The Null Result Penalty

Felix Chopra Ingar Haaland Christopher Roth
Andreas Stegmann*

March 6, 2023

Abstract

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 pre-results review.

Keywords: Null Results, Publication Bias, Learning, Information, Scientific Communication

*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. Chopra: University of Copenhagen, CEBI, Felix.Chopra@econ.ku.dk; Haaland: NHH Norwegian School of Economics, Ingar.Haaland@nhh.no; Roth: University of Cologne, ECONtribute, CEPR, briq, Max-Planck Institute for Collective Goods Bonn, roth@wiso.uni-koeln.de; Stegmann: University of Warwick, CEPR, briq, Andreas.Stegmann@warwick.ac.uk.

1 Introduction

The scientific method is characterized by researchers testing hypotheses with empirical evidence (Popper, 1934). Evidence accumulates with the publication of studies in scientific journals. Scientific progress thus requires a well-functioning publication system that evaluates research studies without bias. However, the publication system may favor research studies reporting large and statistically significant results over research papers documenting small results that are not statistically significant (Camerer et al., 2016; Greenwald, 1975; Simonsohn et al., 2014a). 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 (Camerer et al., 2019; Kasy, 2021; Miguel, 2021; Nosek et al., 2012).

In this paper, we investigate whether researchers penalize 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.

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 randomize whether the point estimate of the main treatment effect was sizable 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-randomize 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 evalu-

ation of research studies depends on expert priors, we cross-randomize whether the vignette includes expert forecasts of the treatment effect. For vignettes including expert forecasts, we further randomize 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-randomize 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-randomize 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 assessment of the research studies—irrespective of how other researchers and editors may evaluate such studies. The data on second-order beliefs furthermore allows 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% C.I. [-16.2,-11.9]; $p < 0.001$) as well as 37.3% of a standard deviation lower quality (95% C.I. [-49.6,-25]; $p < 0.001$) and 32.5% of a standard deviation lower importance (95% C.I. [-43,-21.9]; $p < 0.001$). Our respondents also think that other researchers would associate studies that yield a null result with 46% of a standard deviation lower quality (95% C.I. [-58.1,-33.8]; $p < 0.001$) and 41.7% of a standard deviation (95% C.I. [-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 artifact 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 longstanding 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% C.I. [-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 a standard deviation lower quality (95% C.I. [-56.5,-2.8]; $p = 0.03$) and also makes them think that other researchers will associate the study with 28.6% of a standard deviation lower quality (95% C.I. [-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 favors 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 operationalization of policy relevance grounded in economic theory. In their model, policy relevance is maximized by selectively publishing studies with surprising findings

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

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 points lower chance of being published (95% C.I. [-11.6,-1.2]; $p = 0.016$). These patterns are inconsistent with the conjecture that respondents believe that the publication process favors 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 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 19.8 percentage point (or 32.5%) lower chance of being published (95% C.I. [-24.3,-15.2]; $p < 0.001$). Furthermore, they associate studies with null results with a 126.7% of a standard deviation lower precision (95% C.I. [-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, 2020; Card et al., 2020; Ersoy and Pate, 2021; Frankel and Kasy, 2022; Kasy, 2019) and in particular publication bias (Blanco-Perez and Brodeur, 2020; Brodeur et al., 2016, 2020; Dwan et al., 2008; Franco et al., 2014; Gerber and Malhotra, 2008; Ioannidis, 2005).[3] This literature has examined the extent to which null results

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

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

are less likely to be published (Simonsohn et al., 2014a). Brodeur et al. (2016) study 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 favored in the peer review system. Brodeur et al. (2021) show that initial submissions display significant bunching in p-values, suggesting 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 show that reviewer recommendations are affected significantly by statistical thresholds, consistent with marginally significant results being favored 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 (Christensen and Miguel, 2018; Dufwenberg et al., 2014; Miguel et al., 2014; Nosek et al., 2015), such as the effect of editorial statements emphasizing the potential merit of scientific studies irrespective of the statistical significance of their main empirical estimates (Blanco-Perez and Brodeur, 2020). Our experimental approach allows us to study additional, potential measures to mitigate the null result penalty, such as providing expert forecasts

[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) examine publication bias using audit studies with medical scientists and psychologists, respectively. Berinsky et al. (2021) employ conjoint experiments with a sample of political scientists, focusing on publication biases in the context of replication studies.

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 (Andre and Falk, 2021; Andre et al., 2022a,b; Casey et al., 2012; DellaVigna and Pope, 2018; Dreber et al., 2015) and policymakers (Hjort et al., 2021; Vivalt and Coville, 2020, 2019). 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 2 describes the sample and the experimental design. In Section 3, we present the main results, and heterogeneous treatment effects. We present evidence on mechanisms in Section 4. Section 5 examines the robustness of our results. Finally, Section 6 discusses the implications of our findings for the publication system and the production of research.

2 Experimental design and data

2.1 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-minute 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%. 24.4% of our respondents 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, sizable 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 labor economics (21.1%), econometrics (14.1%), development economics (17.9%), political economy (16.7%), finance (10.5%), behavioral 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.

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.

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.

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 5 publications	1.27		462	0.34	0
Number of top 5s refereed for	1.17		397		
Repeated top 5 referee	0.30		397	0.12	0
Research evaluation:					
Current editor	0.07		443	0.03	0
Current associate editor	0.13		441		
Ever editor	0.15		444		
Ever associate editor	0.19		441		
Professional memberships:					
NBER affiliate	0.08		454		
CEPR affiliate	0.17		451		
Academic fields:					
Labor	0.21		418		
Public	0.13		418		
Development	0.18		418		
Political	0.17		418		
Finance	0.11		418		
Experimental	0.06		418		
Behavioral	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). Section B.1 contains a description of each variable. Data on the underlying sampling population was shared by Peter Andre and Armin Falk (see Andre and Falk, 2021). Section B.2 describes how our measures differ from those obtained from Andre and Falk (2021), in particular, "Years since PhD" and "H-index."

2.2 Design

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 (labor, education, economic history, behavioral economics, development economics, and household finance) and methods (randomized controlled trials, 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:** 3 Professors from Brown University conducted an RCT in Texas in the years 2015–2019. 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. 594 of those students 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.

Null result treatment We next provide respondents with information on the main result of the study (randomized 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 sizable 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 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 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 percent.

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 percent.

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

[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.

[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.

statistical significance of the displayed treatment effect estimate. We also randomize the order of vignettes.

Expert predictions We cross-randomize 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* treatment, respectively. We also provide the standard deviation of the expert forecast to communicate the degree of disagreement among experts. To ensure that there is scope for substantial updating, we set the standard deviation 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.

[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.

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 Lazar, 2016). To examine how the communication of statistical uncertainty affects the evaluation of studies with null results, we also cross-randomize 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 minimize the scope for experimenter demand effects (de Quidt et al., 2018), we cross-randomized 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-randomized 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 percent.

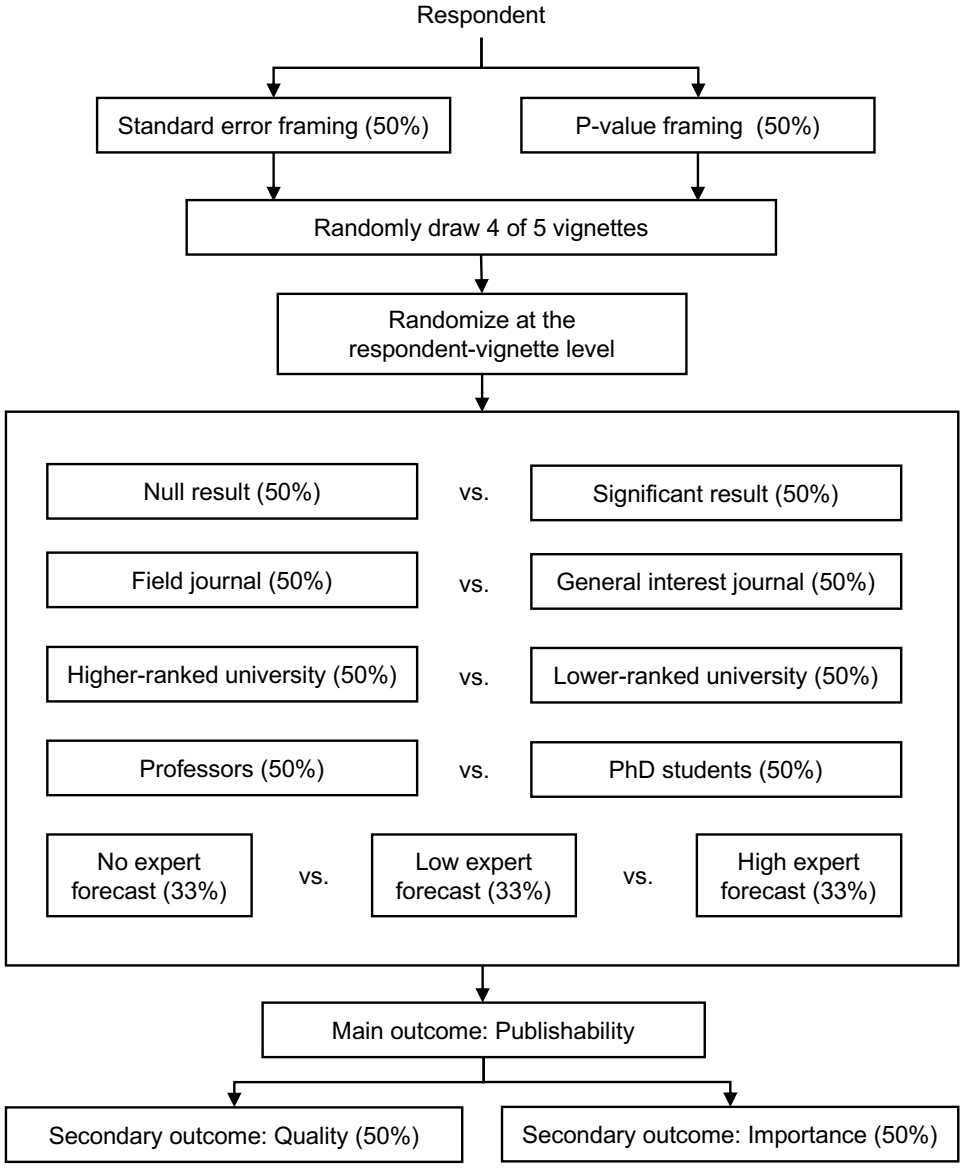
In contrast, respondents in the *null result* treatment cross-randomized to the standard error treatment receive the following instructions:

> **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 percent.

Obfuscation treatments We further cross-randomize 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-randomized 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

Figure 1: Overview of the factorial design



Note: This figure presents an overview of the factorial design, including the between-subject randomization of the scientific communication treatment (standard errors vs. *p*-value framing) and the between-subject randomization of secondary outcomes (first- and second-order perceptions of quality vs importance). The figure also presents the randomization of vignette features at the respondent-vignette level. The respondent-vignette level randomization includes five factors: statistical significance (2 levels), type of journal (2 levels), rank of the university that the researcher team is affiliated with (2 levels), seniority of the researcher team (2 levels), and the presence and magnitude of expert forecasts (3 levels). In the mechanism experiment (introduced in Section 4.3), respondents were shown all five hypothetical vignettes but the cross-randomization of features is otherwise identical.

Table 2: Overview of the vignettes

	Marginal effects of merit aid for low-income students	Long-term effects of equal land sharing	Female empowerment program	Financial literacy program	Salience of poverty and patience
	(1)	(2)	(3)	(4)	(5)
Panel A: General information					
Fields	Labor, Education	Economic History	Behavioral, Labor, Development	Development, Household Finance	Behavioral Economics
Country	USA	Germany	Sierra Leone	India	USA
Type of study	RCT	Regression discontinuity	RCT	RCT	Online experiment
Outcome	Completion of a 4-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 a std., 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
Standard deviation	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; 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; Science; PNAS: Proceedings of the National Academy of Sciences.

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 results treatment and to test whether respondents paid attention to the instructions.

Summary of 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), that yield $2 \times 2 \times 3 \times 2 \times 2 = 48$ factorial combinations. The 480 respondents in our main study complete 4 out of 5 vignettes providing us with 1,920 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 a standard deviation (corresponding to a 4 percentage point difference in perceived publication chances) at 80% power in our main specification and a significance threshold of 5%.

2.3 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-randomize 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-randomize whether respondents estimate the likelihood that the paper will

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

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 assessment of the research studies—irrespective of how other researchers and editors may evaluate such studies. The data on second-order beliefs furthermore allows 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-randomize 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' assessment of the importance of the research study using a similar wording as in the quality condition.

[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.

3 Main results

3.1 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:

$$y_{iv} = \alpha + \beta \text{null result}_{iv} + X_{iv}'\gamma + \delta_v + \tau_i + \varepsilon_{iv} \tag{1}$$

where "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-randomized 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-randomized 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-randomized features once at a time. Specifically, we estimate the following specification to examine heterogeneous treatment effects:

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

where β_z captures the effect of the interaction between the null result indicator and another cross-randomized 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.

3.2 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% C.I. [-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 a standard deviation lower quality (95% C.I. [-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 a standard deviation lower quality (95% C.I. [-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 sizable treatment effects on the perceived importance of the studies. Respondents in the *null result* treatment associate the studies with 32.5% of a standard deviation lower importance (95% C.I. [-43,-21.9]; $p < 0.001$) and think other researchers would associate the studies with 41.7% of a standard deviation lower importance (95% C.I. [-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 emphasize that the updating about quality and importance is large and negative both for personal beliefs and the beliefs of others.

3.3 Heterogeneity

Heterogeneity by respondent characteristics As discussed in Section 2.1, 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 analyze treatment heterogeneity by respondent characteristics. Figure 2 shows that the treatment effects are similar across different subgroups. For instance, the null

[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.

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 FE					
Null result treatment	-14.474***	-0.401***	-0.455***	-0.305***	-0.367***
	(1.224)	(0.069)	(0.072)	(0.062)	(0.069)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250

Note: The table shows regression estimates of our treatment effects on our key outcomes of interest from equation (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-randomized conditions in addition to vignette fixed effects.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

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 specialization, including among respondents who specialize 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]

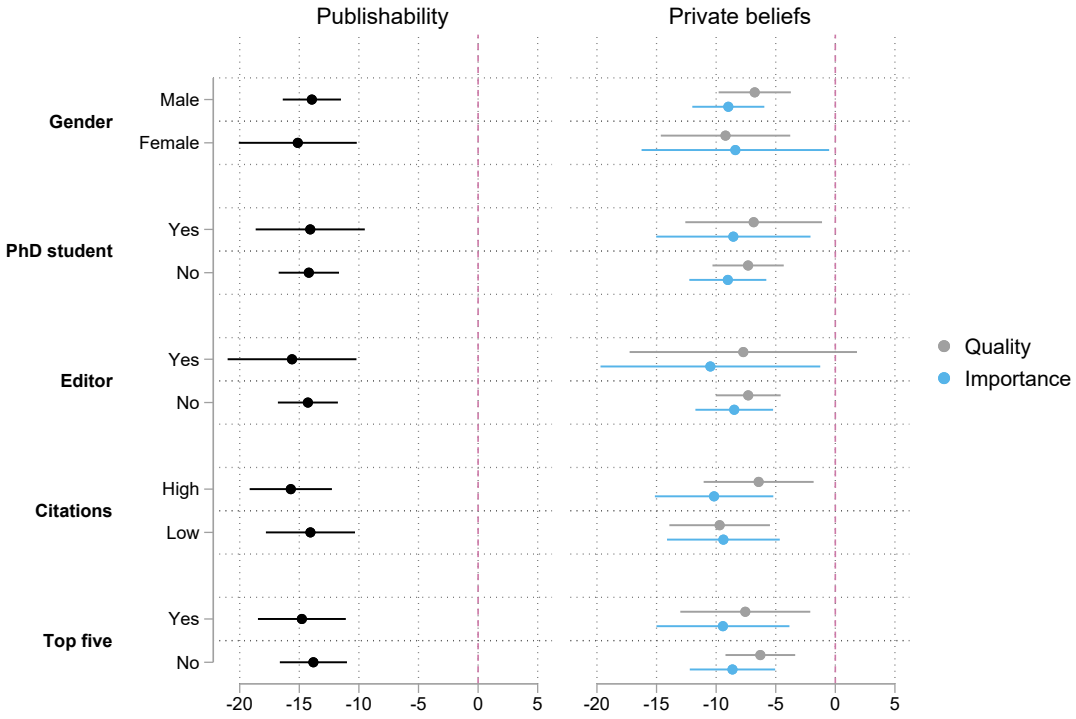
As shown in Figure 2, we also see a very similar belief in a null result penalty—both in terms of perceived publishability and personal beliefs about quality and importance of the research studies—among researchers with experience as editors of scientific journals and those who have never held an editorial position (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 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]

Heterogeneity by vignette characteristics While we included a set of cross-randomized conditions primarily to obfuscate the purpose of our study, the analysis of heterogeneous treatment effects of these cross-randomized "obfuscation treatments" also provides valuable insights on the determinants of the null result penalty. As shown in Panel B of

[12] We also see homogeneous effects across sub-groups in the mechanism experiment introduced in Section 4.3 (results available upon request).

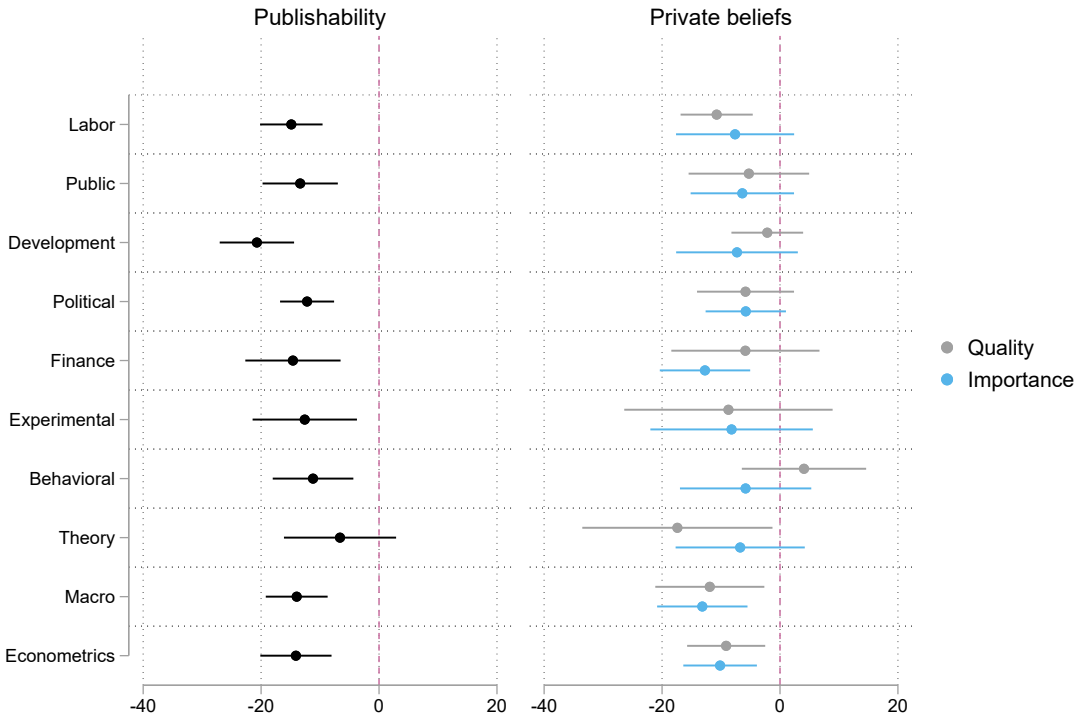
[13] A similar screening mechanism likely operates when evaluating papers for conference submissions or when choosing which job market candidates to interview.

Figure 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 sub-group indicated in the figure. Citations are measured using Google Scholar data as of May 2022 and "low" and 'high" refer to, respectively, 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 5" economics journals. All regressions include controls for the other cross-randomized features at the vignette level as well as respondent fixed effects. Standard errors are clustered at the respondent level. 95% confidence intervals are indicated in the figure.

Figure 3: Heterogeneity by field of specialization



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 sub-group indicated in the figure. The regressions include controls for all cross-randomized features at the vignette level as well as respondent fixed effects. Standard errors are clustered at the respondent level. 95% confidence intervals are indicated in the figure.

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 results 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 points higher probability of being published in field journals compared to general interest journals (95% C.I. [9.5,15]; $p < 0.001$). Moreover, they expect studies authored by PhD students rather than professors to have a 4.5 percentage points lower probability of being published (95% C.I. [-7.3,-1.8]; $p < 0.001$). Similarly, they expect studies conducted by researchers affiliated with lower-ranked universities to have 4.0 percentage points lower publication chances (95% C.I. [-6.7,-1.3]; $p = 0.004$). These results suggest that respondents read the vignettes attentively.

3.4 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% C.I. [-10,-0.4]; $p = 0.032$). This effect is robust across a wide range of alternative specifications (as shown in 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 assessment are further increased by 29.6% of a standard deviation (95% C.I. [-56.5,-2.8]; $p = 0.03$) and 28.6% of a standard deviation (95% C.I. [-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]

4 Mechanisms

4.1 A preference for publishing surprising findings?

The scarcity of available journal space necessitates the adoption of publication rules that maximize a chosen social objective. While several social objectives are plausible, a key trade-off arises between the objective of maximizing 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) provide a first formalization of policy relevance grounded in economic theory. In their framework, maximizing the policy impact of published findings requires the publication process to favor 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 rationalize 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 no expert forecast do not differentially update their beliefs about the publishability of the study in a statistically significant way (95% C.I. [-6.9,2.9]; $p = 0.42$). In contrast, respondents in the *null result* treatment who receive the high expert forecast instead of no expert

[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.

Table 4: Heterogeneity by vignette characteristics

	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: 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-randomized 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-randomized 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 predicted, respectively, 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.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

forecast think the studies have a 6.4 percentage points lower chance of being published (95% C.I. [-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 maximizing the policy impact of published findings would suggest.[16]

4.2 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

[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* maximize one of the two social objectives.

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 (column 1 and 2 in Panels C and D of Table 4). Yet, as discussed in Section 3.3, 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]

4.3 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.

Sample In May 2022, we invited 509 graduate students and early career researchers in the field of economics to participate in a 10-minute 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.

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

[17] Similarly, 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 formalization of its requirements and an empirical test.

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 5-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.

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% C.I. [-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 a standard deviation lower statistical precision (95% C.I. [-155.2,-98.2]; $p < 0.001$).[19]

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.

[19]For ease of interpretation, we z-score the 5-point Likert scale outcome using the *significant result* treatment group mean and standard deviation. Treatment effects are therefore reported in terms of standard deviations.

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 FE		
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 4.3). 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-randomized conditions in addition to vignette fixed effects.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

5 Robustness

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

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 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 5 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]

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 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 ($p = 0.660$).

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 with underpowered studies with less precisely 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

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

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 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 Figure A.2).

Multiple hypothesis adjustment To examine the robustness of our results to multiple testing, we also report adjusted p-values in 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).

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 Table A.8). In this specification, which was not

[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 a standard deviation, which is a commonly accepted threshold across experimental fields.

pre-specified but has the advantage of being more efficient (Lin, 2013; Tsiatis et al., 2008), 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.

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 vignettes. The estimates in Panel C of Table A.6 indicate that we also obtain similar point estimates for this sample. Similarly, 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]

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 Table A.10, the median time spent on the overall survey was 451 seconds, while respondents spent a median of 94 seconds on each vignette. We have relatively few respondents speeding through the survey. For instance, only 33 respondents completed the survey in under four minutes. Second, Table A.11 shows that respondents spent more time on vignettes with longer instructions, such as those including expert forecasts. Third, 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, Table A.12 shows that we obtain virtually identical and, if anything, slightly larger

[22]This robustness check also addresses the concern that researchers who are familiar with the research fields of our vignettes might have recognized 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 which all reported statistically significant findings).

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).

6 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 pre-results review in which the decision on publication is taken before the empirical results are known (Bogdanoski et al., 2020; Camerer et al., 2019; 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 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 emphasize 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 behavioral 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.

References

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

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

_ **, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart**, "Subjective models of the macroeconomy: Evidence from experts and representative samples," *The Review of Economic Studies*, 2022, *89* (6), 2958–2991.

_ **, Ingar Haaland, Chrisotpher Roth, and Johannes Wohlfart**, "Narratives about the Macroeconomy," *CEPR Discussion Paper No. DP17305*, 2022.

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

Benjamin, Daniel J., Sebastian A. Brown, and Jesse M. Shapiro, "Who is 'behavioral'? Cognitive ability and anomalous preferences," *Journal of the European Economic Association*, 2013, *11* (6), 1231–1255.

Berinsky, Adam J., James N. Druckman, and Teppei Yamamoto, "Publication Biases in Replication Studies," *Political Analysis*, 2021, *29* (3), 370–384.

Blanco-Perez, Cristina and Abel Brodeur, "Publication Bias and Editorial Statement on Negative Findings," *The Economic Journal*, 01 2020, *130* (629), 1226–1247.

Bogdanoski, Aleksandar, Andrew Foster, Dean Karlan, and Edward Miguel, "Pre-results Review at the Journal of Development Economics: Lessons learned," *MetaArXiv*, 2020.

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

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

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

Camerer, Colin F., Anna Dreber, and Magnus Johannesson, "Replication and other practices for improving scientific quality in experimental economics," *Handbook of Research Methods and Applications in Experimental Economics*, 2019.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan et al., "Evaluating replicability of laboratory experiments in economics," *Science*, 2016, *351* (6280), 1433–1436.

_ , _ , Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer et al., "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature human behaviour*, 2018, *2* (9), 637–644.

Card, David and Stefano DellaVigna, "Nine facts about top journals in economics," *Journal of Economic Literature*, 2013, *51* (1), 144–61.

_ and _ , "What do editors maximize? Evidence from four economics journals," *Review of Economics and Statistics*, 2020, *102* (1), 195–217.

_ , _ , Patricia Funk, and Nagore Iriberri, "Are referees and editors in economics gender neutral?," *Quarterly Journal of Economics*, 2020, *135* (1), 269–327.

Casey, Katherine, Rachel Glennerster, and Edward Miguel, "Reshaping institutions: Evidence on aid impacts using a preanalysis plan," *Quarterly Journal of Economics*, 2012, *127* (4), 1755–1812.

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

_ , Jeremy Freese, and Edward Miguel, "Transparent and reproducible social science research," in "Transparent and Reproducible Social Science Research," University of California Press, 2019.

Clarke, Damian, "RWOLF2: Stata Module to Calculate Romano-Wolf Stepdown p-Values for Multiple Hypothesis Testing," Statistical Software Components, Boston College Department of Economics 7 2021.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, "Measuring and Bounding Experimenter Demand," *American Economic Review*, November 2018, *108* (11), 3266–3302.

DellaVigna, Stefano and Devin Pope, "Predicting Experimental Results: Who Knows What?," *Journal of Political Economy*, 2018, *126* (6), 2410–2456.

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

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A Nosek, and Magnus Johannesson, "Using prediction markets to estimate the reproducibility of scientific research," *Proceedings of the National Academy of Sciences*, 2015, *112* (50), 15343–15347.

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

Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, Carrol Gamble et al., "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias," *PloS One*, 2008, *3* (8), e3081.

Elson, Malte, Markus Huff, and Sonja Utz, "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*, 2020, *3* (1), 53–65.

Emerson, Gwendolyn B., Winston J. Warme, Fredric M. Wolf, James D. Heckman, Richard A. Brand, and Seth S. Leopold, "Testing for the Presence of Positive-Outcome Bias in Peer Review: A Randomized Controlled Trial," *Archives of Internal Medicine*, 2010, *170* (21), 1934–1939.

Ersoy, Fulya and Jennifer Pate, "Invisible Hurdles: Gender and Institutional Bias in the Publication Process in Economics," *Available at SSRN 3870368*, 2021.

Franco, Annie, Neil Malhotra, and Gabor Simonovits, "Publication bias in the social sciences: Unlocking the file drawer," *Science*, 2014, *345* (6203), 1502–1505.

Frankel, Alexander and Maximilian Kasy, "Which findings should be published?," *American Economic Journal: Microeconomics*, 2022, *14* (1), 1–38.

Gerber, Alan and Neil Malhotra, "Do Statistical Reporting Standards Affect What is Published? Publication Bias in Two Leading Political Science Journals," *Quarterly Journal of Political Science*, 2008, *3* (3), 313–326.

Glaeser, Edward L., "Researcher Incentives and Empirical Methods," Technical Report 2006.

Greenwald, Anthony G., "Consequences of Prejudice Against the Null Hypothesis.," *Psychological Bulletin*, 1975, *82* (1), 1.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature*, 2023, *61* (1), 3–40.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini, "How research affects policy: Experimental evidence from 2,150 brazilian municipalities," *American Economic Review*, 2021, *111* (5), 1442–80.

Ioannidis, John P.A., "Why most published research findings are false," *PLoS Medicine*, 2005, *2* (8), e124.

Kasy, Maximilian, "Selective publication of findings: Why does it matter, and what should we do about it?," *MetaArXiv*, 2019.

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

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr, Stepan Bahnik, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh et al., "Investigating variation in replicability," *Social psychology*, 2014.

Klein, Richard A, Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R. Axt, Mayowa T. Babalola, Stepan Bahnik et al., "Many Labs 2: Investigating variation in replicability across samples and settings," *Advances in Methods and Practices in Psychological Science*, 2018, *1* (4), 443–490.

Lin, Winston, "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *The Annals of Applied Statistics*, 2013, *7* (1), 295–318.

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

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

Nosek, Brian A., George Alter, George C. Banks, Denny Borsboom, Sara D. Bowman, Steven J. Breckler, Stuart Buck, Christopher D. Chambers, Gilbert Chin, Garret Christensen et al., "Promoting an open research culture," *Science*, 2015, *348* (6242), 1422–1425.

_ **, Jeffrey R. Spies, and Matt Motyl**, "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability," *Perspectives on Psychological Science*, 2012, *7* (6), 615–631.

Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science*, 2015, *349* (6251), aac4716.

Pasek, Josh, Matthew Debell, and Jon A. Krosnick, "Standardizing and democratizing survey weights: The ANES weighting system and anesrake," *Working Paper*, 2014.

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

Romano, Joseph P. and Michael Wolf, "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 2005, *73* (4), 1237–1282.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons, "P-Curve: A Key to the File-Drawer.," *Journal of Experimental Psychology: General*, 2014, *143* (2), 534.

_ , _ , and _ , "p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results," *Perspectives on Psychological Science*, 2014, *9* (6), 666–681.

Tsiatis, Anastasios A., Marie Davidian, Min Zhang, and Xiaomin Lu, "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach," *Statistics in Medicine*, 2008, *27* (23), 4658–4677.

Vivalt, E. and A. Coville, "Policy-makers Consistently Overestimate Program Impacts," Technical Report, Working Paper 2020.

Vivalt, Eva and Aidan Coville, "How do Policymakers Update?," 2019.

Wasserstein, Ronald L. and Nicole A. Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 2016, *70* (2), 129–133.

For online publication only:

The Null Result Penalty

Felix Chopra, Ingar Haaland, Christopher Roth and Andreas Stegmann

Section A contains additional figures and tables.

Section B provides details on the background of the expert sample and the re-weighting procedure.

Section C provides a description of how we obtained the numerical features that vary across experimental conditions in our vignettes.

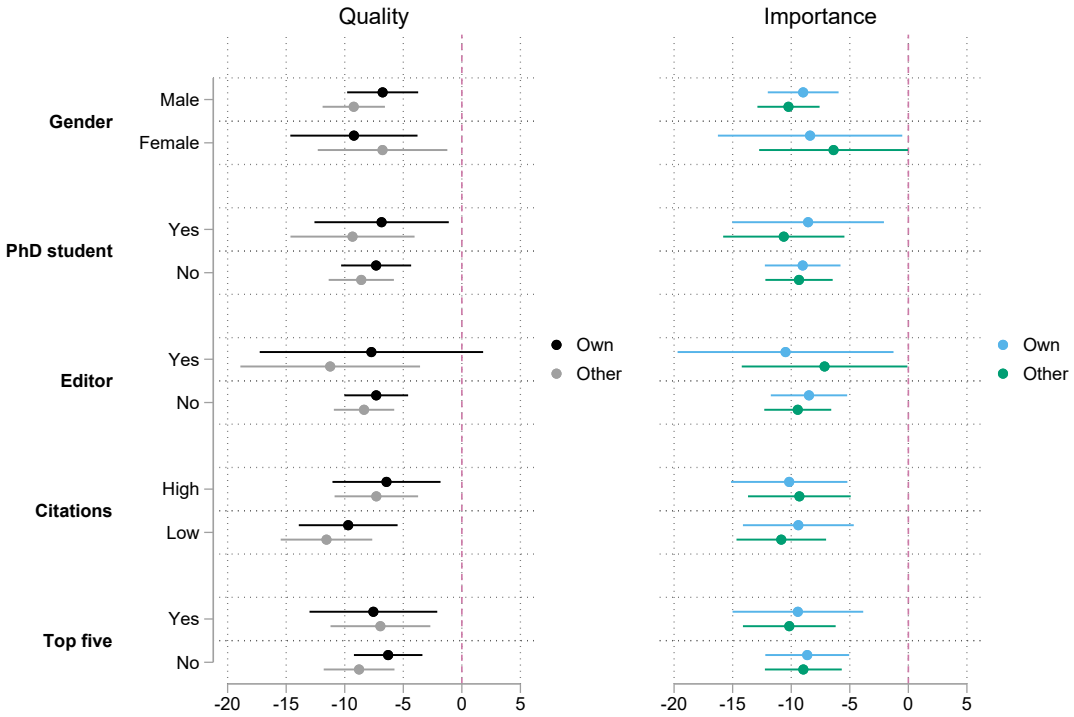
Section D provides a discussion of learning about quality from study results in the presence of expert forecasts.

Section E provides screenshots of the experimental instructions.

Section F includes the pre-analysis plans from the AsPredicted registry.

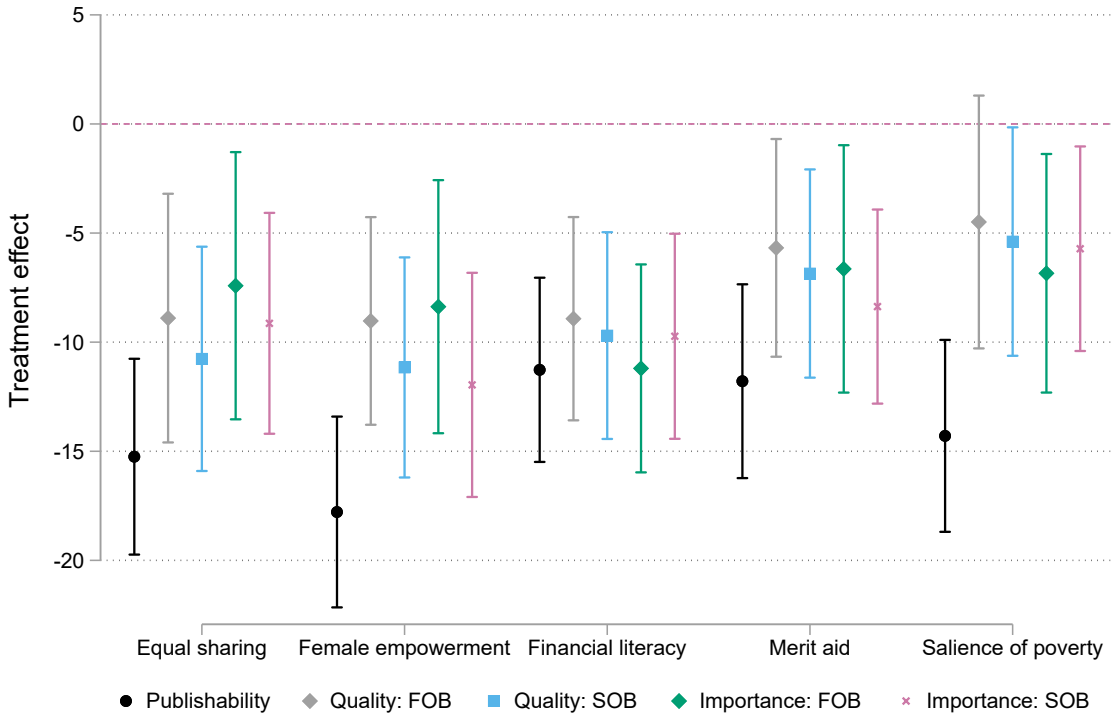
A Additional figures and tables

Figure A.1: Heterogeneity in treatment effects: First versus second-order beliefs



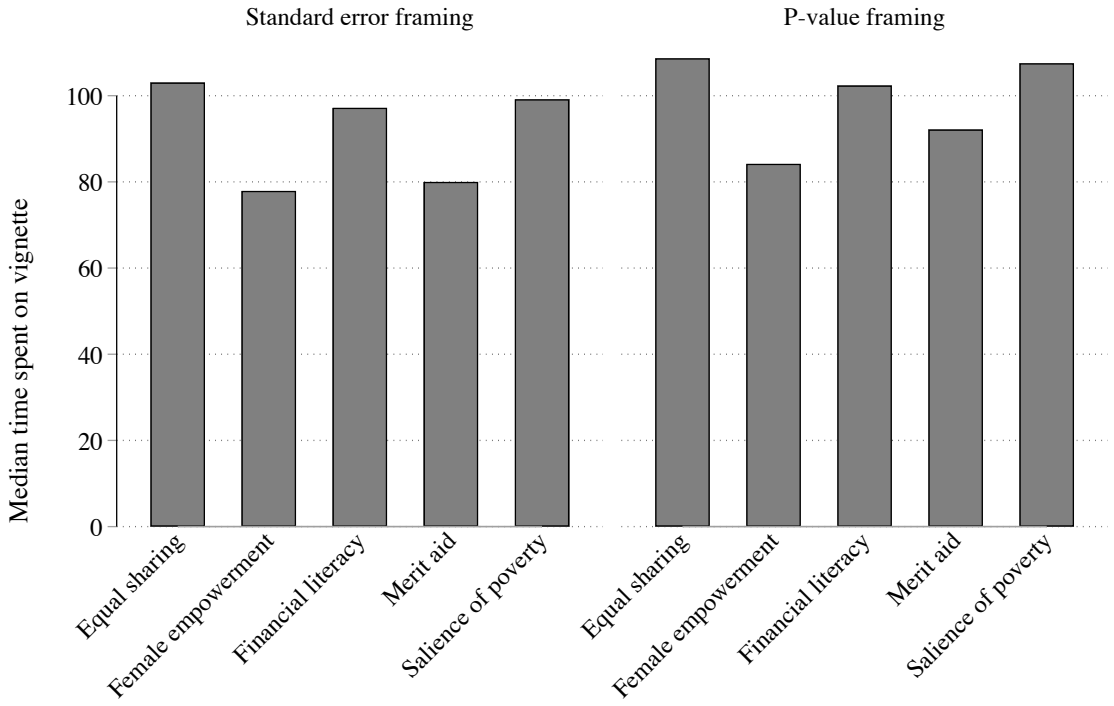
Note: This figure shows regression estimates in which first-order ("own") and second-order beliefs ("other") about the importance and quality of the study (both measured on a scale from 0 to 100) are regressed on the "null result treatment" indicator, separately for each sub-group indicated in the figure. Citations are measured using Google Scholar data as of May 2022 and "low" and 'high" refer to, respectively, 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 5" economics journals. All regressions include controls for the other cross-randomized features as well as respondent fixed effects. Standard errors are clustered at the respondent level. 95% confidence intervals are indicated in the figure.

Figure A.2: Robustness: Vignette-specific treatment effects



Note: This figure shows regression estimates of our treatment effects in which the "Null result treatment" indicator has been interacted with the five vignette-indicators. The regressions include controls for all cross-randomized features at the vignette level as well as respondent fixed effects. All outcomes are measured on a scale from 0 to 100. The publishability questions refers to beliefs about the percent chance of being published. Quality and importance of the studies are measured on a scale where 0 indicates the lowest possible quality/importance and 100 indicates the highest possible quality/importance. "FOB" (first-order beliefs) refers to personal beliefs while "SOB" (second-order beliefs) refers to beliefs about how other researchers in the field responded to the question on average. Standard errors are clustered at the respondent level. 95% confidence intervals are indicated in the figure.

Figure A.3: Median response times across vignettes



Note: This figure shows the median response time (in number of seconds) by vignette and treatment status in our main experiment.

Table A.1: Robustness: Heterogeneity by the *p*-value framing

	Dependent variable: Publishability (in %)				
	(1)	(2)	(3)	(4)	(5)
Panel A: No individual FE					
Null result	-11.754***	-11.702***	-12.009***	-11.960***	-11.161***
	(1.783)	(1.777)	(1.745)	(1.736)	(3.063)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Null result × P-value framing	-5.193**	-5.416**	-4.996**	-5.214**	-4.951**
	(2.504)	(2.515)	(2.420)	(2.430)	(2.425)
	[0.039]	[0.032]	[0.039]	[0.032]	[0.042]
Observations	1,920	1,920	1,920	1,920	1,920
Respondents	480	480	480	480	480
Respondent fixed effects	No	No	No	No	No
Vignette fixed effects		Yes		Yes	Yes
Controls: Other treatment arms			Yes	Yes	Yes
All treatment arms × Null result					Yes
Panel B: Individual FE					
Null result	-11.924***	-11.828***	-12.305***	-12.193***	-11.072***
	(1.585)	(1.533)	(1.491)	(1.432)	(2.681)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Null result × P-value framing	-4.253*	-4.473**	-3.616*	-3.854*	-3.652*
	(2.287)	(2.268)	(2.185)	(2.167)	(2.164)
	[0.064]	[0.049]	[0.099]	[0.076]	[0.092]
Observations	1,920	1,920	1,920	1,920	1,920
Respondents	480	480	480	480	480
Respondents with null variation	414	414	414	414	414
Respondent fixed effects	Yes	Yes	Yes	Yes	Yes
Vignette fixed effects		Yes		Yes	Yes
Controls: Other treatment arms			Yes	Yes	Yes
All treatment arms × Null result					Yes

Note: This table shows regression estimates of our treatment effects on publishability (beliefs in %). 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. "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. *p*-values are shown in square brackets. "Respondents with null variation" is the number of respondents who were presented with at least one vignette that included a null result and at least one vignette with a statistically significant result.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

Table A.2: Treatment effects for vignettes with research teams consisting of professors from higher-ranked universities

	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
Null result treatment	-15.735***	-0.325**	-0.404***	-0.264*	-0.369***
	(2.232)	(0.131)	(0.142)	(0.134)	(0.131)
Observations	502	260	260	242	242
Respondents	345	174	174	171	171

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 includes only observations where the vignette describes a research team consisting of professors from higher-ranked universities. "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. We include respondent and vignette fixed effects in all regressions and control for all other cross-randomized vignette features.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

Table A.3: Main results: Robustness to re-weighting

	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: Baseline with FEs					
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: Baseline with FEs, re-weighted					
Null result treatment	-14.084***	-0.298***	-0.396***	-0.365***	-0.434***
	(1.282)	(0.069)	(0.078)	(0.063)	(0.062)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250
Panel C: OLS					
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
Panel D: OLS, re-weighted					
Null result treatment	-14.567***	-0.360***	-0.420***	-0.299***	-0.362***
	(1.471)	(0.079)	(0.088)	(0.077)	(0.084)
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. 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) respondent fixed effects. All regressions in both panels include treatment indicators for the cross-randomized conditions in addition to vignette fixed effects. Panel A and C replicate the baseline results from Table 3. Panel B and D show analogous estimates when reweighing respondents to match the sampling population along several dimensions (gender, region, editorial position, top 5 referee). For details on the construction of weights, see Section B.3.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

Table A.4: Descriptive statistics: Reweighted for representativeness

	Survey sample		Sampling population
	Original mean	Reweighted mean	Mean
Demographics:			
Female*	0.220	0.229*	0.236
Years since PhD	14.805	15.664	16.091
PhD student	0.244	0.309	
Region of institution:			
Europe*	0.544	0.365*	0.362
North America*	0.406	0.524*	0.527
Australia*	0.033	0.078*	0.078
Asia*	0.017	0.033*	0.033
Academic output:			
H-index	17.216	16.960	8.831
Citations	4,348.341	4,604.263	
Number of top 5 publications	1.268	1.192	0.342
Number of top 5s refereed for	1.166	0.663	
Repeated top 5 referee*	0.305	0.158*	0.125
Research evaluation:			
Current editor*	0.072	0.035*	0.032
Current associate editor	0.127	0.106	
Ever editor	0.151	0.119	
Ever associate editor	0.193	0.162	
Professional memberships:			
NBER affiliate	0.084	0.075	
CEPR affiliate	0.171	0.106	
Academic fields:			
Labor	0.211	0.207	
Public	0.129	0.113	
Development	0.179	0.151	
Political	0.167	0.165	
Finance	0.105	0.103	
Experimental	0.062	0.066	
Behavioral	0.091	0.086	
Theory	0.067	0.062	
Macro	0.141	0.132	
Econometrics	0.141	0.139	

Note: This table displays background characteristics of the participants in the main experiment. These data are not matched with individual responses and are externally collected (i.e., not self-reported). The reweighted mean is obtained using post-stratification weights to match the sampling population on seven dimensions, indicated by an asterik. Weights are obtained using the R package `anesrake` (see Section B.3 for details). Section B.1 contains a description of each variable. Data on the total sampling population was shared by Peter Andre (see Andre and Falk, 2021). Section B.2 describes how our measures differ from those obtained from Andre and Falk (2021), in particular "Years since PhD" and "H-index".

Table A.5: Robustness: OLS using only the first observation

	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
Null result treatment	-16.242***	-0.295**	-0.396***	-0.282**	-0.291**
	(2.133)	(0.125)	(0.129)	(0.120)	(0.125)
Observations	480	230	230	250	250
Respondents	480	230	230	250	250
Controls	Yes	Yes	Yes	Yes	Yes

Note: The table shows OLS regression estimates of our treatment effects on our key outcomes of interest using only the first vignette the respondents were randomly assigned to. "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. We include treatment indicators for the other cross-randomized conditions in addition to vignette fixed effects in all regressions.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table A.6: Robustness: Treatment effects for high-powered studies and among empirical microeconomics researchers

	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: Baseline					
Null result treatment	-14.058***	-0.373***	-0.460***	-0.325***	-0.417***
	(1.090)	(0.062)	(0.062)	(0.054)	(0.056)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250
Panel B: High power					
Null result treatment	-11.486***	-0.286***	-0.338***	-0.362***	-0.370***
	(1.569)	(0.085)	(0.087)	(0.070)	(0.075)
Observations	1,156	543	543	613	613
Respondents	480	230	230	250	250
Panel C: Low power					
Null result treatment	-15.336***	-0.487***	-0.535***	-0.389***	-0.472***
	(2.270)	(0.144)	(0.143)	(0.127)	(0.133)
Observations	568	294	294	274	274
Respondents	284	147	147	137	137
Panel D: Empirical micro sample					
Null result treatment	-13.417***	-0.246***	-0.308***	-0.378***	-0.425***
	(1.897)	(0.089)	(0.097)	(0.092)	(0.092)
Observations	837	420	420	417	417
Respondents	348	176	176	172	172

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. "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. Panel A reports the baseline treatment effects from our main specification. Panel B focuses on the three vignettes that have a power of at least 80% to detect a treatment effect of 20% of a standard deviation (see Table 2), while Panel C uses only observations from vignettes that had a comparatively lower statistical power. Panel D restricts the sample to high-powered vignettes and researchers that are more likely to have experience with empirical microeconomic research by excluding researchers in the field of macro, finance, international economics, and theory. We include respondent and vignette fixed effects in all regressions and control for all other cross-randomized features.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

Table A.7: Heterogeneity by vignette characteristics: Separate interaction terms with multiple hypothesis adjustment

	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: Expert forecast					
Null result	-11.239***	-0.281**	-0.506***	-0.351***	-0.432***
	(1.913)	(0.113)	(0.112)	(0.094)	(0.085)
	[0.001]***	[0.022]**	[0.001]***	[0.001]***	[0.001]***
Null result x Low expert forecast	-2.002	-0.168	0.128	0.032	0.065
	(2.478)	(0.161)	(0.161)	(0.120)	(0.116)
	[0.993]	[0.939]	[0.993]	[1.000]	[1.000]
Null result x High expert forecast	-6.383**	-0.104	0.009	0.048	-0.020
	(2.646)	(0.166)	(0.154)	(0.124)	(0.126)
	[0.029]**	[1.000]	[1.000]	[1.000]	[1.000]
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	-14.571***	-0.366***	-0.446***	-0.343***	-0.418***
	(1.465)	(0.093)	(0.086)	(0.072)	(0.075)
	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***
Null result x Field journal	1.025	-0.014	-0.027	0.036	0.003
	(1.965)	(0.129)	(0.122)	(0.101)	(0.103)
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
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	-14.945***	-0.291***	-0.358***	-0.300***	-0.362***
	(1.491)	(0.085)	(0.082)	(0.081)	(0.081)
	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***
Null result x PhD student	1.745	-0.166	-0.206*	-0.047	-0.104
	(2.049)	(0.117)	(0.107)	(0.102)	(0.097)
	[0.991]	[0.670]	[0.176]	[1.000]	[0.937]
PhD student	-4.543***	-0.025	-0.042	0.066	0.019
	(1.403)	(0.091)	(0.081)	(0.071)	(0.069)
Panel D: Lower-ranked university					
Null result	-14.320***	-0.381***	-0.474***	-0.317***	-0.408***
	(1.480)	(0.094)	(0.093)	(0.073)	(0.076)
	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***
Null result x Lower-ranked university	0.518	0.017	0.030	-0.014	-0.017
	(1.985)	(0.121)	(0.124)	(0.108)	(0.105)
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
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	-11.960***	-0.243**	-0.302***	-0.366***	-0.405***
	(1.736)	(0.095)	(0.101)	(0.081)	(0.095)
	[0.001]***	[0.017]**	[0.004]***	[0.001]***	[0.001]***
Null result x P-value framing	-5.214**	-0.296**	-0.286**	0.140	0.088
	(2.430)	(0.136)	(0.141)	(0.124)	(0.135)
	[0.080]*	[0.068]*	[0.113]	[0.917]	[1.000]
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 presents estimates that are exactly analogous to Table 4. Im addition, Romano and Wolf (2005) p-values adjusted for multiple hypothesis testing are shown in square brackets with corresponding significance stars for the key coefficients of interest (Clarke, 2021), excluding interactants.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

Table A.8: Treatment heterogeneity by vignette characteristics: Fully interacted model with multiple hypothesis adjustment

	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
Main treatment:					
Null result	-11.072***	-0.029	-0.219	-0.330**	-0.390***
	(2.681)	(0.151)	(0.160)	(0.132)	(0.135)
	[0.001]***	[1.000]	[0.755]	[0.026]**	[0.006]***
Interaction effects:					
Null result x Low expert forecast	-1.862	-0.169	0.130	0.030	0.058
	(2.470)	(0.162)	(0.159)	(0.120)	(0.117)
	[0.997]	[0.958]	[0.994]	[1.000]	[1.000]
Null result x High expert forecast	-6.251**	-0.083	0.033	0.048	-0.025
	(2.632)	(0.165)	(0.152)	(0.124)	(0.127)
	[0.042]**	[1.000]	[1.000]	[1.000]	[1.000]
Null result x Field journal	0.871	0.003	-0.027	0.038	0.006
	(1.966)	(0.131)	(0.121)	(0.101)	(0.103)
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Null result x PhD student	1.707	-0.165	-0.196*	-0.047	-0.101
	(2.054)	(0.121)	(0.108)	(0.102)	(0.098)
	[0.994]	[0.755]	[0.305]	[1.000]	[0.960]
Null result x Low-ranked university	0.408	0.021	0.028	-0.011	-0.018
	(1.965)	(0.121)	(0.124)	(0.108)	(0.106)
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Null result x P-value framing	-3.652*	-0.344***	-0.362***	-0.021	0.049
	(2.164)	(0.122)	(0.120)	(0.109)	(0.112)
	[0.407]	[0.006]***	[0.004]***	[1.000]	[1.000]
Interactants:					
Low expert forecast	-0.876	0.200*	-0.007	-0.076	-0.041
	(1.666)	(0.108)	(0.107)	(0.092)	(0.081)
High expert forecast	1.977	0.108	0.110	-0.049	-0.019
	(1.789)	(0.113)	(0.096)	(0.085)	(0.088)
Field journal	12.204***	0.120	0.097	0.108	0.100
	(1.396)	(0.097)	(0.090)	(0.073)	(0.069)
PhD student	-4.600***	-0.036	-0.056	0.068	0.016
	(1.407)	(0.094)	(0.082)	(0.071)	(0.069)
Low-ranked university	-3.986***	-0.105	-0.235***	0.006	-0.046
	(1.363)	(0.081)	(0.076)	(0.077)	(0.073)
N	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 from a specification that includes the full interactions between the null treatment indicator and indicators for all cross-randomized features. The data set is at the vignette-respondent level and contains four observations for each respondent. We include individual and vignette fixed effects in all regressions. "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 predicted, respectively, 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. Romano and Wolf (2005) p-values adjusted for multiple hypothesis testing are shown in square brackets with corresponding significance stars for the key coefficients of interest (Clarke, 2021), excluding interactants.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

Table A.9: Robustness: Familiarity with the research study's research field

	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					
Null result treatment	-13.202***	-0.389***	-0.454***	-0.329***	-0.430***
	(1.253)	(0.072)	(0.074)	(0.060)	(0.068)
Matching field	0.883	0.041	0.053	0.185*	0.025
	(1.922)	(0.118)	(0.114)	(0.105)	(0.102)
Null result treatment x Matching field	-2.764	0.042	-0.019	0.018	0.048
	(2.192)	(0.130)	(0.125)	(0.121)	(0.114)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250
Panel B: Non-matching fields					
Null result treatment	-13.206***	-0.366***	-0.493***	-0.332***	-0.510***
	(1.457)	(0.080)	(0.082)	(0.070)	(0.081)
Observations	988	468	468	520	520
Respondents	247	117	117	130	130
Panel C: Matching fields					
Null result treatment	-18.067***	-0.414***	-0.493***	-0.395***	-0.416***
	(2.117)	(0.119)	(0.114)	(0.126)	(0.112)
Observations	566	307	307	259	259
Respondents	183	98	98	85	85

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. "Matching field" is a binary indicator taking value one if the vignette presented to the respondent belongs to a research field that overlaps with the respondent's fields of specialization, and zero otherwise. Panel A includes all observations, while Panel B includes respondents that are specialized in research fields that are unrelated to the studies they encountered in the survey. Panel C uses only observations where respondents are familiar with the field of specialization that the study presented in the hypothetical vignette belongs to. "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. We include vignette fixed effects and respondent fixed effects in all regressions and control for all other cross-randomized vignette features.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

Table A.10: Descriptive statistics on time use

	count	p25	p50	p75	min	max	mean	sd
Duration in seconds	480	339.0	450.5	694.0	78.0	93322.0	1306.8	7218.4
Duration (winsorized)	480	339.0	450.5	694.0	78.0	1333.0	552.3	300.7
Page time (in seconds)	1920	65.5	93.6	140.1	7.7	92747.4	177.3	2120.5
Page time (winsorized)	1920	65.5	93.6	140.1	7.7	301.5	113.6	69.1

Note: This table displays summary statistics on time use (in seconds) in the main experiment. "Duration" refers to the total time spent on the whole survey (in seconds), including the introductory screen and the feedback screen. "Page time" refers to the total time spent on each vignette screen (in seconds). We also include winsorized versions of these variables where all values above the 95th percentile are set to the 95th percentile.

Table A.11: Time spent on vignettes

	Dependent variable: Page time (in seconds)			
	Non-winsorized		Winsorized	
	(1)	(2)	(3)	(4)
Equal sharing (80 extra words)	-210.678	-216.822	24.405***	23.832***
	(245.382)	(250.097)	(4.122)	(4.088)
Financial literacy (35 extra words)	-221.535	-228.240	16.881***	16.267***
	(246.060)	(251.303)	(4.259)	(4.275)
Salience of poverty (22 extra words)	-216.988	-223.385	22.347***	21.921***
	(246.354)	(252.054)	(4.128)	(4.110)
Merit aid (1 extra word)	-225.516	-231.957	5.393	4.688
	(246.359)	(251.248)	(4.160)	(4.077)
Null result treatment		98.624		8.330***
		(92.092)		(2.786)
Low expert forecast (43 extra words)		157.127		8.209**
		(142.739)		(3.919)
High expert forecast (43 extra words)		20.294*		13.589***
		(12.139)		(3.818)
Field journal		-91.886		-0.295
		(94.881)		(3.198)
PhD student		-73.173		2.023
		(81.304)		(3.056)
Low-ranked university		-96.333		-0.312
		(95.015)		(3.199)
Observations	1,920	1,920	1,920	1,920
Respondents	480	480	480	480
Dep. var. mean	177.280	177.280	114.242	114.242

Note: The table shows OLS regression estimates of our treatment effects on the time that respondents spent on a vignette in the main experiment. The data is at the respondent-vignette level. The dependent variables in columns 3 and 4 are winsorized at the 5th and 95th percentile. "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. We include treatment indicators for the other cross-randomized conditions in addition to vignette fixed effects in all regressions. The omitted category for the vignettes is the "Female empowerment" vignette with a 116 word description of the study. For each vignette, the word count difference relative to the female empowerment vignette is indicated.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

Table A.12: Treatment effects for respondents who spent at least k minutes on the survey

	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: 4+ minutes					
Null result treatment	-14.685***	-0.420***	-0.498***	-0.344***	-0.419***
	(1.126)	(0.063)	(0.062)	(0.056)	(0.057)
Observations	1,788	864	864	924	924
Respondents	447	216	216	231	231
Panel B: 6+ minutes					
Null result treatment	-15.518***	-0.407***	-0.502***	-0.356***	-0.460***
	(1.320)	(0.071)	(0.072)	(0.064)	(0.066)
Observations	1,360	656	656	704	704
Respondents	340	164	164	176	176
Panel C: 8+ minutes					
Null result treatment	-16.654***	-0.482***	-0.556***	-0.310***	-0.413***
	(1.575)	(0.086)	(0.085)	(0.075)	(0.079)
Observations	884	412	412	472	472
Respondents	221	103	103	118	118
Panel D: 10+ minutes					
Null result treatment	-16.986***	-0.581***	-0.569***	-0.285***	-0.449***
	(1.853)	(0.109)	(0.106)	(0.085)	(0.089)
Observations	640	284	284	356	356
Respondents	160	71	71	89	89

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 each panel includes only observations where the respondent spent at least k minutes on the overall survey. The panel header indicates the value of k. "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. We include vignette fixed effects and respondent fixed effects in all regressions and control for all other cross-randomized vignette features.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered at the respondent level in parentheses.

B Expert characteristics

B.1 Survey sample

Below, we describe the observed expert characteristics in our expert sample and how we obtained this information. If not otherwise indicated, the data was obtained from researchers' publicly available CV.

- *Female*: Binary indicator for female experts.
- *Years since PhD*: This variable is the number of calendar years between 2022 and the year the experts obtained their PhD.
- *PhD student*: Binary indicator for PhD students.
- *Region of institution*: Regional indicators taking the values "Asia", "Australia", "Europe", and "North America" depending on where the institution the researcher works for is based.
- *H-index*: The researcher's H-index as taken from their Google Scholar profile (as of May 2022).
- *Citations*: The researcher's total citation count as taken from their Google Scholar profile (as of May 2022).
- *Number of top 5 publications*: This variable is the number of publications in five highly cited general-interest economics journals (the American Economic Review, the Quarterly Journal of Economics, the Journal of Political Economy, Econometrica, and the Review of Economic Studies).
- *Number of top 5s refereed for*: This variables is the number of highly cited general-interest economics journals (the American Economic Review, the Quarterly Journal of Economics, the Journal of Political Economy, Econometrica, and the Review of Economic Studies) that the researcher has refereed for in the past.
- *Repeated top 5 referee*: Binary indicator for whether the researcher has refereed for at least two of the five highly cited general-interest economics journals (the American Economic Review, the Quarterly Journal of Economics, the Journal of Political Economy, Econometrica, and the Review of Economic Studies).

- *Current editor*: Binary indicator for current editors.
- *Current associate editor*: Binary indicator for current associate editors.
- *Ever editor*: Binary indicator for having ever been an editor of a scientific journal.
- *Ever associate editor*: Binary indicator for having ever been an associate editor of a scientific journal.
- *NBER affiliate*: Binary indicator for holding an NBER affiliation.
- *CEPR affiliate*: Binary indicator for holding a CEPR affiliation.
- *Academic field*: A series of binary indicators for different research fields in economics. The information is obtained from the expert's Google Scholar profile. Specifically, we obtain all research fields that the researcher listed on his Google Scholar profile. We then construct indicators depending on whether a field (e.g. "Labor") was included in the list of research fields. Therefore, experts can belong to several academic fields within economics.

B.2 Sampling population

We obtained summary statistics for the full population of our sampling frame from Andre and Falk (2021). Specifically, Peter Andre derived the mean and median for a set of expert characteristics based on the population of active research economists affiliated with the top 200 institutions according to RePEc (as of March 2022). For a definition of active research economists, see Andre and Falk (2021). Below, we describe how the variables are constructed.

- *Female*: Binary indicator for female experts, derived from their first and last name using the Gender API algorithm.
- *Year of first publication*: This is the number of years since the first publication, as recorded in the Scopus database.
- *Region of institution*: Regional indicators are based on the country of the institution a researcher is affiliated with. The information is obtained from the Scopus database.

- *H-index*: The Scopus h-index, which is derived from Scopus data on the citations of all publications of an author as of December 2019.
- *Number of top 5 publications*: This variable is the number of publications in five highly cited general-interest economics journals (the American Economic Review, the Quarterly Journal of Economics, the Journal of Political Economy, Econometrica, and the Review of Economic Studies).
- *Repeated top 5 referee*: Binary indicator for whether the researcher has repeatedly refereed for the five highly cited general-interest economics journals (the American Economic Review, the Quarterly Journal of Economics, the Journal of Political Economy, Econometrica, and the Review of Economic Studies) in the years from 2015-2020. This variable is derived from the list of referees that have refereed for these five journals. These lists are published on an annual basis. For more details, see Appendix A.4 of Andre and Falk (2021).
- *Current editor*: Binary indicator for having been an editor at one of the top 100 journals in economics at some point in the years from 2015-2020. For more details, see Appendix A.4 of Andre and Falk (2021).

Measurement differences: Note that we use "Year of first publication" as a benchmark for our measure of "Years since PhD" in Table 1. Moreover, note that we obtained the H-index from Google Scholar, while Andre and Falk (2021) use the Scopus H-index. Google Scholar H-indices are typically higher than those reported by Scopus.

B.3 Weights

We use the R package `anesrake` to obtain respondent-level weights for our sample (Pasek et al., 2014). We construct weights such that the reweighted sample matches the marginal distribution of four characteristics in the study population of research economists at the top 200 institutions according to RePEc (as of March 2022):

- Gender (two groups)
- Region of institution (four groups)
- Repeated top 5 referee (two groups)

- Current editor (two groups)

The frequencies of different groups in the total study population were shared by Peter Andre (see Andre and Falk, 2021). The distribution of weights does not exhibit extreme outliers. 94.5% of weights are between 0.3 and 2. Moreover, weights range from a minimum of 0.18 to a maximum weight of 2.89.

C Numerical features used in the vignettes

One of our design goals was to remain close to the parameters of the original studies on which we based our vignettes. At the same time, we wanted to vary key features (e.g. the magnitude of the main effect) in a way such that the numerical values of the features that we ultimately report in each vignette are internally consistent irrespective of the condition to which respondents are assigned. This section describes how we proceed to achieve this.

For each vignette, we first discuss the features that are *constant* across respondents. Below, we provide details on how we determined the numerical values of these features:

- Standard error: We conducted a simulation exercise to obtain an estimate of the standard error that one would obtain based on the number of observations, the control group mean, the assignment to treatment and control groups, and the main effect from the original studies on which we based our vignettes. This ensures that our reported standard errors and p-values are internally consistent with the description of the sample and the empirical strategy.
- Number of experts: This is an integer drawn uniformly from the interval $[20, 35]$.
- Standard deviation of the expert prior: We multiplied the standard error (see above) with a number drawn uniformly from the interval $[1, 2]$. This ensures that a Bayesian with the experts' prior should put a weight of at least 0.5 on the study's findings when updating his belief about the underlying "true" effect. This implies that the study findings are informative relative to the experts' prior, irrespective of whether the main effect is statistically significant or not.

For each vignette, we determined the numerical values of the features that we *vary* across respondents as follows:

- Main effect (statistically significant): We draw a hypothetical t-statistic from a uniform distribution $t \sim \text{Unif}([2, 3])$. The main effect is then set to the product of t and the standard error (see above).
- Main effect (statistically non-significant): We draw a hypothetical t-statistic from a uniform distribution $t \sim \text{Unif}([0.1, 0.5])$. The main effect is then set to the

product of t and the standard error (see above). Note that we deliberately opted for an approach that produces standard errors and p-values associated with the main effect that allow respondents to quickly identify whether a vignette includes a statistically significant main effect. We chose this data generating process over one where t-statistics are drawn uniformly around a cutoff such as 1.96 for power reasons.

- p-values (high and low): The p-values are obtained from the hypothetical t-statistic used to generate the statistically (non-)significant main effect.
- Expert prior (high mean): This number is equal to $\mu_{\text{high}} + 0.25x$. Here, μ_{high} is the statistically significant main effect and $x \sim N(0, S)$ where S is the standard error (see above).
- Expert prior (low mean): This number is equal to the high expert prior minus the absolute difference between the statistically significant and the statistically non-significant main effect. This ensures that the absolute difference between the high and low expert mean is equal to the absolute difference between the statistically significant and non-significant main effects.

D Updating about quality

This section provides a discussion of updating about the quality of research studies with null results in the presence of expert forecasts. We start with a general discussion of the empirical evidence from our experiment in Section D.1. This discussion draws on model predictions that we formalize in Section D.2.

D.1 Discussion

One potential explanation for the null result penalty is that researchers might rationally draw negative inference about the quality of studies with null results, thus lowering their perceived publication chances. This mechanism requires several ingredients. First, the quality of a research study cannot be perfectly observed. Second, high-quality studies are more likely to uncover whether there is a true causal relationship between two variables. Note that this entails that high-quality studies are more likely to yield null results if there is no true causal relationship, and that they are at the same time more likely to yield a statistically significant result if there is a causal relationship.[1] Third, prior to observing the study results, researchers are sufficiently confident that there is a causal relationship. In this case, observing a null result will cause them to infer that the research study is more likely to be of low quality—as a high-quality study would have been more likely to yield a statistically significant result consistent with their prior.

Note that the directional predictions reverse if researchers are sufficiently certain that there is no causal relationship. In this case, observing that a study yielded a null result is a positive signal of its quality. This suggests an empirical test based on exogenously varying the prior belief in a causal relationship and studying how this affects the inferences people draw about the quality of studies with null results compared to studies with statistically significant results.

Empirically, we find that respondents that do not receive an expert forecast update negatively about the quality of research studies with null results (column 2 of Table 3).

[1] For example, high-quality studies might be more likely to yield null results if there is no true causal relationship because they are less likely to engage in *p*-hacking, or because their experimental design is less likely to be confounded. Note that these are potentially unobservable features of quality that are not captured by, for instance, observable characteristics such as the standard error and the sample size that influence the *nominal* Type I and Type II error rates.

This would be consistent with the mechanism outlined above if respondents were sufficiently certain that there is a true causal effect. Next, we exploit the exogenous variation in prior beliefs induced by the high vs low expert forecasts. Note that the expert forecasts affect respondents' assessment of the publishability of null results (column 1 of Table 4, Panel A), suggesting that participants—at least to a certain degree—internalize the information about the expert predictions. Column 2 in Panel A of Table 4 shows that respondents update negatively about the quality of research studies independent of whether they were informed that experts predict a large or a small effect. If anything, participants update somewhat more negatively about the quality of research studies with unsurprising nulls (relative to the experts' prior) compared to surprising nulls. While these effects are somewhat noisily estimated, they are directionally at odds with rational inference about the quality of a study.

D.2 Formalization

We present a formalization of the above mechanism that causes observers to negatively update about the unobserved quality of a research study after observing that the study yielded a null result.

Setup A research study tests whether there is a causal relationship between two variables of interest. Let $\omega \in \{0,1\}$ denote whether there is a causal relationship ($\omega = 1$) or not ($\omega = 0$). The research study provides a binary signal $s \in \{0,1\}$ which we interpret as the result from a statistical test for whether the causal parameter of interest is statistically significantly different from zero. Research studies differ in their characteristics $(\theta_{\text{obs}}, \theta)$, where θ_{obs} denotes characteristics that are perfectly observed such as the sample size; and θ denotes unobserved quality characteristics, such as the clarity of the experimental instructions or the soundness of the statistical analysis.

We will examine how a Bayesian researcher will rationally revise his belief about the unobserved quality θ of a study depending on whether it yielded a statistically significant main effect or not. The thought experiment we are interested in is one where there are two studies, A and B, with the same observable characteristics θ_{obs}, but study A yielded a null result while study B yielded a statistically significant result. This thought experiment mirrors our experimental design, where we fix observable characteristics such as the sample size, and then exogenously vary the statistical significance of the

main finding. How will a Bayesian observer revise his beliefs about the quality of study A compared to study B? In the following, we suppress θ_{obs} from the notation to simplify the exposition as we are only interested in the observers' inferences about the unobserved study features.

Heterogeneity in unobserved quality Suppose that studies either have a high quality ($\theta = H$) or a low quality ($\theta = L$) and the quality of a study is drawn independently from ω. The quality of a study determines the probability of correctly diagnosing ω. We denote these probabilities by

$$\pi^1(\theta) = P(s = 1 \mid \omega = 1, \theta) \tag{3}$$

$$\pi^0(\theta) = P(s = 0 \mid \omega = 0, \theta) \tag{4}$$

Note that $1 - \pi^1(\theta)$ and $1 - \pi^0(\theta)$ will be different from the *nominal* Type II and Type I error rates of a study with quality θ, as the latter are statistical concepts that, for example, do not account for the potential confoundedness of experimental designs or fraudulent research practices that might be more prevalent among low-quality studies. We model the quality of a study in our setting by assuming that $\pi_H^j \geq \pi_L^j$ for $j \in \{0, 1\}$ where $\pi_\theta^j \equiv \pi^j(\theta)$, which means that high-quality studies are more likely to yield the "correct" result in both states of the world.

Inference about study quality Suppose that an outside researcher observes the results s of a research study. The prior belief of the researcher is that there is a chance of $\rho \in (0, 1)$ that a study is of high quality. At the same time, the researcher starts from a prior $p = P(\omega = 1) \in (0, 1)$ that there is a causal relationship. How will the researcher revise his beliefs about the quality of the study when the study yielded a null result ($s = 0$) or a statistically significant result ($s = 1$)?

The proposition below establishes that a null result will yield to more pessimism about the quality of a study if the researcher's prior belief in a causal relationship p is sufficiently high. Conversely, a statistically significant finding will make the researcher more confident that the study is of high-quality if he believes in a causal relationship.

Proposition 1. Let $\hat{\rho}(s)$ denote the posterior probability of a researcher with prior ρ

that a study is of high-quality after observing the results $s \in \{0,1\}$ of a study. Then

$$\hat{\rho}(s=0) \leq \rho \leq \hat{\rho}(s=1) \tag{5}$$

if the following inequality holds

$$\frac{(\pi_H^0 - \pi_L^0)}{(\pi_H^0 - \pi_L^0) + (\pi_H^1 - \pi_L^1)} \leq p \tag{6}$$

where p is the researcher's prior belief that there is a causal effect. Importantly, the strength of the negative updating about the quality of the study after observing a null depends on his prior belief:

$$\frac{\partial \hat{\rho}(s=0)}{\partial p} < 0 < \frac{\partial \hat{\rho}(s=1)}{\partial p} \tag{7}$$

Proof. We first characterize $\hat{\rho}(s=0)$, which is determined by Bayes' rule:

$$\hat{\rho}(0) = \frac{P(H)P(s=0|H)}{P(L)P(s=0|L) + P(H)P(s=0|H)} \tag{8}$$

$$= \frac{P(H)\left(P(s=0|\omega=0,H)P(\omega=0) + P(s=0|\omega=1,H)P(\omega=1)\right)}{\left(\begin{array}{r} P(H)\left(P(s=0|\omega=0,H)P(\omega=0) + P(s=0|\omega=1,H)P(\omega=1)\right) + \\ P(L)\left(P(s=0|\omega=0,L)P(\omega=0) + P(s=0|\omega=1,L)P(\omega=1)\right) \end{array}\right)} \tag{9}$$

$$= \frac{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right)}{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) + (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right)} \tag{10}$$

It then follows that $\hat{\rho}(0) \leq \rho$ if and only if

$$\frac{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right)}{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) + (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right)} \leq \rho \tag{11}$$

$$\iff (1-\rho)\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) \leq (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right) \tag{12}$$

$$\iff \pi_H^0(1-p) + (1-\pi_H^1)p \leq \pi_L^0(1-p) + (1-\pi_L^1)p \tag{13}$$

$$\iff (\pi_H^0 - \pi_L^0)(1-p) \leq \left(\pi_H^1 - \pi_L^1\right)p \tag{14}$$

As we assumed that $\pi_L^j \leq \pi_H^j$ for both $j = 0,1$, it follows that the above inequality holds

if and only if

$$\frac{(\pi_H^0 - \pi_L^0)}{(\pi_H^0 - \pi_L^0) + (\pi_H^1 - \pi_L^1)} \leq p, \tag{15}$$

which concludes the proof of the first part of the proposition.

Next, we characterize $\hat{\rho}(s=1)$, which is again determined by Bayes' rule, using analogous steps:

$$\hat{\rho}(1) = \frac{P(H)P(s=1|H)}{P(L)P(s=1|L) + P(H)P(s=1|H)} \tag{16}$$

$$= \frac{P(H)\left(P(s=1|\omega=0,H)P(\omega=0) + P(s=1|\omega=1,H)P(\omega=1)\right)}{\left(\begin{array}{c} P(H)\left(P(s=1|\omega=0,H)P(\omega=0) + P(s=1|\omega=1,H)P(\omega=1)\right) + \\ P(L)\left(P(s=1|\omega=0,L)P(\omega=0) + P(s=1|\omega=1,L)P(\omega=1)\right) \end{array}\right)} \tag{17}$$

$$= \frac{\rho\left((1-\pi_H^0)(1-p) + \pi_H^1 p\right)}{\rho\left((1-\pi_H^0)(1-p) + \pi_H^1 p\right) + (1-\rho)\left((1-\pi_L^0)(1-p) + \pi_L^1 p\right)} \tag{18}$$

It then follows that $\rho \leq \hat{\rho}(1)$ if and only if

$$\frac{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right)}{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) + (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right)} \geq \rho \tag{19}$$

$$\iff (1-\rho)\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) \geq (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right) \tag{20}$$

$$\iff \pi_H^0(1-p) + (1-\pi_H^1)p \geq \pi_L^0(1-p) + (1-\pi_L^1)p \tag{21}$$

$$\iff (\pi_H^0 - \pi_L^0)(1-p) \geq \left(\pi_H^1 - \pi_L^1\right)p \tag{22}$$

The above condition is equivalent to equation (14). The same argument as above then shows that $\rho \leq \hat{\rho}(1)$ if and only if equation (15) holds.

We can now examine how the strength of the updating is related to the prior belief by computing the derivative of $\hat{\rho}$ with respect to p. Let

$$f \equiv \rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) \tag{23}$$

$$g \equiv \rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) + (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right) \tag{24}$$

and let $sgn(x)$ denote the signum of x. Then

$$sgn\left(\frac{\partial\hat{\rho}(s=0)}{\partial p}\right) = sgn\left(\frac{\partial(f/g)}{\partial p}\right) = sgn\left(\frac{\partial f}{\partial p}g - f\frac{\partial g}{\partial p}\right) \tag{25}$$

$$=sgn(\rho(1-\pi_H^0-\pi_H^1)(1-\rho)(\pi_L^0(1-p)+(1-\pi_L^1)p)- \tag{26}$$

$$(1-\rho)(1-\pi_L^0-\pi_L^1)\rho(\pi_H^0(1-p)+p(1-\pi_H^1))) \tag{27}$$

$$= sgn\left((1-\pi_H^0-\pi_H^1)\pi_L^0-(1-\pi_L^0-\pi_L^1)\pi_H^0\right) = -1 \tag{28}$$

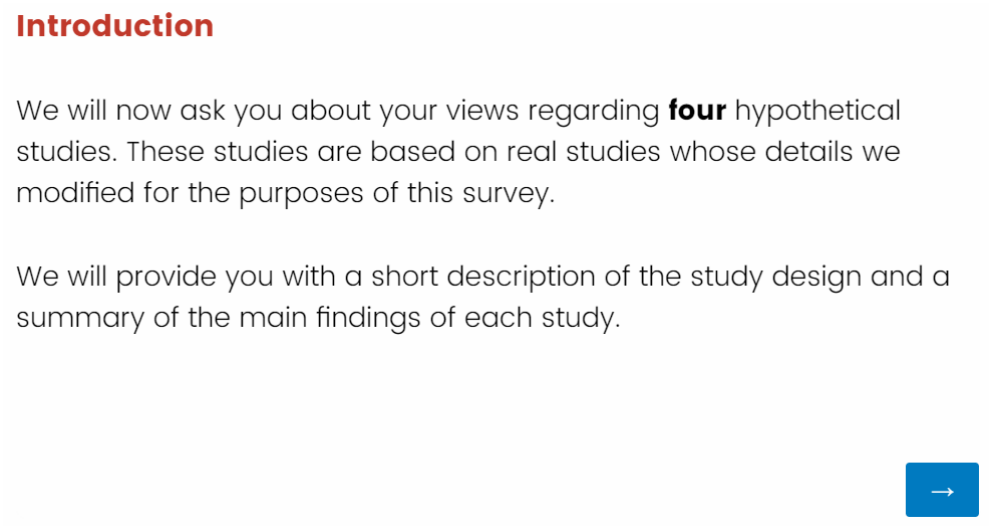
where the last equality follows from $\pi_L^0 \leq \pi_H^0$ and $\pi_H^0+\pi_H^1 > \pi_L^0+\pi_L^1$. An analogous calculation establishes that $sgn\left(\frac{\partial\hat{\rho}(s=1)}{\partial p}\right) = 1$, which concludes the proof. □

E Screenshots

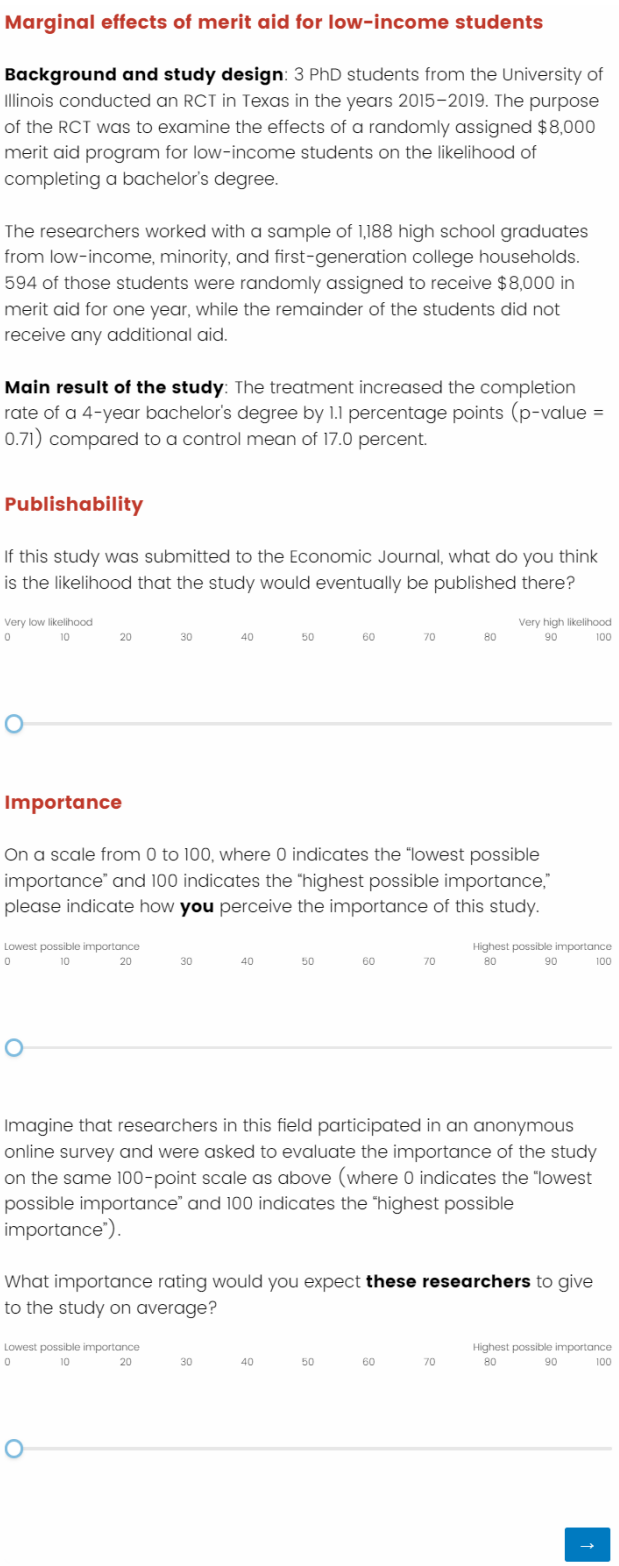
E.1 Main experiment

Respondents in the main experiment were randomly shown four of the five vignettes (in random order). We experimentally vary six features across vignettes (the communication of scientific findings, the statistical significance of the results, whether it includes a high or low or no expert forecast, seniority of the research team, university of the research team, and whether the journal is a general interest or field journal). Five features vary at the respondent-by-vignette level, and one feature varies at the respondent level (whether the main finding includes the p-value or the standard error associated with the main effect). The conditions shown in the following screenshots include a random draw of these six cross-randomized conditions.

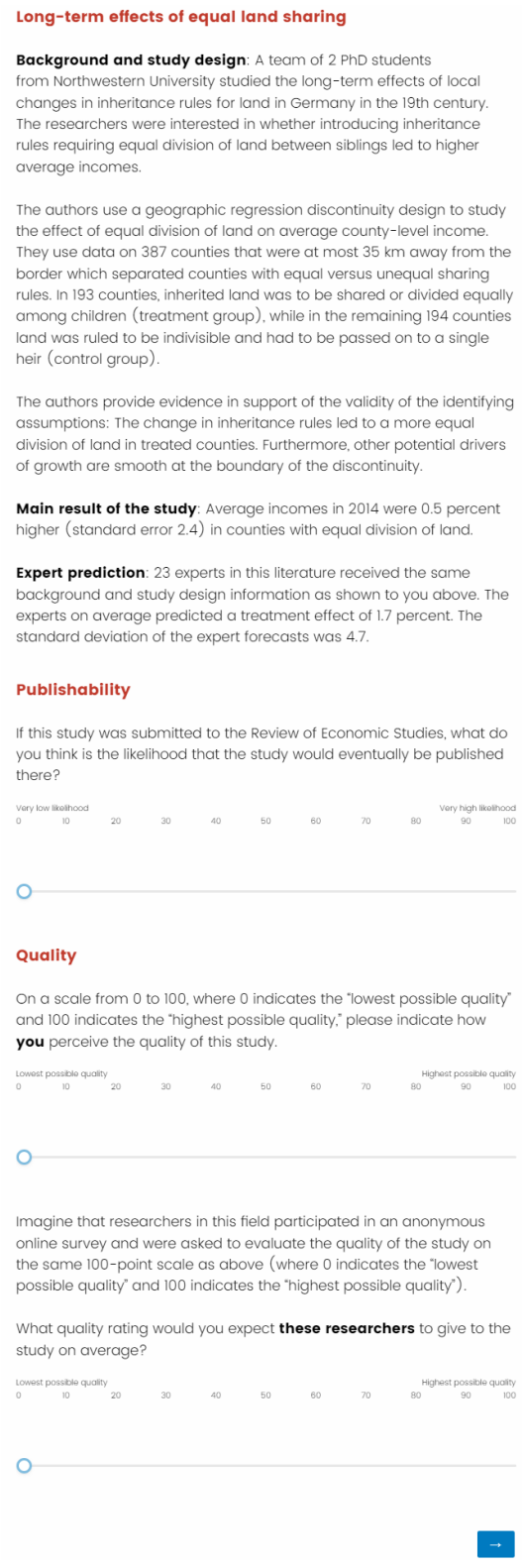
E.1.1 Pre-treatment information



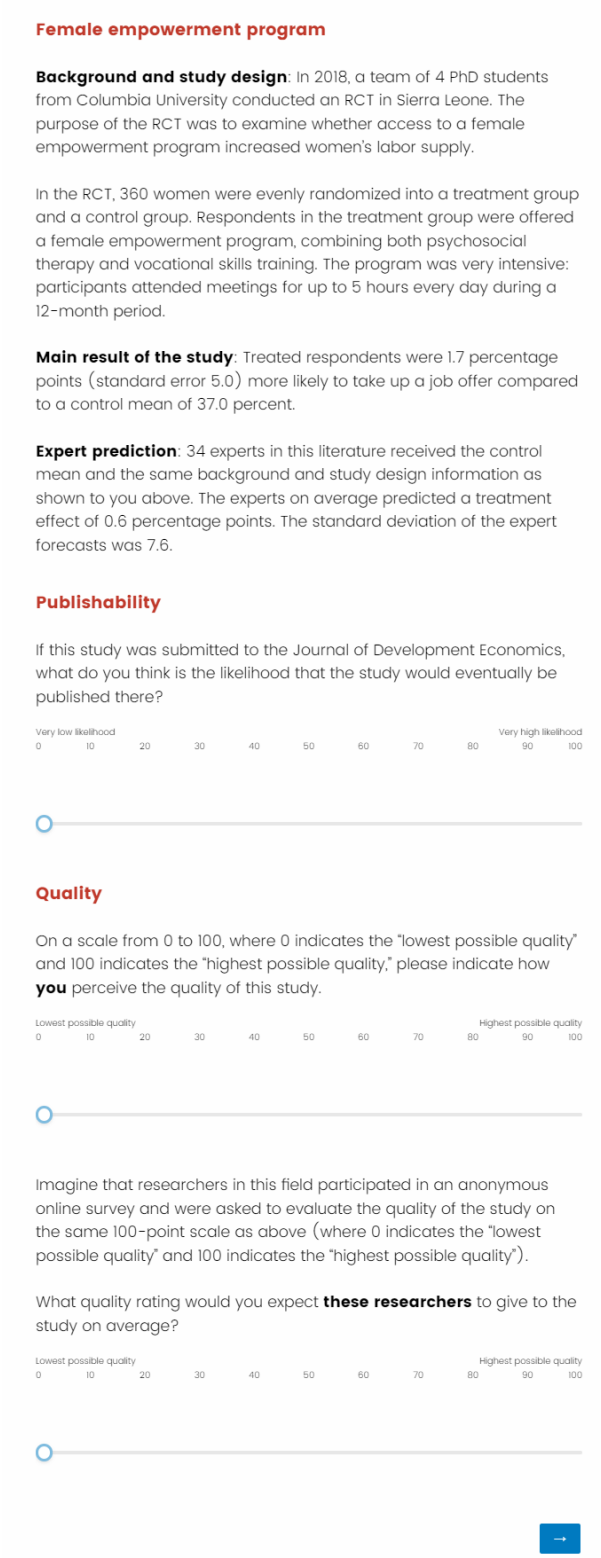
E.1.2 Marginal effects of merit aid for low-income students



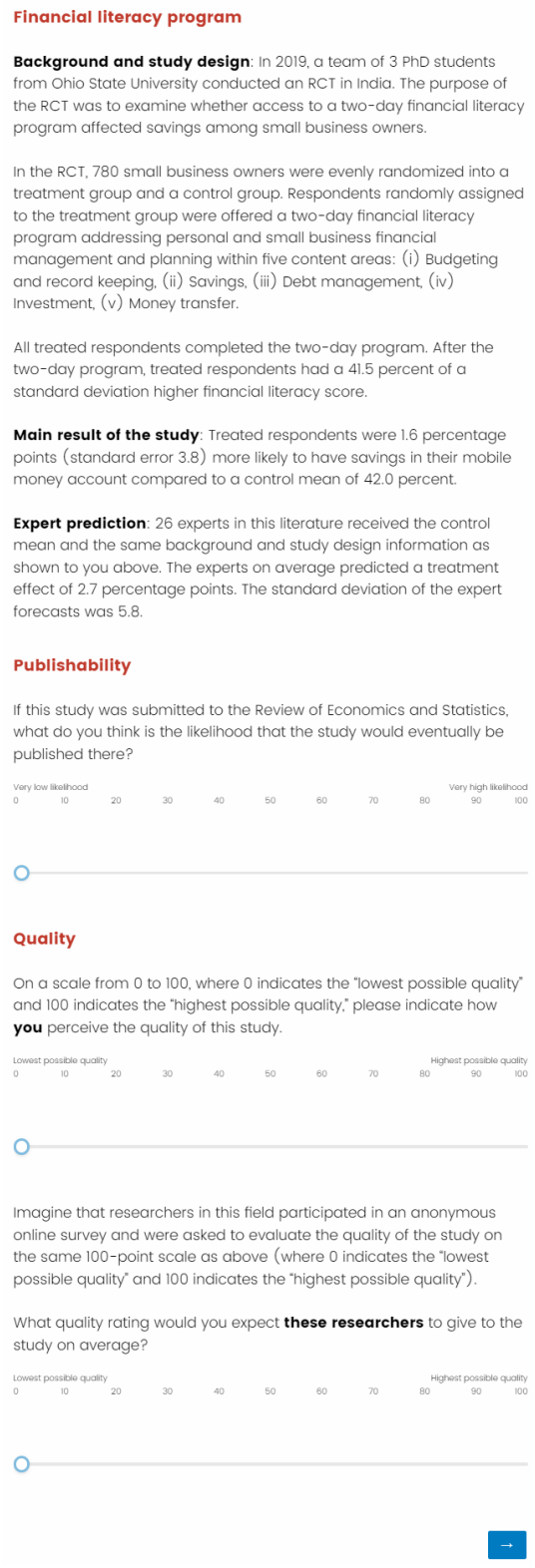
E.1.3 Long-term effects of equal land sharing



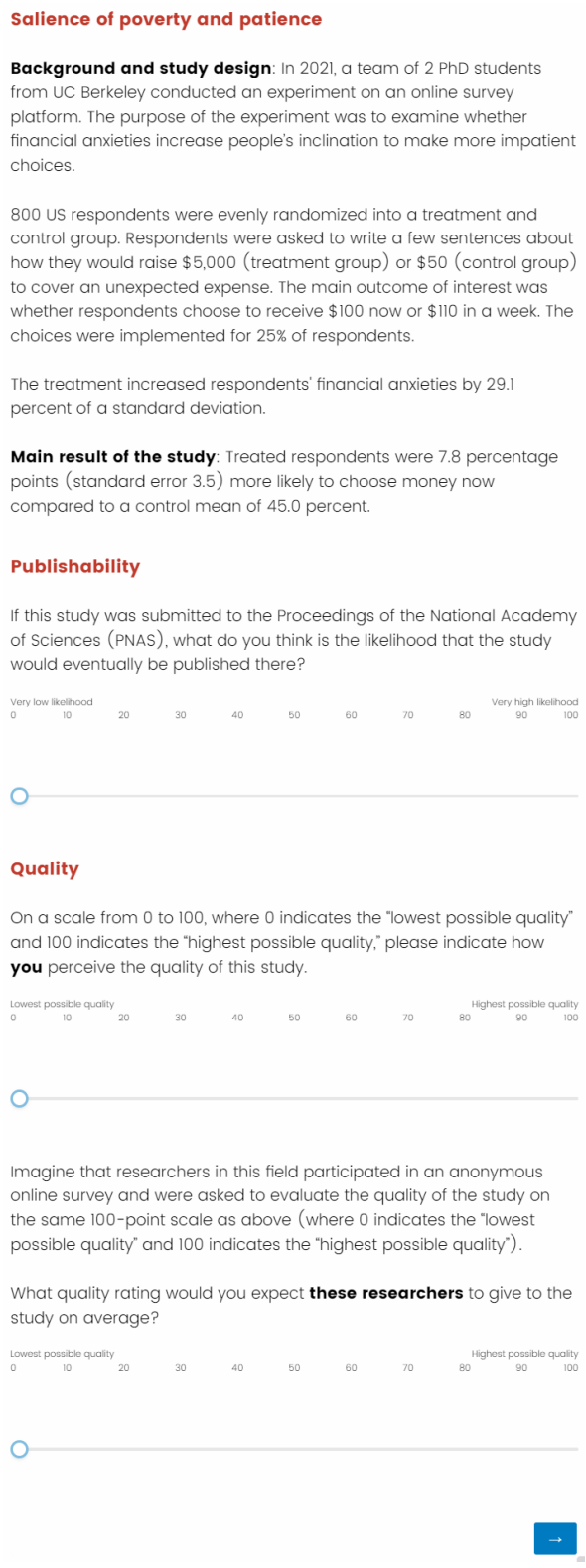
E.1.4 Female empowerment program



E.1.5 Financial literacy program



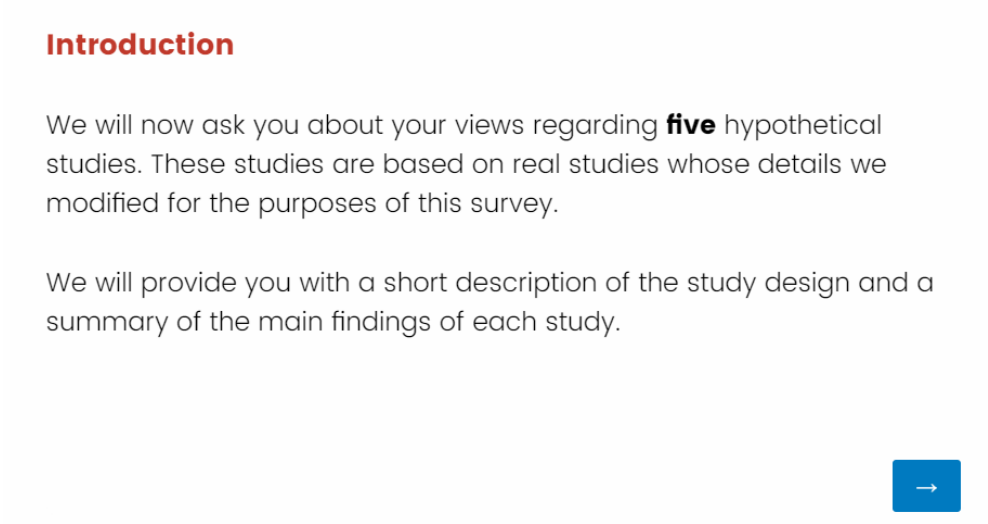
E.1.6 Salience of poverty and patience



E.2 Mechanism experiment

The mechanism experiment was identical to the main experiment except that respondents were shown all five vignettes and that we asked about the precision of the study instead of its quality or importance. Since the wording of the vignettes was identical across the experiments, we only show screenshots of one of the vignettes for the mechanism experiment (the female empowerment program vignette).

E.2.1 Pre-treatment information



E.2.2 Female empowerment program

Female empowerment program

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

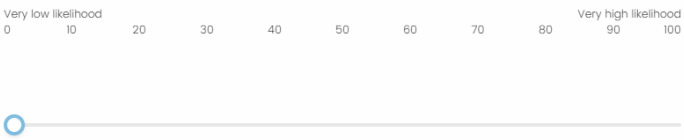
In the RCT, 360 women were evenly randomized into a treatment group and a control group. Respondents in the treatment group were offered a female empowerment program, combining both psychosocial therapy and vocational skills training. The program was very intensive: participants attended meetings for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

Main result of the study: Treated respondents were 1.7 percentage points (p-value = 0.73) more likely to take up a job offer compared to a control mean of 37.0 percent.

Expert prediction: 34 experts in this literature received the control mean and the same background and study design information as shown to you above. The experts on average predicted a treatment effect of 0.6 percentage points. The standard deviation of the expert forecasts was 7.6.

Publishability

If this study was submitted to the Journal of Development Economics, what do you think is the likelihood that the study would eventually be published there?



Precision

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

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

→

F Pre-analysis plans

The data collections were pre-registered in the AsPredicted registry (#95235 and #96599). All of our analyses follow the pre-analysis plans unless otherwise noted. The pre-analysis plans for the main and mechanism experiments are available on the following links:

- Main experiment: `https://aspredicted.org/su6dj.pdf`
- Mechanism experiment: `https://aspredicted.org/83i25.pdf`[2]

[2] For the mechanism experiment, we erroneously wrote that we would invite approximately 150 graduate students and early-career researchers. Our aim with this collection was to obtain a final sample size of approximately 150 graduate students and early-career researchers, which led us to send out 509 invitations.

Do Results Shape the Evaluation of Research? - April 2022 (#95235)

Created: 04/26/2022 07:29 AM (PT)
Public: 05/25/2022 02:08 AM (PT)

Author(s)
Felix Chopra (University of Bonn) - felix.chopra@uni-bonn.de
Ingar Haaland (University of Bergen) - Ingar.Haaland@uib.no
Christopher Roth (University of Cologne) - roth@wiso.uni-koeln.de
Andreas Stegmann (University of Warwick) - andreas.stegmann@warwick.ac.uk

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
We conduct an expert survey using hypothetical vignettes to study how economists' evaluation of scientific research depends on the statistical significance of the main finding.

3) Describe the key dependent variable(s) specifying how they will be measured.
We constructed a total of 5 hypothetical vignettes describing research studies. These vignettes are based on actual research studies whose details we manipulated for the purpose of this survey.

Each vignette contains background information about the research team (seniority and institution). In addition, each vignette provides respondents with a brief description of the research question, the study design, and the main findings.

We ask respondents to evaluate the research studies described in four randomly chosen vignettes based on the information provided. For each of these four hypothetical vignettes, we then measure the following main outcome:

Publishability: We elicit beliefs about the likelihood that the research study will be published in a vignette-specific journal on a scale from 0 to 100.

In addition, we measure four secondary outcomes:

Half of our respondents receive the following two secondary outcomes:

First-order belief about quality: We elicit respondents' perception of the quality of the research study on a scale from 0 (lowest possible quality) to 100 (highest possible quality).

Second-order belief about quality: We ask respondents to imagine that researchers participated in an anonymous online survey and were asked to evaluate the quality of the study on the same 100-point scale as in the previous question. We then ask respondents for the quality rating they would expect these researchers to give to the study on average.

The other half of our respondents receive the following two secondary outcomes:

First-order belief about importance: We elicit respondents' perception of the importance of the research study on a scale from 0 (lowest possible importance) to 100 (highest possible importance).

Second-order belief about importance: We ask respondents to imagine that researchers participated in an anonymous online survey and were asked to evaluate the importance of the study on the same 100-point scale as in the previous question. We then ask respondents for the importance rating they would expect these researchers to give to the study on average.

4) How many and which conditions will participants be assigned to?
We experimentally vary six features across vignettes. Five features vary at the respondent-by-vignette level, and one feature varies at the respondent level.

1) Communication of scientific findings: We exogenously vary whether the statistical significance of the main finding presented in the vignettes is reported by indicating the (a) the main treatment effect estimate along with the associated standard error or (b) the main treatment effect associated with the corresponding p-value implied by the standard error. This feature is varied between subjects.

2) Statistical significance of results: We exogenously vary the effect size of the main finding of the study such that it is either statistically significant at the 5% level, or not. We hold the associated standard error constant across conditions.

3) Expert forecast: We vary whether the vignette includes (a) no expert forecast, (b) information that experts predicted a large effect, or (c) information that experts predicted a small/no effect. The magnitude of the large/small expert prediction is in line with the magnitude of the large/small treatment effect estimate.

4) Seniority: We vary whether the researchers involved in the study are PhD students or Professors.

5) University: We vary whether the researchers involved in the study are affiliated with a top or a lower ranked institution.

6) Journal: We vary the identity of the journal for which we elicit the publishability belief (see section 3). The journal is either a top field journal or a general interest journal.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
5.1 Variable construction: We construct a binary "non-significant" indicator taking value one whenever the main finding reported in a vignette is not statistically significant at the 5% level, and zero otherwise. In addition, we will construct indicator variables for all other features that vary across vignettes (as listed in section 4 of this document).

5.2 Main specification: We will then use OLS regressions where the unit of observation is a respondent-vignette. We will regress our outcome measure on the null finding indicator. In addition, we will include vignette fixed effects and individual fixed effects when pooling observations across vignettes. We will also include indicators for all other features that we experimentally vary across vignettes (as described above). Standard errors will be clustered at the respondent level.

5.3 Heterogeneity analysis: To investigate whether the main treatment has heterogeneous effects, we will separately add interaction terms between the non-significant indicator and an additional dummy variable for other cross-randomized features (seniority, journal, expert forecast, university) to the main specification.

The analysis of heterogeneity in treatment effects as a function of whether non-significant results are communicated by displaying the estimate and the associated standard error or the p-value instead relies on between-subject variation. Consequently, we are not able to include respondent fixed effects as additional controls.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will not exclude any responses from the analysis. There will be no outliers in the remaining survey data as all outcomes are bounded (measured on a 0 to 100 scale).

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
We collected the email addresses of about approximately 16,500 researchers in the field of economics from the top 200 institutions according to RePEc (as of March 2022). We will invite these researchers to participate in our online survey using a Qualtrics invitation email. Our final sample size will depend on the overall response rate.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Nothing else to pre-register.

Do Results Shape the Evaluation of Research? - May 2022 (#96599)

Created: 05/10/2022 04:48 AM (PT)
Public: 05/25/2022 02:07 AM (PT)

Author(s)
Felix Chopra (University of Bonn) - felix.chopra@uni-bonn.de
Ingar Haaland (University of Bergen) - Ingar.Haaland@uib.no
Christopher Roth (University of Cologne) - roth@wiso.uni-koeln.de
Andreas Stegmann (University of Warwick) - andreas.stegmann@warwick.ac.uk

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
We conduct an expert survey using hypothetical vignettes to study how economists' evaluation of scientific research depends on the statistical significance of the main finding.

3) Describe the key dependent variable(s) specifying how they will be measured.
We constructed a total of 5 hypothetical vignettes describing research studies. These vignettes are based on actual research studies whose details we manipulated for the purpose of this survey.

Each vignette contains background information about the research team (seniority and institution). In addition, each vignette provides respondents with a brief description of the research question, the study design, and the main findings.

We ask respondents to evaluate the research studies described in the five vignettes based on the information provided. For each of these five hypothetical vignettes, we then measure the following main outcome:

Publishability: We elicit beliefs about the likelihood that the research study will be published in a vignette-specific journal on a scale from 0 to 100.

In addition, we measure the following secondary outcome:

Perceived precision: We elicit respondents' perception of the precision of the research study's main finding on a 5-point Likert scale (very precisely estimated, precisely estimated, somewhat precisely estimated, not precisely estimated, not at all precisely estimated).

4) How many and which conditions will participants be assigned to?
We experimentally vary six features across vignettes. Five features vary at the respondent-by-vignette level, and one feature varies at the respondent level.

1) Communication of scientific findings: We exogenously vary whether the statistical significance of the main finding presented in the vignettes is reported by indicating the (a) the main treatment effect estimate along with the associated standard error or (b) the main treatment effect associated with the corresponding p-value implied by the standard error. This feature is varied between subjects.

2) Statistical significance of results: We exogenously vary the effect size of the main finding of the study such that it is either statistically significant at the 5% level, or not. We hold the associated standard error constant across conditions.

3) Expert forecast: We vary whether the vignette includes (a) no expert forecast, (b) information that experts predicted a large effect, or (c) information that experts predicted a small/no effect. The magnitude of the large/small expert prediction is in line with the magnitude of the large/small treatment effect estimate.

4) Seniority: We vary whether the researchers involved in the study are PhD students or Professors.

5) University: We vary whether the researchers involved in the study are affiliated with a top or a lower ranked institution.

6) Journal: We vary the identity of the journal for which we elicit the publishability belief (see section 3). The journal is either a top field journal or a general interest journal.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
5.1 Variable construction: We construct a binary "non-significant" indicator taking value one whenever the main finding reported in a vignette is not statistically significant at the 5% level, and zero otherwise. In addition, we will construct indicator variables for all other features that vary across vignettes (as listed in section 4 of this document).

5.2 Main specification: We will then use OLS regressions where the unit of observation is a respondent-vignette. We will regress our outcome measure on the null finding indicator. In addition, we will include vignette fixed effects and individual fixed effects when pooling observations across vignettes. We will also include indicators for all other features that we experimentally vary across vignettes (as described above). Standard errors will be clustered at the respondent level.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will not exclude any responses from the analysis. There will be no outliers in the remaining survey data as all outcomes are bounded (measured on a 0 to 100 scale).

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
We will invite approximately 150 graduate students and early-career researchers in Economics studying at different institutions in Europe to participate in our online survey using a Qualtrics invitation email. Our final sample size will depend on the overall response rate.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Nothing else to pre-register.