DISCUSSION PAPER SERIES

DP17331

The Null Result Penalty

Felix Chopra, Ingar Haaland, Christopher Roth and Andreas Stegmann

PUBLIC ECONOMICS

CEPR

ISSN 0265-8003

The Null Result Penalty

Felix Chopra, Ingar Haaland, Christopher Roth and Andreas Stegmann

Discussion Paper DP17331
Published 28 May 2022
Submitted 27 May 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

The Null Result Penalty

Abstract

In experiments with economists, we measure how the evaluation of research studies depends on whether the 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 statistically significant results, even when holding constant all other study features, including the precision of estimates. The penalty for null results is of similar magnitude for various subgroups of researchers, from PhD students to editors. The null result penalty is larger when experts predict a non-null result and when statistical uncertainty is communicated in terms of p-values rather than standard errors. Our findings have implications for understanding mechanisms underlying publication bias and the communication of research findings.

JEL Classification: N/A

Keywords: N/A

Felix Chopra - felix.chopra@uni-bonn.de
University of Bonn

Ingar Haaland - ingar.haaland@gmail.com
University of Bergen

Christopher Roth - cproth89@gmail.com
University of Cologne and CEPR

Andreas Stegmann - andreas.stegmann@warwick.ac.uk
University of Warwick and CEPR

The Null Result Penalty

Felix Chopra Ingar Haaland Christopher Roth
Andreas Stegmann*

May 27, 2022

Abstract

In experiments with economists, we measure how the evaluation of research studies depends on whether the 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 statistically significant results, even when holding constant all other study features, including the precision of estimates. The penalty for null results is of similar magnitude for various subgroups of researchers, from PhD students to editors. The null result penalty is larger when experts predict a non-null result and when statistical uncertainty is communicated in terms of *p*-values rather than standard errors. Our findings have implications for understanding mechanisms underlying publication bias and the communication of research findings.

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

*We thank all participants of this study for generously sharing their time. We also thank Peter Andre, Isaiah Andrews, Lukas Hensel, Johannes Hermle, Max Kasy, Matt Lowe, Nick Otis, Jesse Shapiro, and Dmitry Taubinsky for excellent suggestions. 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 Bonn, Felix.Chopra@uni-bonn.de; Haaland: University of Bergen, Ingar.Haaland@uib.no; Roth: University of Cologne, ECONtribute, roth@wiso.uni-koeln.de; Stegmann: University of Warwick, 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 relies on a well-functioning publication system that evaluates research studies without bias. However, the publication system may favor research papers reporting large and statistically significant results over research papers documenting small results that are not statistically significant (Camerer et al., 2016; Simonsohn et al., 2014). Selection of this type can lead to biased estimates and misleading confidence sets in published studies (Andrews and Kasy, 2019).

In this paper, we investigate whether there is a penalty in the publication system for research studies with null results and, if so, what mechanisms can explain the penalty. To address these questions, we conduct an experiment with a sample of about 500 researchers from the leading top 200 economics departments in the world. Our main treatment varies whether a research study reports a statistically significant main finding. Studies that obtain and do not obtain statistically significant results might differ systematically in important dimensions. For instance, studies that obtain statistically significant results might, on average, have higher precision of the estimates, rendering a statistically significant effect more likely. To study whether there is a *penalty* for null results, we therefore rely on an experimental approach that keeps all other study characteristics constant. In the main experiment, 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 four vignettes, we randomize whether the coefficient estimate of the treatment effect was sizable and statistically significant or close to zero and not statistically significant. We keep the standard error of the main finding constant across treatments such that null findings are not associated with lower statistical precision of the estimates.

To examine how the evaluation of research results depends on expert priors (DellaVigna et al., 2019), we cross-randomize whether the vignette includes an expert forecast of the treatment effect. For vignettes including an expert forecast, we further randomize whether the experts predict a null or a non-null result. 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. Finally, to obfuscate the purpose of the study, we further cross-randomize a series of other salient study characteristics, including the seniority of the researcher 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 private 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 document that studies with a null result are perceived to be less likely to be published, 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, a null result makes our respondents associate the study 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 yielding 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 null result penalty is of similar magnitude for various subgroups of researchers, from PhD students to editors. 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 and of lower quality.

A longstanding concern in the academic community is that an excessive focus on p-values could amplify problems related to the reproducibility and 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 of treatment effects by whether respondents were given information about the p-value or the standard error of the main estimate presented in the study vignettes. We find that the null result penalty is 3.7 percentage points larger

when the main results are reported with p-values (95% C.I. [-7.9,0.60]; $p = 0.092$). Moreover, reporting results with p-values instead of the standard error further leads our respondents to associate a null result study with 34.4% of a standard deviation lower quality (95% C.I. [-58.3,-10.4]; $p = 0.005$) and also makes them think other researchers will associate the study with 36.2% of a standard deviation lower quality (95% C.I. [-59.8,-12.6]; $p = 0.003$).

The null result penalty can lead to biased estimates and misleading confidence sets in published studies (Andrews and Kasy, 2019). However, a null result penalty might be optimal depending on the social objectives underlying the publication process. For example, if the decision of which studies to publish is driven by a desire to maximize the policy impact, Frankel and Kasy (2022) formally show that this prescribes a benchmark publication rule that favors research studies with surprising findings relative to the prior in the literature.[1] 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. Second, we find that the negative effect of null results on publishability is aggravated when a null result is at odds with expert forecasts: Respondents evaluate a study with a null result as having a further 6.3 percentage points lower chance of being published (95% C.I. [-11.4,-1.1]; $p = 0.018$). These patterns are inconsistent with the conjecture that respondents believe that the publication process favors research findings with surprising results. In addition, we find that receiving an expert forecast predicting a treatment effect does not significantly differentially affect perceptions of quality (95% C.I. [-40.9,24.2]; $p = 0.615$). This suggests that respondents do not make differential inferences about unobservable study characteristics when receiving different expert forecasts, but instead believe that surprising null results are more strongly discounted in the publication process.

Finally, we conduct an additional experiment that focuses on PhD students and 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 statistical precision as studies with larger and statistically significant effects. We employ the same vignettes and main outcomes 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 null results with 19.8

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

percentage point (or 32.5%) lower probability of being published (95% C.I. [-24.3,-15.2]; $p < 0.001$). Furthermore, they associate null results studies with 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. In particular, the data indicates that researchers might penalize null results in the publication system partly because of errors in statistical reasoning.

Our study relates to a growing literature on the publication process (Card and DellaVigna, 2013, 2020; Card et al., 2020; Frankel and Kasy, 2022; Kasy, 2019) and in particular publication bias (Brodeur et al., 2016, 2020; Franco et al., 2014; Gerber et al., 2008; Ioannidis, 2005).[2] This literature has examined the extent to which null results are less likely to be published (Simonsohn et al., 2014). 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 a spike just below the 0.05 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.

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 statistically significant results might be systematically different from studies with null 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.[3] We also shed light on the mechanisms underlying the null result penalty with rich data on how null

[2] A related literature has examined the replicability of research finding (Camerer et al., 2016, 2018) and has discussed research transparency efforts (Christensen et al., 2019).

[3] Identifying the role of selection up until the submission stage versus selection conditional upon submission is challenging with observational data.

results shape perceptions of the quality, importance, and precision of the studies. Our finding that null results are perceived to be more noisily estimated suggests some role for errors in statistical reasoning. Finally, our experimental approach also allows us to study potential measures to mitigate the null results penalty, such as providing expert forecasts and expressing statistically uncertainty in terms of standard errors rather than p-values.

Moreover, 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) and policymakers (Vivalt and Coville, 2020). Vivalt and Coville (2019) study how policy makers update beliefs from evidence and Vivalt et al. (2021) examine which attributes policy-makers care about. We assess how academic economists' perceptions of the publishability, quality, importance, and precision of research studies hinge on the results of the study.

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 two hundred institutions according to RePEc (as of March 2022) to participate in a 10-minute online survey. In total, 480 researchers follow our invitation and complete the online survey, implying an overall response rate of 3.4%.

Table 2 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 placed 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 (the American Economic Review, Econometrica, the Journal of Political Economy, the Quarterly Journal of Economics, and the Review of Economic Studies). 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 4348.3 (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. These summary statistics underscore that our expert sample is mostly comprised of experienced researchers with substantial academic impact in the field of economics. They 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%), and macroeconomics (14.1%).

Pre-specification The data collections were pre-registered in the AsPredicted registry (#95235 and #96599). We pre-specified the sampling procedure, the main outcomes of interest, the main right-hand-side variable of interest, as well as the baseline specifications. The pre-analysis plans can be found in Section E.

2.2 Design

Baseline design For the purpose of this experiment, we created five hypothetical vignettes describing different research studies. Each vignette is loosely based on an actual research paper in economics. 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). Table 1 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 vignette level): Half of the respondents are informed that the study had a sizable main effect (*positive result* treatment), while the other half of respondents are informed that the study had a main effect close to zero (*null result* treatment). Importantly, while we vary the point estimate between treatments, we keep the sample size and the standard error of the estimate constant across treatments. We construct the standard errors such that the main effect is statistically significant in the positive result treatment (at conventional significance thresholds) and not significant in the null result treatment.[4]

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 *positive results* treatment of this vignette receive the following instructions:

[4] Appendix Section C 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. Our approach ensures that the numerical values (e.g. effect size, standard error, number of observations) are internally consistent for each version of a vignette that is presented to our respondents.

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

Expert predictions We cross-randomize at the vignette level whether we provide respondents with expert predictions of the main treatment effect estimate, allowing us to examine whether the evaluation of null results depends on whether the result is surprising to experts or in line with expert predictions.[5] 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 high and low expert forecasts such that they are close, but not identical, to the magnitude of the coefficient estimate in the positive and null result treatments, respectively. We also provide the standard deviation of the expert forecast to communicate the degree of disagreement among experts. To ensure that the main result of the study is informative from a Bayesian perspective, 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.[6]

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.

[5] 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).

[6] 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.

Similarly, 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.

Communication treatment A common view in the academic community is that an excessive focus on *p*-values might amplify problems related to the reproducibility and replicability of scientific findings (Camerer et al., 2016; Wasserstein and Lazar, 2016). To examine how the statistical communication of findings affects the evaluation of 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 estimate. To minimize the scope for experimenter demand effects (de Quidt et al., 2018), we cross-randomized this feature between subjects.

In the context of the vignette on merit aid for low-income students, respondents assigned to the *p*-value treatment and 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 (p-value = 0.73) compared to a control mean of 17 percent.

By contrast, respondents assigned to the standard error treatment and the *positive 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.

Obfuscation treatments Finally, 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.[7] Both of these pieces of information are featured at the beginning of the "Background and study design" part 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 related social desirability effects (Haaland et al., 2021). This additional within-respondent variation substantially increases the difficulty to guess the true purpose of our research study as respondents only observe variation in vignette features for a relatively small number of vignettes. For instance, by making the university affiliation and the seniority of the research team salient, respondents could have guessed that we wanted to study discrimination against younger researchers or researchers from lower-ranked institutions in the publication process. 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.

2.3 Main outcomes

After the presentation of each vignette, we ask our respondents three questions. Our main outcome of interest is 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.[8] For example, for the vignette on merit aid for low-income students, we cross-randomize whether respondents estimate the likelihood that the paper will eventually be published in the Economic Journal or the Journal of Human Resources. 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 zero and one hundred.

Second, to examine mechanisms, we then measure respondents' first and second

[7] The higher ranked institutions we use in the vignettes are: Harvard University, Columbia University, UC Berkeley, Northwestern University and Brown University. The somewhat lower ranked institutions we use are: Ohio State University, the University of Pittsburgh, Boston College, the University of Arkansas and the University of Illinois.

[8] We include the following general interest journals: The Economic Journal, The Review of Economic Studies, Science, Review of Economics and Statistics, Proceedings of the National Academy of Sciences. We include the following field journals: Journal of Human Resources, Journal of Economic Growth, Journal of Development Economics, Journal of Public Economics, Experimental Economics.

order beliefs either 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. The data on second-order beliefs allows us to shed light on a potential wedge between private beliefs and the perceived beliefs of other researchers in the field. To reduce concerns about survey fatigue, we cross-randomize at the individual 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 beliefs about other researchers' assessment of the quality of the research study 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 as above. 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.

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 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 not statistically significant, and zero otherwise; X_{iv} is a vector of binary indicators for other cross-randomized treatment conditions (described in more detail in Section 2.2); 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 other

cross-randomized features, we additionally include all interaction terms between the *null result* indicator and the set of indicators X_{iv} for all other cross-randomized features.

3.2 Main treatment effects

Table 3 shows the average treatment effects on our main outcomes of interest. 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.

Column 2 shows the effects on private beliefs about the quality of the studies. Respondents in the *null result* treatment associate the studies with 37.3% of a standard deviation lower quality (95% C.I. [-49.6,-25]; $p < 0.001$). Furthermore, as shown in column 3, respondents in the *null result* treatment similarly think that other researchers would associate the studies with 46% of a standard deviation lower quality (95% C.I. [-58.1,-33.8]; $p < 0.001$).

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]; $p < 0.001$).

Robustness We find a very similar null result penalty irrespective of whether the relevant journal is a field journal or a general interest journal (see Table B.2). The treatment effects are also robust to alternative approaches of analyzing the data. Panel B of Table 3 shows that we get virtually identical results when excluding individual-level fixed effects.[9] Moreover, Figure B.1 shows that treatment effects are very homogeneous across the vignettes, suggesting that the null result penalty is not specific to any particular vignette.

Furthermore, Table B.1 shows that we obtain quantitatively similar point estimates if we restrict the sample to the first, randomly selected vignette presented to each

[9] In contrast to the main specification, this specification also uses respondents who are always shown studies with null results and those always exposed to studies with non-null results.

respondent. This robustness check exploits only the between-subject variation in treatment assignment from the first vignette, which addresses several concerns. First, it mitigates the potential concern that 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 higher at the beginning of the survey. Using only variation from the first vignette thus mitigates concerns that respondents' response behavior could increasingly rely on heuristics—such as focusing on the statistical significance of findings—at later stages in the survey.[10]

As discussed in Section 2.1, there is substantial heterogeneity in experience with the production and evaluation of research studies among our respondents. In Figure 1, we show that our treatment effects are similar across different subgroups. In particular, the null result penalty is of similar magnitude among experienced researchers, such as editors or those with above median number of citations on Google Scholar, and less experienced researchers, such as respondents with below median citations and PhD students. We also see similar treatment effects among males and females and depending on whether respondents have published in one of the top five economics journals. 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 on the publication process. Furthermore, as shown in Figure B.2, we also see that the null result penalty is homogeneous across different fields of specialization, including among respondents who specialize in econometrics.[11]

3.3 Do p-values aggravate the null result penalty?

To test whether the statistical presentation of results affects the penalty for studies with null results, we vary between subjects whether the statistical uncertainty is communicated in terms of standard errors or p-values. Column 1 of Table 4 shows that communicating the results in terms of p-values rather than standard errors further decreases the perceived publishability and quality of research papers reporting main findings that are not significant. These effects are substantial: The null result penalty

[10] Treatment effects do not interact with the order of the vignette presented ($p = 0.660$).

[11] Interestingly, the only group for which we do not observe a significant null result penalty is among respondents who specialize in economic theory, but the confidence interval is wide and the point estimate is negative.

on perceived publishability is 3.7 percentage points higher when results are presented in terms of p-values rather than standard errors (95% C.I. [-7.9,0.60]; $p = 0.092$). 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 higher order beliefs about others' quality assessment are further increased by 37.3% of a standard deviation (95% C.I. [-49.6,-25]; $p < 0.001$) and 46% of a standard deviation (95% C.I. [-58.1,-33.8]; $p < 0.001$), respectively. 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).[12]

4 Mechanisms

4.1 Do people prefer 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: 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 (Frankel and Kasy, 2022).

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

To test this conjecture, we examine heterogeneity in null result penalties by whether the experts predicted a null result or not. Table 4 shows interaction effects between the *null result* treatment and treatment indicators for being shown a high expert forecast

[12]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, very likely this calculation is too complex for our respondents, leading them to rely on simple heuristics instead.

or a low expert forecast. As shown in column 1, respondents provided with a low expert forecast do not differentially update their beliefs about the publishability of the study in a statistically significant way (95% C.I. [-6.7,3]; $p = 0.451$). In contrast, respondents in the *null result* treatment who also receive the high expert forecast instead of no expert forecast think the studies have a 6.3 percentage points lower chance of being published (95% C.I. [-11.4,-1.1]; $p = 0.018$).[13] In other words, the effect of the *null result* treatment on perceived publishability is even exacerbated when the null result is at odds with expert predictions, suggesting that researchers in our sample do not perceive surprising results to be rewarded in the publication process. In addition, we find that receiving an expert forecast predicting a treatment effect does not significantly differentially affect perceptions of quality (95% C.I. [-40.9,24.2]; $p = 0.615$). This suggests that respondents do not make differential inferences about unobservable study characteristics when receiving different expert forecasts, but instead believe that surprising null results are more strongly discounted in the publication process. Naturally, the conclusions from this analysis are somewhat more suggestive as the interaction effects between expert predictions and the null result treatment are more noisily measured than the main effect.

From a positive standpoint, 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. 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 maximization of policy impact would prescribe (see Frankel and Kasy (2022) for a formalization). Taken together, this suggests that a mechanism in which the publication process rationally maximizes a social objective that requires a null result penalty is unlikely to explain the patterns in our data.[14]

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

[14] As we elicited positive beliefs about the status quo, our participants may of course privately think that the publication process *should* maximize one of the two social objectives.

4.2 Learning about unobservables?

One explanation for the null result penalty is based on inference about the quality of a study. For such a mechanism to operate, we need two ingredients. First, researchers can only imperfectly observe the quality of a research study, where quality determines the likelihood of uncovering the true effect size. Second, researchers start from a prior that favors a positive and significant effect. Observing a null result will then cause a Bayesian to infer that the research study is more likely to be of low quality (which we formally show in Appendix Section A). Consistent with this account, we find that participants negatively update about the quality of research studies with null results (as shown in Table 3), suggesting that our participants start from a prior that there is a causal relationship. We also find that receiving an expert forecast predicting a large effect rather than no effect aggravates the null results penalty (column 1 of Table 4). This suggests that participants internalize the information about the expert predictions to some degree. However, if people start from a prior favoring an effect, a Bayesian mechanism would predict that null results should be evaluated *more* favorably when experts predict no effect—the opposite of what we find. In addition, we find that, if anything, participants update somewhat more negatively about the quality of research studies with unsurprising nulls compared to surprising nulls. While these effects are somewhat noisily estimated, they are directionally at odds with rational inference about unobserved study characteristics.

Heterogeneity by priors about quality 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 who have less research experience. The Bayesian prediction would thus be that researchers should update more strongly about the quality of a study if the authors are either affiliated with lower-ranked institutions or of lower seniority.

We find that researchers expect articles of PhD students and researchers from lower-ranked universities to be less likely to be published even though they do not perceive any quality differences. Yet, we find only very muted interaction effects between the null result treatment and being a PhD student or being affiliated with a lower-ranked

university on the extent of the null result penalty (see column 1 of Table B.2). The lack of heterogeneous treatment effects on our outcome variables 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.

4.3 Perceived statistical precision

We conducted an additional experiment to examine whether researchers' subjective beliefs about the statistical precision of a study depends on whether the study yielded a null result, while providing respondents with information about the actual precision of the estimate.

Sample In May 2022, we invited an additional 509 graduate students and early career researchers in the field of economics to participate in a 10-minute online survey which we designed to complement our main experiment by providing further evidence on underlying mechanisms. In total, 95 graduate students and early career researchers follow our invitation and complete the survey, implying a response rate of 17%. The graduate students and early career researchers in this sample 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 the conjecture that researchers perceive studies with null results to be less precisely estimated, even when they are provided with the standard error of the estimate. The design is identical to our main experiment except for two differences. First, respondents are asked to rate the statistical precision of the main result on a 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 that were included in the main experiment. Second, respondents are exposed to all five vignettes, which gives us 475 observations at the vignette-respondent level.

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]; $p < 0.001$). These findings are robust to not including respondent-level fixed effects (as shown in Panel B).

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. This finding suggests that researchers may use simple heuristics to assess the statistical precision of estimates, such as a heuristic that relies on statistical significance to form beliefs about the precision of coefficient estimates. These results indicate that part of the null result penalty may be due to a behavioral bias in how researchers evaluate the statistical precision of studies.

5 Conclusion

We show that research studies with null results are perceived to be substantially less publishable, of lower quality, of lower importance, and less precisely estimated than studies with statistically significant results, even when holding constant all other study features, including the statistical precision of estimates. The null result penalty is even larger when experts predict a non-null result, inconsistent with people perceiving the publication process to favor surprising results. Communicating the statistical uncertainty of study results in terms of p-values rather than standard errors further decreases the perceived publishability and quality of research papers with findings that are not statistically significant.

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 telling them about the informativeness and importance of null results (Abadie, 2020). Finally, one practical implication of our study is that communicating statistical uncertainty of results in terms of standard errors rather than p-values might help to counteract potential errors in statistical reasoning.

References

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

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

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

_ **, 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.

Bogdanoski, Aleksandar, Andrew Foster, Dean Karlan, and Edward Miguel, "Preresults Review at the Journal of Development Economics: Lessons learned," 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, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu, "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, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu**, "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," *The Quarterly Journal of Economics*, 2012, *127* (4), 1755–1812.

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

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.

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, Neil Malhotra et al., "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.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature*, 2021.

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.

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

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

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons, "p-curve and effect size: Correcting for publication bias using only significant results," *Perspectives on Psychological Science*, 2014, *9* (6), 666–681.

Vivalt, E. and A. Coville, "Policy-makers consistently overestimate program impacts," Technical Report, Working Paper 2020.

Vivalt, Eva, Aidan Coville, and Sampada KC, "Weighing the Evidence: Which Studies Count?," *Working Paper*, 2021.

_ **and** _ , "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.

Main figures and tables

Table 1: 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
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.

Table 2: Summary statistics

	Mean	Standard deviation	Median	Observations
Demographics:				
Male	0.780	0.415	1	477
Years since Ph.D.	14.805	12.899	11	308
PhD student	0.244	0.430	0	467
Region of institution:				
Europe	0.544	0.499	1	478
North America	0.406	0.492	0	478
Australia	0.033	0.180	0	478
Asia	0.017	0.128	0	478
Academic output:				
H-index	17.216	17.654	12	328
Citations	4348.341	10801.581	846	328
Number of top 5 publications	1.268	3.892	0	462
Number of top 5s refereed for	1.166	1.694	0	397
Research evaluation:				
Current editor	0.072	0.259	0	443
Current associate editor	0.127	0.333	0	441
Ever editor	0.151	0.358	0	444
Ever associate editor	0.193	0.395	0	441
Professional memberships:				
NBER affiliate	0.084	0.277	0	454
CEPR affiliate	0.171	0.377	0	451
Academic fields:				
Labor	0.211	0.408	0	418
Public	0.129	0.336	0	418
Development	0.179	0.384	0	418
Political	0.167	0.374	0	418
Finance	0.105	0.307	0	418
Experimental	0.062	0.242	0	418
Behavioral	0.091	0.288	0	418
Theory	0.067	0.250	0	418
Macro	0.141	0.349	0	418
Econometrics	0.141	0.349	0	418

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). "Male" is a binary indicator taking the value one for male respondents, and zero otherwise. 'Years since PhD" is the number of calendar years between 2022 and the year the experts obtained their PhD. "Number of top 5 publications" 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). "H-index" and 'Citations" are, respectively, their H-index and their total number of citations as taken from their Google Scholar profile (as of May 2022). "Asia", "Australia", "Europe", and "North America" are regional indicators taking the value one if the institution the expert works for is based in the region.

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: 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***
	(1.136)	(0.066)	(0.069)	(0.061)	(0.067)
Observations	1,920	920	920	1,000	1,000

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) individual-level fixed effects. All regressions in both panels include treatment indicators for the cross-randomized conditions in addition to vignette-level fixed effects. Standard errors are clustered at the individual level.

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

Table 4: Main results: Heterogeneity by expert forecasts and statistical communication

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Panel A: Fixed effects					
Null result treatment	-11.072***	-0.029	-0.219	-0.330**	-0.390***
	(2.681)	(0.151)	(0.160)	(0.132)	(0.135)
Null result × Low expert forecast	-1.862	-0.169	0.130	0.030	0.058
	(2.470)	(0.162)	(0.159)	(0.120)	(0.117)
Null result × High expert forecast	-6.251**	-0.083	0.033	0.048	-0.025
	(2.632)	(0.165)	(0.152)	(0.124)	(0.127)
Null result × P-value framing	-3.652*	-0.344***	-0.362***	-0.021	0.049
	(2.164)	(0.122)	(0.120)	(0.109)	(0.112)
Panel B: No individual FE					
Null result treatment	-11.161***	-0.004	-0.126	-0.412**	-0.464**
	(3.063)	(0.172)	(0.186)	(0.168)	(0.181)
Null result × Low expert forecast	-1.022	-0.196	0.062	0.135	0.107
	(2.937)	(0.166)	(0.185)	(0.150)	(0.161)
Null result × High expert forecast	-7.744***	-0.273	-0.303*	0.110	0.078
	(2.853)	(0.168)	(0.179)	(0.152)	(0.161)
Null result × P-value framing	-4.951**	-0.291**	-0.272*	0.136	0.086
	(2.425)	(0.136)	(0.141)	(0.124)	(0.134)
Observations	1,920	920	920	1,000	1,000

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. "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). "P-value framing" is a treatment indicator taking the value one if the vignette treatment effect had an associated p-value and zero if had an associated standard error 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 fully interacted with "Null result treatment" in addition to vignette-level fixed effects (the interactions for all the cross-randomized conditions with individual-level fixed effects are shown in Table B.2). Standard errors are clustered at the individual level.

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

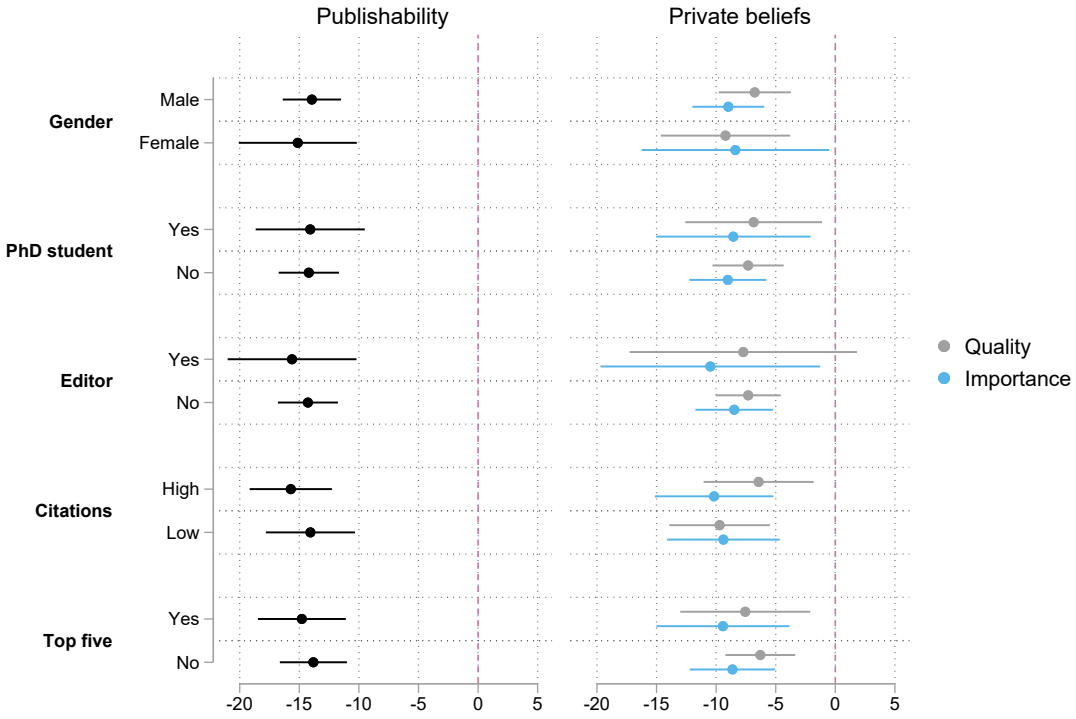
Table 5: Main results: Mechanism experiment on perceived precision

	(1) Publishability (in percent)	(2) Precision (z-scored)
Panel A: 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

Note: The table shows regression estimates of our treatment effects on our key outcomes of interest 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) individual-level fixed effects. All regressions in both panels include treatment indicators for the cross-randomized conditions in addition to vignette-level fixed effects. Standard errors are clustered at the individual level.

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

Figure 1: Heterogeneity in treatment effects



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 private 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 treatments at the vignette level as well as individual fixed effects. Standard errors are clustered at the individual level. 95% confidence intervals are indicated in the figure.

For online publication only:

The Null Result Penalty

Felix Chopra, Ingar Haaland, Christopher Roth and Andreas Stegmann

Section A presents theoretical results.

Section B contains additional tables and figures.

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

Section D provides screenshots of the experimental instructions.

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

A Theoretical framework: Updating about quality

We briefly formalize a Bayesian mechanism where researchers update negatively about the quality of studies with null findings.

A research study tests whether there is a causal relationship between two variables X and Y. Let $\omega \in \{0,1\}$ denote whether there is a causal relationship ($\omega = 1$) or not ($\omega = 0$). An observer has a prior belief $p = P(\omega = 1) \in (0,1)$ that there is a causal relationship. The research study generates a binary signal $s \in \{0,1\}$, e.g. whether the estimate of the causal relationship is statistically significantly different from zero.

Research studies differ in their characteristics $(\theta_{\text{obs}}, \theta)$. Here, θ_{obs} denotes characteristics that are perfectly observed such as the sample size; and θ denotes unobserved quality, such as the clarity of the experimental instructions. These study characteristics determine the precision of the signal provided by a research study:

$$P(s = 1|\omega = 1, (\theta_{obs}, \theta)) = P(s = 0|\omega = 0, (\theta_{obs}, \theta)) = \pi(\theta_{obs}, \theta) \tag{2}$$

We will now examine how a Bayesian observer will update his belief about the characteristics (θ_{obs}, θ) of a study depending on whether it yielded statistically significant results. As there is no uncertainty about the observable characteristics of a study (θ_{obs}), the observer will only about his belief about θ. We will therefore suppress θ_{obs} from the notation to simplify the exposition.

Suppose that studies either have a high quality ($\theta = H$) or a low quality ($\theta = L$). Then

$$P(s = 1|\omega = 1) = P(s = 0|\omega = 0) = \pi(\theta) \tag{3}$$

We assume that $\pi \in (0.5, 1]$ to ensure that the signals are informative. The observer has a prior belief $\rho \in (0,1)$ that a study is of high quality.

How will the observer update his prior belief ρ after observing s? Let $\hat{\rho}(s)$ denote the posterior belief about ρ after observing the study's findings. Then

$$\hat{\rho}(0) = \frac{P(s = 0|H)P(H)}{P(s = 0|L)P(L) + P(s = 0|H)P(H)} \tag{4}$$

$$= \frac{(p(1-\pi_H) + (1-p)\pi_H)\,\rho}{(p(1-\pi_L) + (1-p)\pi_L)\,(1-\rho) + (p(1-\pi_H) + (1-p)\pi_H)\,\rho} \tag{5}$$

One can then show algebraically that $\hat{\rho}(0) < \rho$ if and only if $p > 0.5$. In other words, an observer that believes in a causal relationship will revise his belief that a study is of high-quality downwards after observing a main finding that is not significant.

Analogously, one can show that

$$\hat{\rho}(1) = \frac{(p\pi_H + (1-p)(1-\pi_H))\,\rho}{(p\pi_L + (1-p)(1-\pi_L))\,(1-\rho) + (p\pi_H + (1-p)(1-\pi_H))\,\rho}. \tag{6}$$

Here, one can show that $\hat{\rho}(1) > \rho$ if and only if $p > 0.5$. Thus, a Bayesian observer will revise his belief about the quality of a study upwards after observing a significant finding.

B Additional tables and figures

Table B.1: 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
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-level fixed effects in all regressions. Standard errors are clustered at the individual level.

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

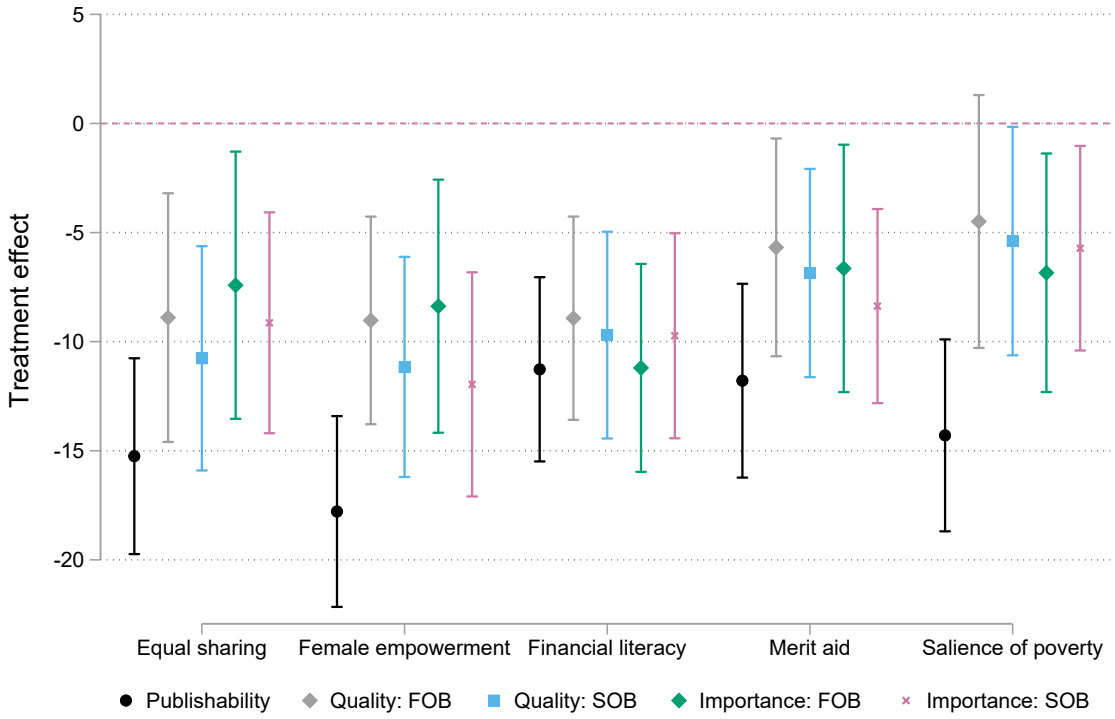
Table B.2: Treatment heterogeneity with all conditions displayed

	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 treatment	-11.072***	-0.029	-0.219	-0.330**	-0.390***
	(2.681)	(0.151)	(0.160)	(0.132)	(0.135)
Interaction effects:					
Null result × Low expert forecast	-1.862	-0.169	0.130	0.030	0.058
	(2.470)	(0.162)	(0.159)	(0.120)	(0.117)
Null result × High expert forecast	-6.251**	-0.083	0.033	0.048	-0.025
	(2.632)	(0.165)	(0.152)	(0.124)	(0.127)
Null result × Field journal	0.871	0.003	-0.027	0.038	0.006
	(1.966)	(0.131)	(0.121)	(0.101)	(0.103)
Null result × PhD student	1.707	-0.165	-0.196*	-0.047	-0.101
	(2.054)	(0.121)	(0.108)	(0.102)	(0.098)
Null result × Low-ranked university	0.408	0.021	0.028	-0.011	-0.018
	(1.965)	(0.121)	(0.124)	(0.108)	(0.106)
Null result × P-value framing	-3.652*	-0.344***	-0.362***	-0.021	0.049
	(2.164)	(0.122)	(0.120)	(0.109)	(0.112)
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)
Observations	1920	920	920	1000	1000

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. "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 had an associated standard error estimate. We include individual and vignette fixed effects in all regressions. Standard errors are clustered at the individual level.

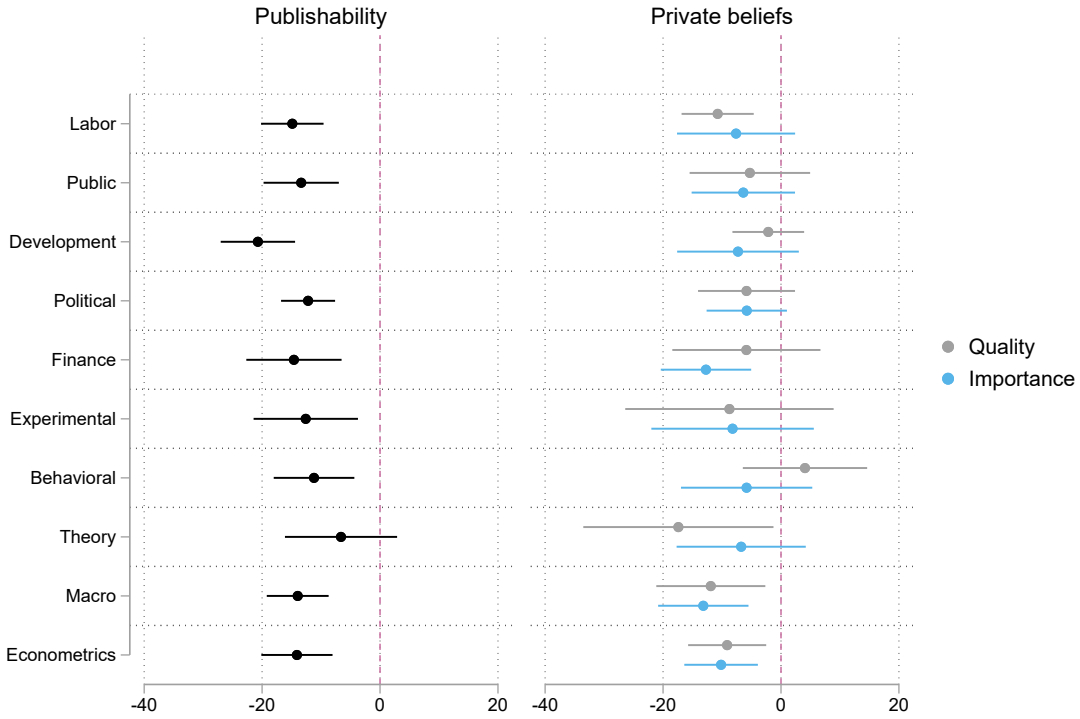
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Figure B.1: Robustness: Vignette-specific treatment effects



Note: This figure shows regression estimates of our treatment effects in which the "Null result treatment" treatment indicator has been interacted with the five vignette-indicators. The regressions include controls for all cross-randomized treatments at the vignette level as well as individual 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 private 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 individual level. 95% confidence intervals are indicated in the figure.

Figure B.2: 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 private beliefs about importance and quality of the study (both measured on a scale from 0 to 100) are regressed n the "Null result treatment" indicator, separately for each sub-group indicated in the figure. The regressions include controls for all cross-randomized treatments at the vignette level as well as individual fixed effects. Standard errors are clustered at the individual level. 95% confidence intervals are indicated in the figure.

C Numerical features used in the vignettes

One of our design goal was to remain close 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 tried 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).

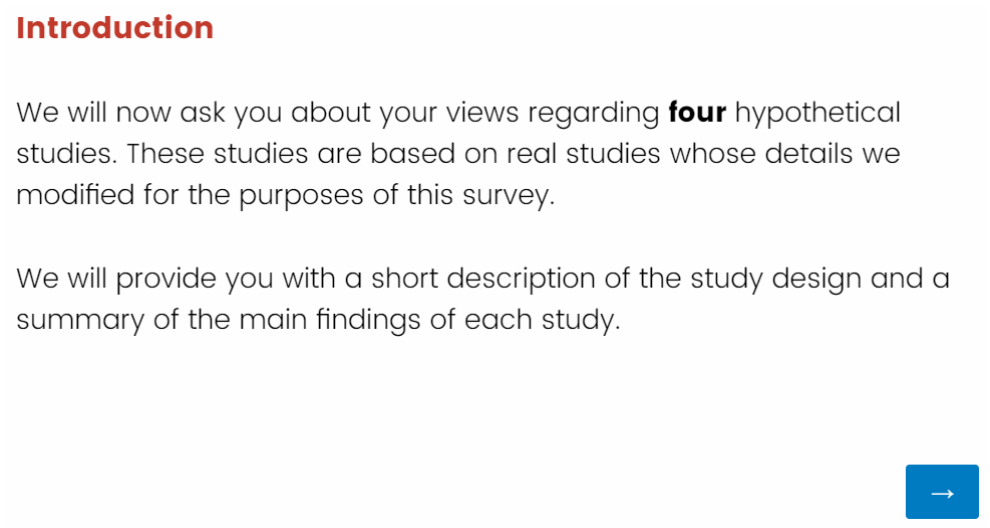
- p-values (high and low): The p-values are obtained from the hypothetical t-statistic used to generate the statistically (non-)significant main effect.
- Expert prior (high mean): This number is equal to $\mu_{\text{high}} + 0.25x$. Here, μ_{high} is the statistically significant main effect and $x \sim N(0, S)$ 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 Screenshots

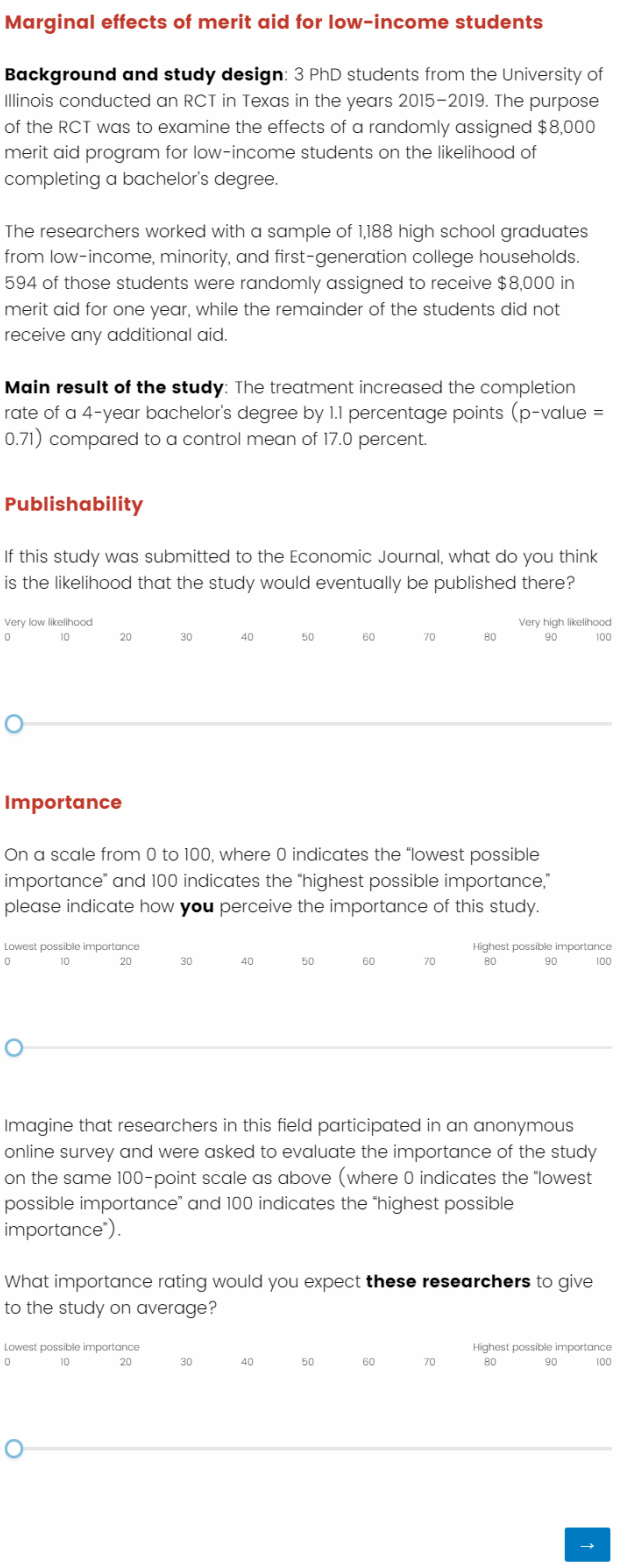
D.1 Main experiment

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.

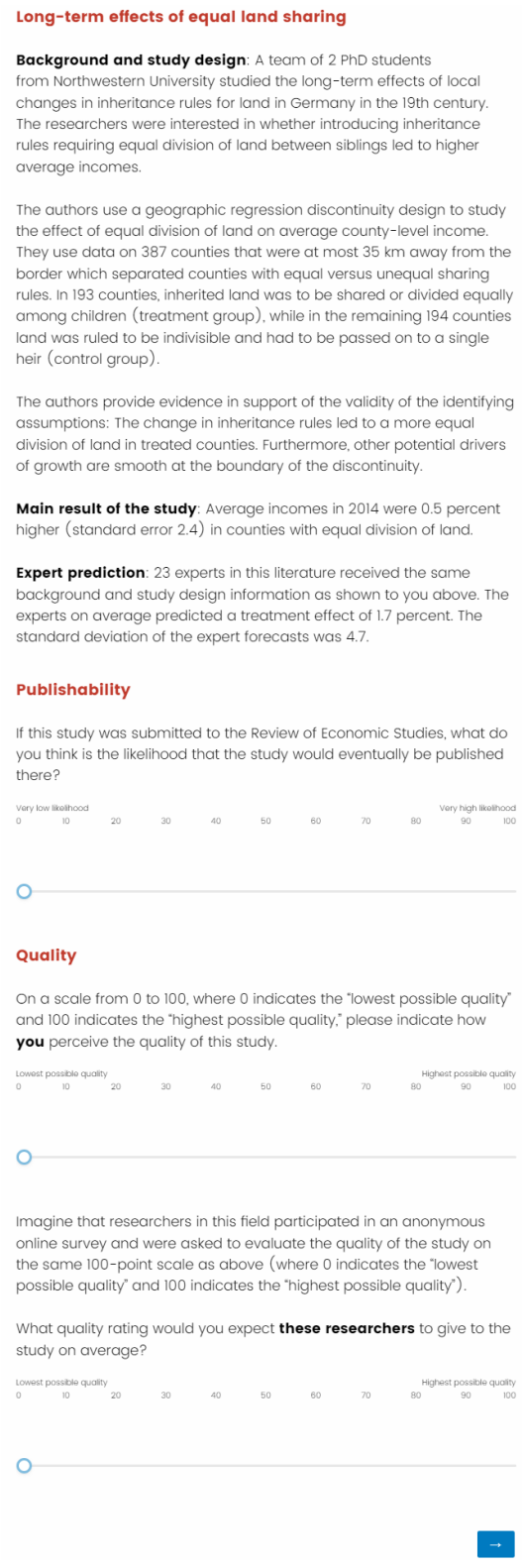
D.1.1 Pre-treatment information



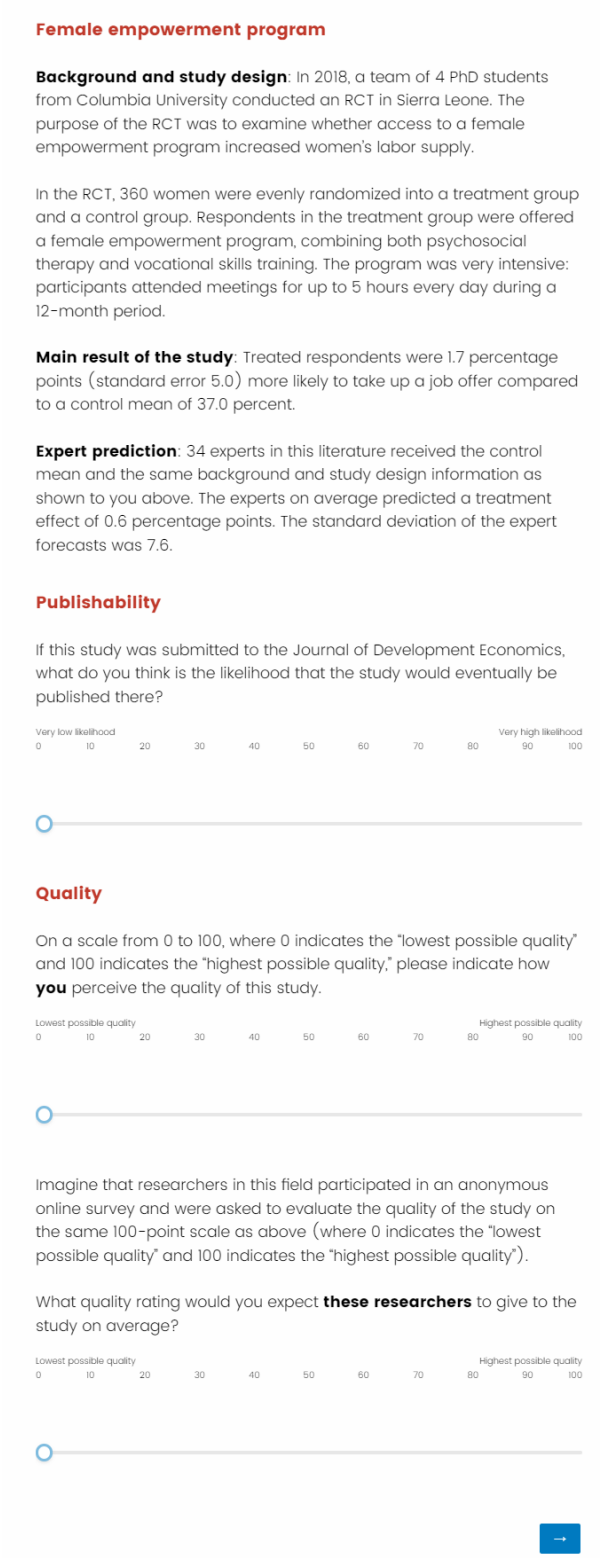
D.1.2 Marginal effects of merit aid for low-income students



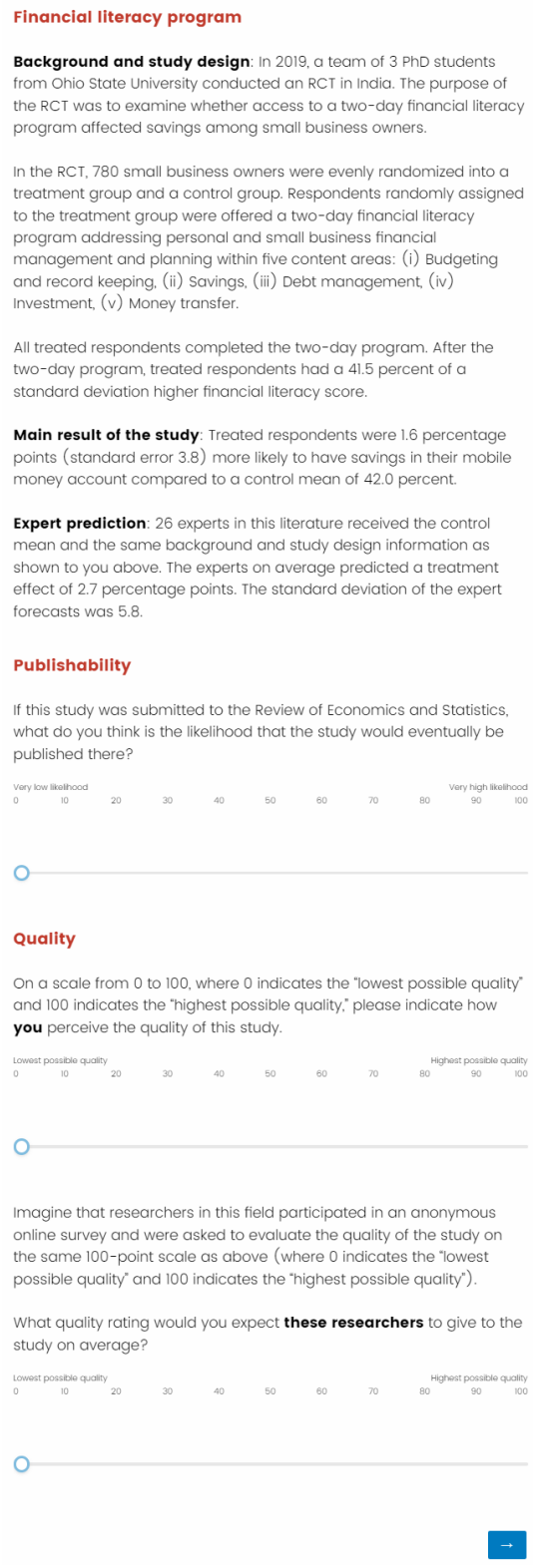
D.1.3 Long-term effects of equal land sharing



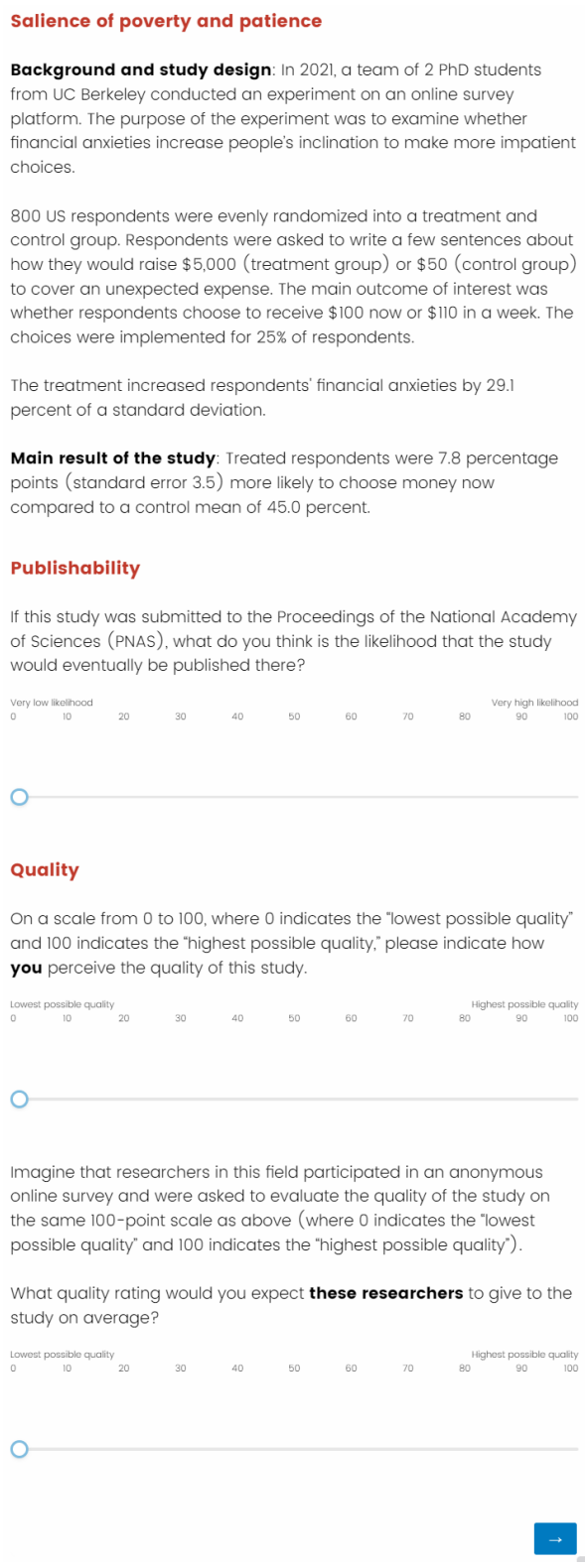
D.1.4 Female empowerment program



D.1.5 Financial literacy program



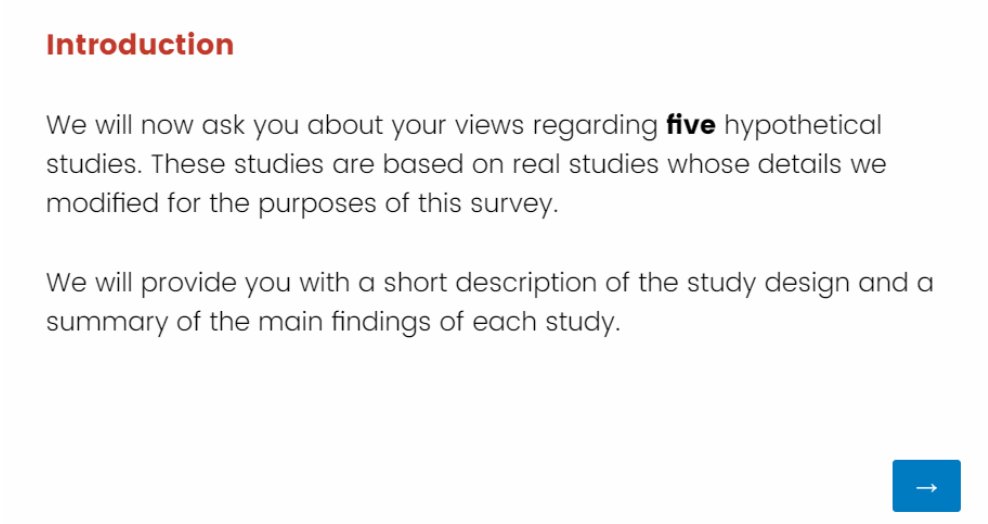
D.1.6 Salience of poverty and patience



D.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).

D.2.1 Pre-treatment information



D.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.

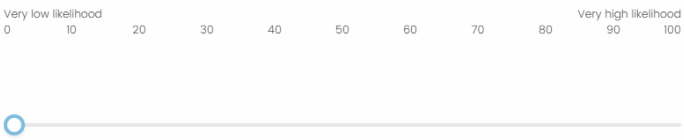
In the RCT, 360 women were evenly randomized into a treatment group and a control group. Respondents in the treatment group were offered a female empowerment program, combining both psychosocial therapy and vocational skills training. The program was very intensive: participants attended meetings for 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

→

E Pre-analysis plans

The data collections were pre-registered in the AsPredicted registry (#95235 and #96599). The pre-analysis plans for the main experiment and the mechanism experiment are available on the following links: `https://aspredicted.org/su6dj.pdf` and `https://aspredicted.org/83i25.pdf`.

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.