	−11.239***	−0.281**	−0.506***	−0.351***	−0.432***
	(1.913)	(0.113)	(0.112)	(0.094)	(0.085)
Null result × low expert forecast	−2.002	−0.168	0.128	0.032	0.065
	(2.478)	(0.161)	(0.161)	(0.120)	(0.116)
Null result × high expert forecast	−6.383**	−0.104	0.009	0.048	−0.020
	(2.646)	(0.166)	(0.154)	(0.124)	(0.126)
Low expert forecast	−0.890	0.190*	−0.015	−0.076	−0.042
	(1.671)	(0.108)	(0.108)	(0.092)	(0.081)
High expert forecast	1.959	0.115	0.121	−0.049	−0.018
	(1.800)	(0.112)	(0.099)	(0.085)	(0.088)
Panel B: field journal					
Null result treatment	−14.571***	−0.366***	−0.446***	−0.343***	−0.418***
	(1.465)	(0.093)	(0.086)	(0.072)	(0.075)
Null result × field journal	1.025	−0.014	−0.027	0.036	0.003
	(1.965)	(0.129)	(0.122)	(0.101)	(0.103)
Field journal	12.218***	0.141	0.108	0.108	0.101
	(1.397)	(0.095)	(0.089)	(0.072)	(0.069)
Panel C: PhD student					
Null result treatment	−14.945***	−0.291***	−0.358***	−0.300***	−0.362***
	(1.491)	(0.085)	(0.082)	(0.081)	(0.081)
Null result × PhD student	1.745	−0.166	−0.206*	−0.047	−0.104
	(2.049)	(0.117)	(0.107)	(0.102)	(0.097)
PhD student	−4.543***	−0.025	−0.042	0.066	0.019
	(1.403)	(0.091)	(0.081)	(0.071)	(0.069)
Panel D: low-ranked university					
Null result treatment	−14.320***	−0.381***	−0.474***	−0.317***	−0.408***
	(1.480)	(0.094)	(0.093)	(0.073)	(0.076)
Null result × low-ranked university	0.518	0.017	0.030	−0.014	−0.017
	(1.985)	(0.121)	(0.124)	(0.108)	(0.105)
Low-ranked university	−3.998***	−0.093	−0.230***	0.007	−0.046
	(1.371)	(0.082)	(0.077)	(0.077)	(0.072)
Panel E: p-value framing					
Null result treatment	−11.960***	−0.243**	−0.302***	−0.366***	−0.405***
	(1.736)	(0.095)	(0.101)	(0.081)	(0.095)
Null result × p-value framing	−5.214**	−0.296**	−0.286**	0.140	0.088
	(2.430)	(0.136)	(0.141)	(0.124)	(0.135)
p-value framing	−2.824	0.022	−0.032	−0.066	−0.104
	(2.091)	(0.114)	(0.118)	(0.114)	(0.120)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250

Note: This table shows regression estimates of our treatment effects on our key outcomes of interest. The data set is at the vignette-respondent level and contains four observations for each respondent. Each panel reports results from a separate set of pre-registered regressions (see equation 2). We include respondent fixed effects in all regressions, except in panel E (which we pre-registered) where the p-value framing only varies between respondents. All regressions include treatment indicators for all other cross-randomised conditions in addition to vignette fixed effects. Each panel includes a separate interaction of the *null result* treatment indicator with indicators related to the cross-randomised feature indicated by the panel's header. 'Null result treatment' is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. 'Low expert forecast' and 'high expert forecast' are treatment indicators taking the value one if the group of experts respectively predicted a low or high treatment effect estimate (and zero otherwise). 'Field journal' is a treatment indicator taking the value one if the vignette included a field journal and zero if it included a general interest journal. 'PhD student' is a treatment indicator taking the value one if the team behind the vignette research study included PhD students and zero if it included professors. 'Low-ranked university' is a treatment indicator taking the value one if the team behind the vignette research study was affiliated with a lower-ranked university and zero if it was affiliated with a higher-ranked university. 'p-value framing' is a treatment indicator taking the value one if the vignette treatment effect had an associated p-value and zero if it had an associated standard error estimate. ***denotes statistical significance at the 0.01 threshold. **denotes statistical significance at the 0.05 threshold. *denotes statistical significance at the 0.10 threshold

2.4. *Do p-Values Aggravate the Null Result Penalty?*

To test whether the presentation of results affects the penalty for studies with null results, we vary between respondents whether the statistical uncertainty associated with the main effect is communicated in terms of standard errors or *p*-values. Column (1) in panel E of Table 4 shows that communicating the results in terms of *p*-values rather than standard errors somewhat decreases the perceived publishability and strongly decreases the perceived quality of research papers reporting main findings that are not significant.

The null result penalty on perceived publishability is 5.2 percentage points higher when results are presented in terms of *p*-values rather than standard errors (95% CI [−10, −0.4]; $p = 0.032$). This effect is robust across a wide range of alternative specifications (as shown in Online Appendix Table A.1). Similarly, when the results are presented displaying *p*-values instead of standard errors, the negative effects of the *null result* treatment on both first-order beliefs about quality and beliefs about others' quality assessments are further increased by 29.6% of an SD (95% CI [−56.5, −2.8]; $p = 0.03$) and 28.6% of an SD (95% CI [−56.4, −0.8]; $p = 0.044$), respectively.

Overall, this evidence suggests that individuals may rely on simple heuristics to evaluate research results, consistent with cognitive constraints playing an important role (Benjamin *et al.*, 2013).[14]

3. Mechanisms

3.1. *A Preference for Publishing Surprising Findings?*

The scarcity of available journal space necessitates the adoption of publication rules that maximise a chosen social objective. While several social objectives are plausible, a key trade-off arises between the objective of maximising the policy impact of published studies and the goal of maintaining the validity of statistical inference about true effect sizes based on published studies. Frankel and Kasy (2022) provided a first formalisation of policy relevance grounded in economic theory. In their framework, maximising the policy impact of published findings requires the publication process to favour research studies that are 'surprising' relative to the profession's prior, while maintaining valid inference requires that the publication process does not condition publication on the statistical significance of a study's findings.

One could thus potentially rationalise the null result penalty if referees and editors mainly care about the policy impact of published studies. If respondents expect such a preference to be common, the null result penalty we document in the previous section should be more severe for null results that are predicted by experts and attenuated for those null results that conflict with expert priors.

To test this conjecture, we examine heterogeneity by whether the experts predicted a large and significant effect or a small and statistically non-significant effect. Panel A of Table 4 shows interaction effects between the *null result* treatment and treatment indicators for being shown a high expert forecast or a low expert forecast. As shown in column (1), respondents provided with a low expert forecast instead of a no expert forecast do not differentially update their beliefs about the publishability of the study in a statistically significant way (95% CI [−6.9, 2.9]; $p = 0.42$). In

[14] Respondents in the *p*-value treatment do not learn about the standard error, but could in principle back out the standard error implied by the coefficient estimate and the associated *p*-value. Yet, this calculation is likely too complex for our respondents, leading them to rely on simple heuristics instead.

contrast, respondents in the *null result* treatment who receive the a high expert forecast instead of a no expert forecast think the studies have a 6.4 percentage point lower chance of being published (95% CI [−11.6, −1.2]; $p = 0.016$).[15] In other words, the negative effect of obtaining a small and not statistically significant result on perceived publication chances is even exacerbated when experts predict a large and statistically significant effect, suggesting that the null result penalty is not driven by a desire to reward surprising findings in the publication process.

Our findings suggest that people *perceive* the publication process not to be perfectly in line with either of the two objectives outlined above. First, the substantial perceived penalty against null results is at odds with the objective of valid inference about true effect sizes based on published studies. Second, our participants believe that 'surprising' null results—i.e., null findings in contexts where experts predict a large treatment effect—have lower publication prospects compared to unsurprising nulls. This is the opposite of what the model by Frankel and Kasy (2022) on maximising the policy impact of published findings would suggest.[16]

3.2. *Learning About Unobservables?*

One explanation for the null result penalty is that researchers might draw negative inference about the quality of studies with null results, thus lowering their perceived publication chances. In our vignettes, we fix beliefs about several study characteristics, such as the sample size or the statistical precision. However, participants might update about unobserved dimensions of quality, such as the quality of the experimental instructions, the adherence to the experimental protocol or the integrity of the statistical analyses.

One potential way to assess whether participants draw inference about unobservable study characteristics is to look at heterogeneous effects of study features that affect the dispersion of priors about quality. In particular, it is plausible that respondents have more diffuse priors about the quality of research studies conducted by researchers who are affiliated with lower-ranked institutions or those who have less research experience. The Bayesian prediction would thus be that researchers should update more strongly about the quality of a study if the authors are either affiliated with lower-ranked institutions or of lower seniority. We find that researchers expect articles of PhD students and researchers from lower-ranked universities to be less likely to be published even though they do not perceive any quality differences (columns (1) and (2) in panels C and D of Table 4). Yet, as discussed in Section 2.3, we find only muted interaction effects between the *null result* treatment and an indicator for vignettes in which the research team is composed of PhD students or is affiliated with a lower-ranked university on our main outcomes of interest (see column (1) of Table 4).[17] The lack of heterogeneous treatment effects thus provides suggestive evidence against learning about unobservables playing a quantitatively important role, though naturally the heterogeneous effects are less precisely estimated compared to the main effects.[18]

[15] The two interaction effects are marginally significantly different from each other ($p = 0.073$).

[16] We elicited beliefs about the status quo rather than normative beliefs. It is thus conceivable that our participants may think that the publication process *should* maximise one of the two social objectives.

[17] Similarly, Online Appendix Table A.2 shows that we obtain virtually identical treatment effects when we restrict the sample to the subset of vignettes that describe the study to be conducted by a research team with ex ante plausibly higher research quality (a research team consisting of professors from higher-ranked universities).

[18] A complementary way of examining this mechanism is to exploit exogenous variation in prior beliefs in a true causal relationship in the context of the study described in a specific vignette. Section D of the Online Appendix provides an extended discussion of this approach, including a formalisation of its requirements and an empirical test.

Table 5. *Main Results: Mechanism Experiment on Perceived Precision.*

	(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 fixed effects		
Null result treatment	−18.134***	−1.086***
	(2.605)	(0.148)
Observations	475	475
Respondents	95	95

Note: The table shows regression estimates of our treatment effects on our key outcomes of interest from equation (1) using data from the mechanism experiment (see Section 3.3). The data set is at the vignette-respondent level and contains five observations for each respondent. 'Null result treatment' is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. The regressions in panel A (panel B) include (do not include) respondent fixed effects. All regressions in both panels include treatment indicators for the cross-randomised conditions in addition to vignette fixed effects. ***denotes statistical significance at the 0.01 threshold.

3.3. *Perceived Statistical Precision*

We conducted an additional pre-registered experiment to examine whether beliefs about the statistical precision of a study depend on whether the study yielded a large and statistically significant result or a small and statistically non-significant result while holding constant information about the actual precision of the estimate.

3.3.1. *Sample*

In May 2022, we invited 509 graduate students and early career researchers in the field of economics to participate in a 10-min online survey. In total, 95 graduate students and early career researchers follow our invitation and complete the survey, implying a response rate of 19%. These respondents are affiliated with one of the following institutions: University of Oxford, Universitat Pompeu Fabra, University of Cologne, University of Bonn, NHH Norwegian School of Economics and the University of Zurich.

3.3.2. *Design*

We examine whether researchers perceive studies with null results to be less precisely estimated, even when they are provided with the standard error of the estimate. The design is identical to our main experiment except for two differences. First, respondents are asked to rate the statistical precision of the main result on a five-point Likert scale ranging from (1) *very imprecisely estimated* to (5) *very precisely estimated*. This measure of perceived statistical precision replaces the questions on perceived quality and importance of the study from the main experiment. Second, respondents are shown all five vignettes.

3.3.3. *Results*

Panel A of Table 5 presents treatment effects on our key outcomes of interest. First, as shown in column (1), we replicate our main finding that research studies with null results are perceived to be less publishable: respondents in the *null result* treatment think that the studies have a 19.8 percentage point lower probability of being published (95% CI [−24.3, −15.2]; $p < 0.001$),

corresponding to a 32.5% reduction in perceived publication chances. Second, column (2) provides support for the hypothesis that null results lead respondents to associate the corresponding studies with lower statistical precision: even though we keep the sample size and standard errors constant across conditions, respondents in the *null result* treatment associate the research studies with 126.7% of an SD lower statistical precision (95% CI [−155.2, −98.2]; $p < 0.001$).[19]

3.3.4. *Is the null result penalty a bias?*

The evidence on the perceived statistical precision is inconsistent with Bayesian explanations of learning about unobservables and suggests that at least some of the penalty may be driven by a bias. Researchers' beliefs about the precision of coefficient estimates are thus influenced by the coefficient's statistical significance, even though standard errors are identical. These findings suggest that researchers may use simple heuristics to assess the statistical precision of estimates.

4. Robustness

This section provides additional tests to examine whether our treatment effects are robust to alternative approaches of analysing the data.

4.1.1. *Selection into the survey*

A potential concern related to our response rate of 3.4% could be that a potential selection bias into the survey could make our results less externally valid. If anything, our sample appears 'positively selected' on observable markers of professional success (as shown in Table 1). Reassuringly, panels B and D of Online Appendix Table A.3 show that our results are robust to re-weighting our sample to match the marginal distribution of four characteristics (gender, region, repeated top five referee dummy and current editor dummy) in the study population of researchers affiliated with a top 200 institution according to RePEc (as of March 2022), mitigating concerns about the external validity of our findings.[20]

4.1.2. *Between versus within variation*

Our main design relies on within-person variation, which raises potential concerns. First, respondents might be more likely to guess the researchers' hypothesis in the process of seeing multiple vignettes. Second, respondents' attention and effort might be somewhat lower at the end of the survey, inducing them to increasingly rely on heuristics in evaluating the research studies. To deal with these concerns, Online Appendix Table A.5 shows estimates if we restrict the sample to the first, randomly selected vignette presented to each respondent. The table shows that we obtain quantitatively similar point estimates, indicating the robustness of our results. More broadly, treatment effects do not interact with the order of the vignette presented ($p = 0.660$).

4.1.3. *Statistical power*

The hypothetical research studies included in our experiments had the statistical power to detect relatively precisely estimated effects. This means that the null results included in our study were informative about the underlying effect sizes. In cases of underpowered studies with less precisely

[19] For ease of interpretation, we z-score the five-point Likert scale outcome using the *significant result* treatment group mean and SD. Treatment effects are therefore reported in terms of SDs.

[20] Online Appendix Table A.4 presents summary statistics for the re-weighted data. Online Appendix B.3 contains details on the construction of weights.

estimated effect sizes, a null result paper might not only be punished in the publication system because referees associate null result studies with lower quality, but also because they are hesitant to publish studies with imprecisely estimated effects. For studies with statistically significant effects, however, referees might put less attention on the statistical precision of the estimates as long as the *p*-value is below the significance threshold. The vignettes included in our experiment were all fairly highly powered, but some of the vignettes were more powered than others. To examine whether our treatment effects are smaller for studies with higher statistical power, panel B of Online Appendix Table A.6 restricts the sample to the subset of vignettes with high levels of statistical power, while panel C shows the results for vignettes with less-powered research studies.[21] We obtain quantitatively similar point estimates for the subset of highly powered vignettes, suggesting that concerns about underpowered studies that yield imprecise null effects are unlikely to explain our main results. This is more broadly consistent with the finding that the null result penalty is fairly homogeneous across the vignettes (see Online Appendix Figure A.2).

4.1.4. *Multiple hypothesis adjustment*

To examine the robustness of our results to multiple testing, we also report adjusted *p*-values in Online Appendix Table A.7 based on our pre-specified specification (Romano and Wolf, 2005). Specifically, we implement a conservative procedure in which we correct for all five outcomes (our main outcome on perceived publishability and the four secondary beliefs on quality and importance of the study) as well as seven hypothesis tests per outcome (the null indicator and six interaction effects).

Online Appendix Table A.7 underscores the robustness of our main treatment effects of this conservative adjustment for multiple hypothesis testing. The heterogeneous treatment effects we document are also largely robust to the multiple hypothesis adjustments. Specifically, the *p*-values of the interaction effects of receiving the high expert forecast remain statistically significant after the adjustment. The interaction effect between the null result treatment and the *p*-value interaction on the publishability outcome is also still marginally significant ($p = 0.08$ after adjusting; $p = 0.03$ before adjusting). On quality perceptions, the effect on first-order beliefs is still marginally significant ($p = 0.07$), while the effect on second-order beliefs is not statistically significant ($p = 0.11$). Results remain similar in a fully interacted model with all treatment indicators included simultaneously in the regressions (see Online Appendix Table A.8). In this specification, which was not pre-specified, but has the advantage of being more efficient (Tsiatis *et al.*, 2008; Lin, 2013), almost all of our main results remain statistically significant at the 5% level. The only exception is the interaction effect between the *p*-value indicator and the null result treatment on the publishability outcome, which is no longer statistically significant at conventional levels in the fully interacted model after adjusting for multiple hypothesis testing.

4.1.5. *Match between researcher expertise and vignette*

One concern about the external validity of our study relates to the match between the field of the study and the field of the evaluators. Given that our studies all leverage methods from applied microeconomics, we restrict the sample to researchers working in empirical microeconomics fields. These researchers have higher exposure to experimental work and are therefore also likely in a better position to judge the statistical power of the research designs presented in our

[21] The minimum detectable effect size at 80% statistical power for the vignettes on the marginal effects of merit aid for low-income students, the financial literacy program and the salience of poverty and patience is (below) 20% of an SD, which is a commonly accepted threshold across experimental fields.

vignettes. The estimates in panel C of Online Appendix Table A.6 indicate that we also obtain similar point estimates for this sample. Similarly, Online Appendix Table A.9 shows that our results are robust to restricting the sample to researchers who work in fields that are covered by our vignettes.[22]

4.1.6. *Effort and attention*

Finally, a potential concern is that respondents might exert little effort when evaluating hypothetical research studies. First, we examine time spent on the survey and on different vignettes as a proxy for respondents' effort and attention. As shown in Online Appendix Table A.10, the median time spent on the overall survey was 451 s, while respondents spent a median of 94 s on each vignette. We have relatively few respondents speeding through the survey. For instance, only 33 respondents completed the survey in under 4 min. Second, Online Appendix Table A.11 shows that respondents spent more time on vignettes with longer instructions, such as those including expert forecasts. Third, Online Appendix Figure A.3 shows that vignette response times are very similar across the *p*-value and standard error treatment arms, suggesting that this treatment variation did not differentially affect respondents' attention to the experimental instructions. Fourth, Online Appendix Table A.12 shows that we obtain virtually identical and, if anything, slightly larger treatment effects on publishability when we restrict the sample to respondents who spent more time on the survey. Taken together, this underscores that most respondents spent a reasonable time carefully evaluating each vignette and that our results are not driven by inattentive respondents.

As a final check to identify low-effort respondents, we examine what fraction of respondents always provide the same answer across vignettes. Consistent with high levels of effort, only 1% of respondents always provide the same response when asked about the publication chances, while only 2.1% of respondents provide responses that differ by less than 5 percentage points across vignettes (the results are robust to excluding these respondents).

5. Conclusion

We show that research studies with small and not statistically significant effects are perceived to be less publishable, of lower quality, of lower importance and less precisely estimated than studies with large and statistically significant results, even when holding constant all other study features, including the statistical precision of estimates. Small and not statistically significant effects are considered even less publishable when experts predict a large effect, suggesting that the null result penalty is not driven by a desire to reward surprising results in the publication process. Communicating the statistical uncertainty of study results in terms of *p*-values rather than standard errors further aggravates the null result penalty.

Our findings highlight the potential value of a pre-result review in which the decision on publication is taken before the empirical results are known (Camerer *et al.*, 2019; Bogdanoski *et al.*, 2020; 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 and importance of null results (Abadie, 2020). Finally, one practical implication of our study is that

[22] This robustness check also addresses the concern that researchers who are familiar with the research fields of our vignettes might have recognised some of the original studies on which the vignettes are based and might have been able to infer the study's main hypothesis more easily in case they were randomly assigned to the *null result* treatment (which represented a deviation from the original studies that all reported statistically significant findings).

communicating statistical uncertainty of estimates in terms of standard errors rather than *p*-values might help to counteract negative updating about the quality of null result studies.

While our paper documents a clear penalty against null results, it is important to emphasise that our design keeps the standard error of the estimates constant, implying equal power across treatments. In the real world, some studies might yield null results because of research designs with low statistical power, leading to point estimates with large confidence intervals that are not very informative about whether there is an underlying effect or not. In such cases, it is more appropriate to talk about a penalty for noisy estimates rather than a bias against null results. Future work should investigate the extent to which studies with low statistical precision are discounted in the publication process and how this depends on whether they yielded a null result. Future work should also examine whether the *p*-value framing leads to a large and robust decrease in perceptions of quality because the *p*-value is a simple heuristic to judge the quality of a study or whether there are other behavioural explanations behind this anomaly.

Improving our understanding of how the design of the publication process affects the magnitude of the null result penalty is important as a widespread belief in such a penalty likely has direct implications for researchers' incentives and therefore the production of scientific research (Glaeser, 2006). Both first-order beliefs and higher-order beliefs in a null result penalty may affect which projects researchers pursue and whether they opt to move projects to the file drawer or submit their findings for publication. Moreover, a null result penalty could also negatively impact projects prior to the publication process. For instance, projects with null results might have more difficulty attracting research funding or getting accepted at scientific conferences. A null result penalty could thus have a negative compounding effect by raising the bar for null result studies prior to the peer review process. Future research could thus examine how beliefs about the null result penalty shape researcher decisions in different contexts and whether interventions shifting the perceived value or publication chances of studies with null results can change the production decisions of scientists. Future research could also examine the robustness of the null result penalty across different contexts. Our vignettes, which gave respondents a short summary of the study, should have a comparatively high external validity for initial screening decisions, such as when an editor decides on whether to send out a paper for review. However, given the lack of incentives in our experiment as well as the relatively short time to evaluate each study, our results might be less informative about outcomes at later stages in the publication process, such as referee recommendations.

University of Copenhagen, Denmark
NHH Norwegian School of Economics, Norway
University of Cologne, ECONtribute, briq & Max-Planck Institute for Collective Goods, Germany & CEPR, UK
University of Warwick, CEPR, UK & briq, Germany

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

References

Abadie, A. (2020). 'Statistical nonsignificance in empirical economics', *American Economic Review: Insights*, vol. 2(2), pp. 193–208.

Andre, P. and Falk, A. (2021). 'What's worth knowing? Economists' opinions about economics', Discussion paper, ECONtribute.

Andre, P., Haaland, I., Roth, C. and Wohlfart, J. (2022a). 'Narratives about the macroeconomy', Discussion Paper DP17305, Centre for Economic Policy Research.

Andre, P., Pizzinelli, C., Roth, C. and Wohlfart, J. (2022b). 'Subjective models of the macroeconomy: Evidence from experts and representative samples', *The Review of Economic Studies*, vol. 89(6), pp. 2958–91.

Andrews, I. and Kasy, M. (2019). 'Identification of and correction for publication bias', *American Economic Review*, vol. 109(8), pp. 2766–94.

Benjamin, D.J., Brown, S.A. and Shapiro, J.M. (2013). 'Who is "behavioral"? Cognitive ability and anomalous preferences', *Journal of the European Economic Association*, vol. 11(6), pp. 1231–55.

Berinsky, A.J., Druckman, J.N. and Yamamoto, T. (2021). 'Publication biases in replication studies', *Political Analysis*, vol. 29(3), pp. 370–84.

Blanco-Perez, C. and Brodeur, A. (2020). 'Publication bias and editorial statement on negative findings', Economic Journal, vol. 130(629), pp. 1226–47.

Bogdanoski, A., Foster, A., Karlan, D. and Miguel, E. (2020). 'Pre-results review at the journal of development economics: Lessons learned', MetaArXiv, doi:10.31222/osf.io/5yacr.

Brodeur, A., Carrell, S., Figlio, D. and Lusher, L. (2021). 'Unpacking p-hacking and publication bias'. Technical Report, University of Ottawa.

Brodeur, A., Cook, N. and Heyes, A. (2020). 'Methods matter: p-hacking and publication bias in causal analysis in economics', *American Economic Review*, vol. 110(11), pp. 3634–60.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016). 'Star wars: The empirics strike back', *American Economic Journal: Applied Economics*, vol. 8(1), pp. 1–32.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016). 'Evaluating replicability of laboratory experiments in economics', *Science*, vol. 351(6280), pp. 1433–6.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J. and Wu, H. (2018). 'Evaluating the replicability of social science experiments in nature and science between 2010 and 2015', *Nature Human Behaviour*, vol. 2(9), pp. 637–44.

Camerer, C.F., Dreber, A. and Johannesson, M. (2019). 'Replication and other practices for improving scientific quality in experimental economics', in (A. Schram and A. Ule, eds.), *Handbook of Research Methods and Applications in Experimental Economics*, pp. 83–103, Cheltenham: Edward Elgar Publishing.

Card, D. and DellaVigna, S. (2013). 'Nine facts about top journals in economics', *Journal of Economic Literature*, vol. 51(1), pp. 144–61.

Card, D. and DellaVigna, S. (2020). 'What do editors maximize? Evidence from four economics journals', *Review of Economics and Statistics*, vol. 102(1), pp. 195–217.

Card, D., DellaVigna, S., Funk, P. and Iriberri, N. (2020). 'Are referees and editors in economics gender neutral?', *Quarterly Journal of Economics*, vol. 135(1), pp. 269–327.

Casey, K., Glennerster, R. and Miguel, E. (2012). 'Reshaping institutions: Evidence on aid impacts using a preanalysis plan', *Quarterly Journal of Economics*, vol. 127(4), pp. 1755–812.

Christensen, G., Freese, J. and Miguel, E. (2019). *Transparent and Reproducible Social Science Research*, Berkeley, CA: University of California Press.

Christensen, G. and Miguel, E. (2018). 'Transparency, reproducibility, and the credibility of economics research', *Journal of Economic Literature*, vol. 56(3), pp. 920–80.

de Quidt, J., Haushofer, J. and Roth, C. (2018). 'Measuring and bounding experimenter demand', *American Economic Review*, vol. 108(11), pp. 3266–302.

DellaVigna, S. and Pope, D. (2018). 'Predicting experimental results: Who knows what?', *Journal of Political Economy*, vol. 126(6), pp. 2410–56.

DellaVigna, S., Pope, D. and Vivalt, E. (2019). 'Predict science to improve science', *Science*, vol. 366(6464), pp. 428–9.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B.A. and Johannesson, M. (2015). 'Using prediction markets to estimate the reproducibility of scientific research', *Proceedings of the National Academy of Sciences*, vol. 112(50), pp. 15343–7.

Dufwenberg, M. and Martinsson, P. (2014). 'Keeping researchers honest: The case for sealed-envelope-submissions', Working Paper 533, Innocenzo Gasparini Institute for Economic Research.

Dwan, K., Altman, D.G., Arnaiz, J.A., Bloom, J., Chan, A.W., Cronin, E., Decullier, E., Easterbrook, P.J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J.P.A., Simes, J. and Williamson, P.R. (2008). 'Systematic review of the empirical evidence of study publication bias and outcome reporting bias', *PloS One*, vol. 3(8), e3081.

Elson, M., Huff, M. and Utz, S. (2020). 'Metascience on peer review: Testing the effects of a study's originality and statistical significance in a field experiment', *Advances in Methods and Practices in Psychological Science*, vol. 3(1), pp. 53–65.
Emerson, G.B., Warme, W.J., Wolf, F.M., Heckman, J.D., Brand, R.A. and Leopold, S.S. (2010). 'Testing for the presence of positive-outcome bias in peer review: A randomized controlled trial', *Archives of Internal Medicine*, vol. 170(21), pp. 1934–9.
Ersoy, F. and Pate, J. (2021). 'Invisible hurdles: Gender and institutional bias in the publication process in economics', Preprint, http://dx.doi.org/10.2139/ssrn.3870368.
Franco, A., Malhotra, N. and Simonovits, G. (2014). 'Publication bias in the social sciences: Unlocking the file drawer', *Science*, vol. 345(6203), pp. 1502–5.
Frankel, A. and Kasy, M. (2022). 'Which findings should be published?', *American Economic Journal: Microeconomics*, vol. 14(1), pp. 1–38.
Gerber, A. and Malhotra, N. (2008). 'Do statistical reporting standards affect what is published? Publication bias in two leading political science journals', *Quarterly Journal of Political Science*, vol. 3(3), pp. 313–26.
Glaeser, E.L. (2006). 'Researcher incentives and empirical methods', Working Paper 0329, National Bureau of Economic Research.
Greenwald, A.G. (1975). 'Consequences of prejudice against the null hypothesis.', *Psychological Bulletin*, vol. 82(1), pp. 1–20.
Haaland, I., Roth, C. and Wohlfart, J. (2023). 'Designing information provision experiments', *Journal of Economic Literature*, vol. 61(1), pp. 3–40.
Hjort, J., Moreira, D., Rao, G. and Santini, J.F. (2021). 'How research affects policy: Experimental evidence from 2,150 Brazilian municipalities', *American Economic Review*, vol. 111(5), pp. 1442–80.
Ioannidis, J.P. (2005). 'Why most published research findings are false', *PLoS Medicine*, vol. 2(8), e124.
Kasy, M. (2019). 'Selective publication of findings: Why does it matter, and what should we do about it?', MetaArXiv, doi:10.31222/osf.io/xwngs.
Kasy, M. (2021). 'Of forking paths and tied hands: Selective publication of findings, and what economists should do about it', *Journal of Economic Perspectives*, vol. 35(3), pp. 175–92.
Klein, R.A., Ratliff, K.A., Vianello, M., Adams, R.B., Jr, Bahnik, S., Bernstein, M.J., Bocian, K., Brandt, M.J., Brooks, B., Brumbaugh, C.C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W.E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E.M., Hasselman, F., Hicks, J.A., Hovermale, J.F., Hunt, S.J., Huntsinger, J.R., IJzerman, H., John, M.-S., Joy-Gaba, J.A., Kappes, H.B., Krueger, L.E., Kurtz, J., Levitan, C.A., Mallett, R.K., Morris, W.L., Nelson, A.J., Nier, J.A., Packard, G., Pilati, R., Rutchick, A.M., Schmidt, K., Skorinko, J.L., Smith, R., Steiner, T.G., Storbeck, J., Van Swol, L.M., Thompson, D., van't Veer, A.E., Vaughn, L.A., Vranka, M., Wichman, A.L., Woodzicka, J.A. and Nosek, B.A. (2014). 'Investigating variation in replicability', *Social Psychology*, vol. 45(3), pp. 142–52.
Klein, R.A. et al. (2018). 'Many labs 2: Investigating variation in replicability across samples and settings', *Advances in Methods and Practices in Psychological Science*, vol. 1(4), pp. 443–90.
Lin, W. (2013). 'Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique', *The Annals of Applied Statistics*, vol. 7(1), pp. 295–318.
Miguel, E. (2021). 'Evidence on research transparency in economics', *Journal of Economic Perspectives*, vol. 35(3), pp. 193–214.
Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K.M., Gerber, A., Glennerster, R., Green, D.P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B.A., Petersen, M., Sedlmayr, R., Simmons, J.P., Simonsohn, U. and der Laan, M.V. (2014). 'Promoting transparency in social science research', *Science*, vol. 343(6166), pp. 30–1.
Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D.P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., Mcnutt, M., Miguel, E., Levy Paluck, E., Simonsohn, U., Soderberg, C., Spellman, B.A., Turitto, J., Vandenbos, G., Vazire, S., Wagenmakers, E.J., Wilson, R. and Yarkoni, T. (2015). 'Promoting an open research culture', *Science*, vol. 348(6242), pp. 1422–5.
Nosek, B.A., Spies, J.R. and Motyl, M. (2012). 'Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability', *Perspectives on Psychological Science*, vol. 7(6), pp. 615–31.
Open Science Collaboration. (2015). 'Estimating the reproducibility of psychological science', *Science*, vol. 349(6251), aac4716.
Popper, K. (1934). *The Logic of Scientific Discovery*, New York: Routledge.
Romano, J.P. and Wolf, M. (2005). 'Stepwise multiple testing as formalized data snooping', *Econometrica*, vol. 73(4), pp. 1237–82.
Simonsohn, U., Nelson, L.D. and Simmons, J.P. (2014a). 'P-curve: A key to the file-drawer', *Journal of Experimental Psychology: General*, vol. 143(2), pp. 534–47.
Simonsohn, U., Nelson, L.D. and Simmons, J.P. (2014b). 'p-curve and effect size: Correcting for publication bias using only significant results', *Perspectives on Psychological Science*, vol. 9(6), pp. 666–81.

Tsiatis, A.A., Davidian, M., Zhang, M. and Lu, X. (2008). 'Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach', *Statistics in Medicine*, vol. 27(23), pp. 4658–77.
Vivalt, E. and Coville, A. (2020). 'Policy-makers consistently overestimate program impacts', Working paper, University of Toronto.
Vivalt, E. and Coville, A. (2023). 'How do policymakers update their beliefs?'.*Journal of Development Economics*, vol. 165, p. 103121.
Wasserstein, R.L. and Lazar, N.A. (2016). 'The asa statement on p-values: Context, process, and purpose', *The American Statistician*, vol. 70(2), pp. 129–33.

The Economic Journal, **134** (*January*), 193–219 https://doi.org/10.1093/ej/uead060 © The Author(s) 2023. Published by Oxford University Press on behalf of Royal Economic Society. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Advance Access Publication Date: 7 August 2023