For online publication only:

The Null Result Penalty

Felix Chopra, Ingar Haaland, Christopher Roth and Andreas Stegmann

Section A contains additional figures and tables.

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

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

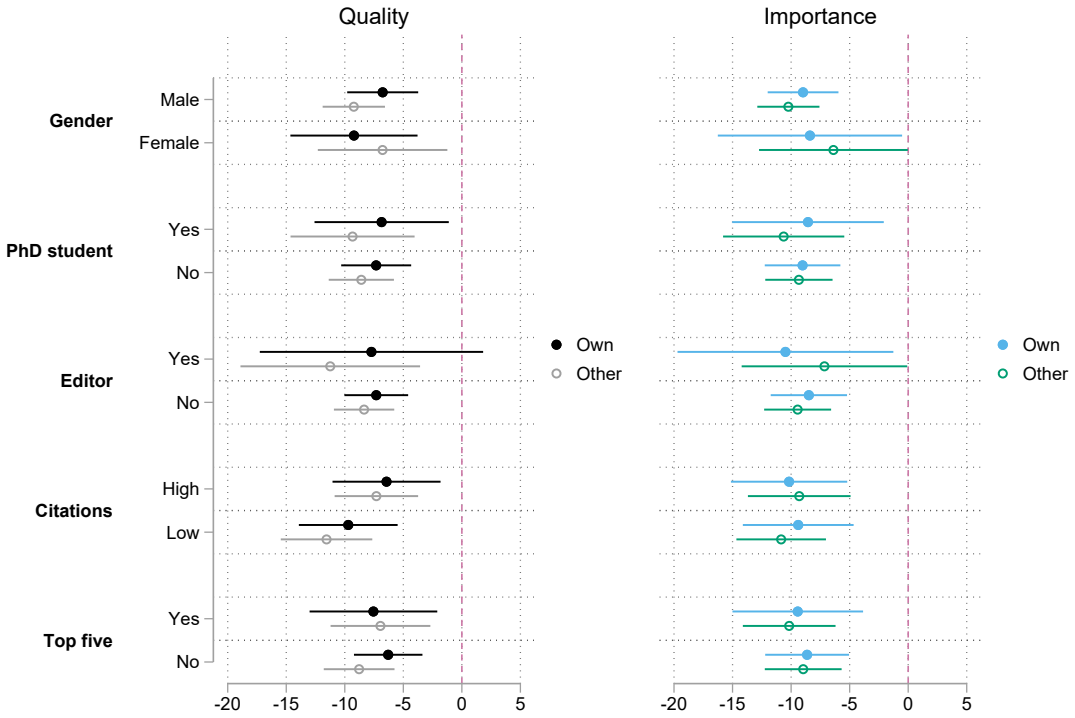
Section D provides a discussion of learning about quality from study results in the presence of expert forecasts.

Section E provides screenshots of the experimental instructions.

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

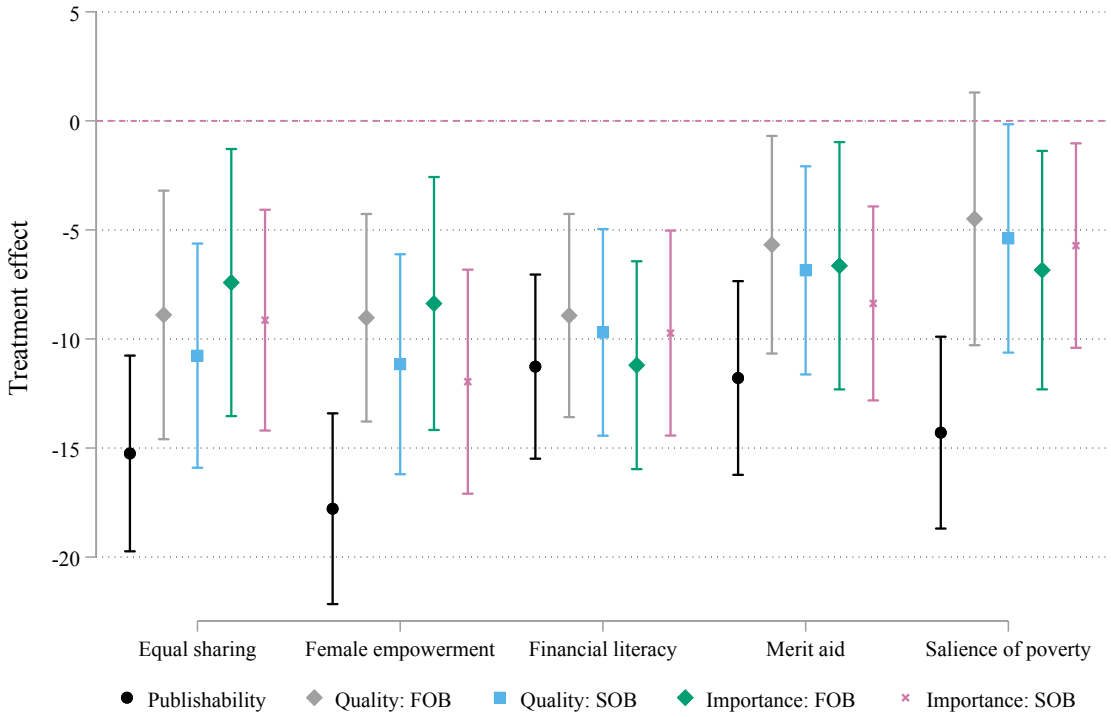
A Additional figures and tables

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



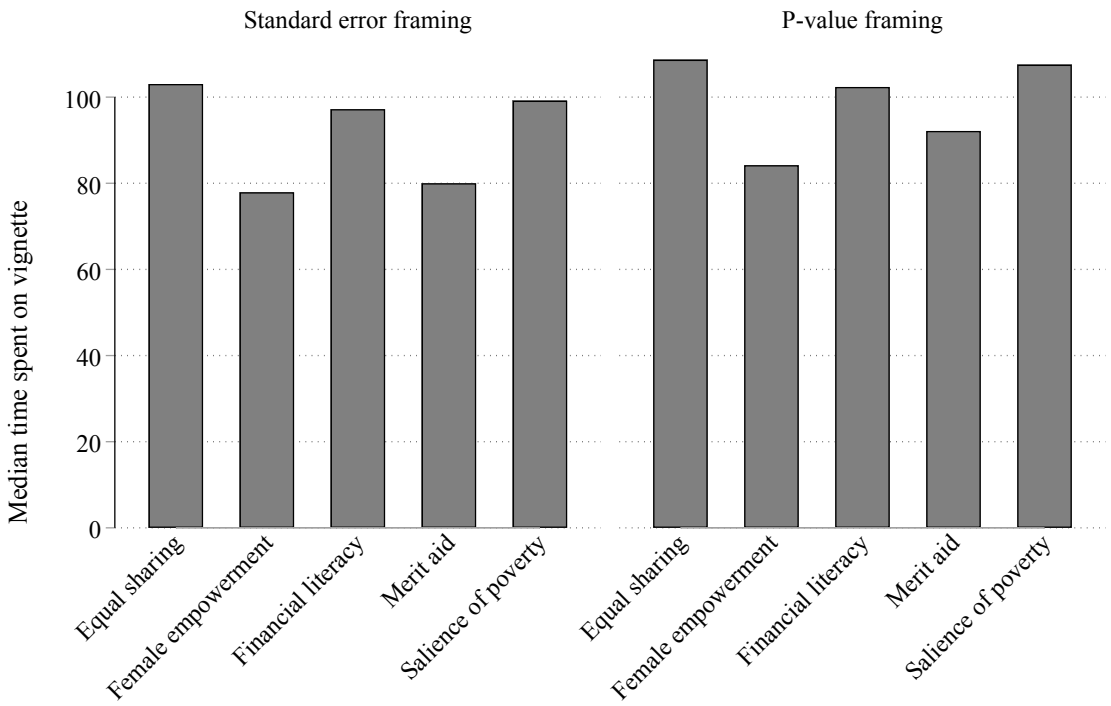
Note: This figure shows regression estimates in which first-order ("own") and second-order beliefs ("other") about the importance and quality of the study (both measured on a scale from 0 to 100) are regressed on the "null result treatment" indicator, separately for each sub-group indicated in the figure. Citations are measured using Google Scholar data as of May 2022 and "low" and 'high" refer to, respectively, below or above median citations in our sample. "Editor" refers to whether the respondent ever has been an editor of a scientific journal. "Top five" refers to whether the respondent has published a paper in any of the "top 5" economics journals. All regressions include controls for the other cross-randomized features as well as respondent fixed effects. Standard errors are clustered at the respondent level. 95% confidence intervals are indicated in the figure.

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



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

Figure A.3: Median response times across vignettes



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

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

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

Note: This table shows regression estimates of our treatment effects on publishability (beliefs in %). The data set is at the vignette-respondent level and contains four observations for each respondent. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. "P-value framing" is a treatment indicator taking the value one if the vignette treatment effect had an associated *p*-value and zero if it had an associated standard error estimate. *p*-values are shown in square brackets. "Respondents with null variation" is the number of respondents who were presented with at least one vignette that included a null result and at least one vignette with a statistically significant result.

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

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

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Null result treatment	-15.735***	-0.325**	-0.404***	-0.264*	-0.369***
	(2.232)	(0.131)	(0.142)	(0.134)	(0.131)
Observations	502	260	260	242	242
Respondents	345	174	174	171	171

Note: This table shows regression estimates of our treatment effects on our key outcomes of interest. The data set is at the vignette-respondent level and includes only observations where the vignette describes a research team consisting of professors from higher-ranked universities. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. We include respondent and vignette fixed effects in all regressions and control for all other cross-randomized vignette features.

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

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

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Panel A: Baseline with FEs					
Null result treatment	-14.058***	-0.373***	-0.460***	-0.325***	-0.417***
	(1.090)	(0.062)	(0.062)	(0.054)	(0.056)
Panel B: Baseline with FEs, re-weighted					
Null result treatment	-14.131***	-0.299***	-0.394***	-0.368***	-0.435***
	(1.286)	(0.069)	(0.079)	(0.063)	(0.062)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250
Panel C: OLS					
Null result treatment	-14.474***	-0.401***	-0.455***	-0.305***	-0.367***
	(1.224)	(0.069)	(0.072)	(0.062)	(0.069)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250
Panel D: OLS, re-weighted					
Null result treatment	-14.628***	-0.360***	-0.417***	-0.300***	-0.362***
	(1.471)	(0.078)	(0.088)	(0.077)	(0.084)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250

Note: The table shows regression estimates of our treatment effects on our key outcomes of interest. The data set is at the vignette-respondent level and contains four observations for each respondent. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. The regressions in Panel A (Panel B) include (do not include) respondent fixed effects. All regressions in both panels include treatment indicators for the cross-randomized conditions in addition to vignette fixed effects. Panel A and C replicate the baseline results from Table 3. Panel B and D show analogous estimates when reweighing respondents to match the sampling population along several dimensions (gender, region, editorial position, top 5 referee). For details on the construction of weights, see Section B.3.

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

Table A.4: Descriptive statistics: Reweighted for representativeness

	Survey sample		Sampling population
	Original mean	Reweighted mean	Mean
Demographics:			
Female*	0.220	0.238*	0.236
Years since PhD	14.805	15.612	16.091
PhD student	0.244	0.310	
Region of institution:			
Europe*	0.544	0.365*	0.362
North America*	0.406	0.524*	0.527
Australia*	0.033	0.078*	0.078
Asia*	0.017	0.033*	0.033
Academic output:			
H-index	17.22	16.905	8.831
Citations	4,348.34	4,579.753	
Number of top 5 publications	1.27	1.184	0.342
Number of top 5s refereed for	1.166	0.662	
Repeated top 5 referee*	0.305	0.157*	0.125
Research evaluation:			
Current editor*	0.072	0.035*	0.032
Current associate editor	0.127	0.105	
Ever editor	0.151	0.118	
Ever associate editor	0.193	0.161	
Professional memberships:			
NBER affiliate	0.084	0.075	
CEPR affiliate	0.171	0.106	
Academic fields:			
Labor	0.211	0.207	
Public	0.129	0.113	
Development	0.179	0.151	
Political	0.167	0.164	
Finance	0.105	0.102	
Experimental	0.062	0.066	
Behavioral	0.091	0.086	
Theory	0.067	0.061	
Macro	0.141	0.131	
Econometrics	0.141	0.139	

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

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

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Null result treatment	-16.242***	-0.295**	-0.396***	-0.282**	-0.291**
	(2.133)	(0.125)	(0.129)	(0.120)	(0.125)
Observations	480	230	230	250	250
Respondents	480	230	230	250	250
Controls	Yes	Yes	Yes	Yes	Yes

Note: The table shows OLS regression estimates of our treatment effects on our key outcomes of interest using only the first vignette the respondents were randomly assigned to. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. We include treatment indicators for the other cross-randomized conditions in addition to vignette fixed effects in all regressions.

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

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

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Panel A: Baseline					
Null result treatment	-14.058***	-0.373***	-0.460***	-0.325***	-0.417***
	(1.090)	(0.062)	(0.062)	(0.054)	(0.056)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250
Panel B: High power					
Null result treatment	-11.486***	-0.286***	-0.338***	-0.362***	-0.370***
	(1.569)	(0.085)	(0.087)	(0.070)	(0.075)
Observations	1,156	543	543	613	613
Respondents	480	230	230	250	250
Panel C: Low power					
Null result treatment	-15.336***	-0.487***	-0.535***	-0.389***	-0.472***
	(2.270)	(0.144)	(0.143)	(0.127)	(0.133)
Observations	568	294	294	274	274
Respondents	284	147	147	137	137
Panel D: Empirical micro sample					
Null result treatment	-13.417***	-0.246***	-0.308***	-0.378***	-0.425***
	(1.897)	(0.089)	(0.097)	(0.092)	(0.092)
Observations	837	420	420	417	417
Respondents	348	176	176	172	172

Note: This table shows regression estimates of our treatment effects on our key outcomes of interest. The data set is at the vignette-respondent level and contains four observations for each respondent. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. Panel A reports the baseline treatment effects from our main specification. Panel B focuses on the three vignettes that have a power of at least 80% to detect a treatment effect of 20% of a standard deviation (see Table 2), while Panel C uses only observations from vignettes that had a comparatively lower statistical power. Panel D restricts the sample to high-powered vignettes and researchers that are more likely to have experience with empirical microeconomic research by excluding researchers in the field of macro, finance, international economics, and theory. We include respondent and vignette fixed effects in all regressions and control for all other cross-randomized features.

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

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

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Panel A: Expert forecast					
Null result	-11.239***	-0.281**	-0.506***	-0.351***	-0.432***
	(1.913)	(0.113)	(0.112)	(0.094)	(0.085)
	[0.001]***	[0.022]**	[0.001]***	[0.001]***	[0.001]***
Null result x Low expert forecast	-2.002	-0.168	0.128	0.032	0.065
	(2.478)	(0.161)	(0.161)	(0.120)	(0.116)
	[0.993]	[0.939]	[0.993]	[1.000]	[1.000]
Null result x High expert forecast	-6.383**	-0.104	0.009	0.048	-0.020
	(2.646)	(0.166)	(0.154)	(0.124)	(0.126)
	[0.029]**	[1.000]	[1.000]	[1.000]	[1.000]
Low expert forecast	-0.890	0.190*	-0.015	-0.076	-0.042
	(1.671)	(0.108)	(0.108)	(0.092)	(0.081)
High expert forecast	1.959	0.115	0.121	-0.049	-0.018
	(1.800)	(0.112)	(0.099)	(0.085)	(0.088)
Panel B: Field journal					
Null result	-14.571***	-0.366***	-0.446***	-0.343***	-0.418***
	(1.465)	(0.093)	(0.086)	(0.072)	(0.075)
	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***
Null result x Field journal	1.025	-0.014	-0.027	0.036	0.003
	(1.965)	(0.129)	(0.122)	(0.101)	(0.103)
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Field journal	12.218***	0.141	0.108	0.108	0.101
	(1.397)	(0.095)	(0.089)	(0.072)	(0.069)
Panel C: PhD student					
Null result	-14.945***	-0.291***	-0.358***	-0.300***	-0.362***
	(1.491)	(0.085)	(0.082)	(0.081)	(0.081)
	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***
Null result x PhD student	1.745	-0.166	-0.206*	-0.047	-0.104
	(2.049)	(0.117)	(0.107)	(0.102)	(0.097)
	[0.991]	[0.670]	[0.176]	[1.000]	[0.937]
PhD student	-4.543***	-0.025	-0.042	0.066	0.019
	(1.403)	(0.091)	(0.081)	(0.071)	(0.069)
Panel D: Lower-ranked university					
Null result	-14.320***	-0.381***	-0.474***	-0.317***	-0.408***
	(1.480)	(0.094)	(0.093)	(0.073)	(0.076)
	[0.001]***	[0.001]***	[0.001]***	[0.001]***	[0.001]***
Null result x Lower-ranked university	0.518	0.017	0.030	-0.014	-0.017
	(1.985)	(0.121)	(0.124)	(0.108)	(0.105)
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Low-ranked university	-3.998***	-0.093	-0.230***	0.007	-0.046
	(1.371)	(0.082)	(0.077)	(0.077)	(0.072)
Panel E: P-value framing					
Null result	-11.960***	-0.243**	-0.302***	-0.366***	-0.405***
	(1.736)	(0.095)	(0.101)	(0.081)	(0.095)
	[0.001]***	[0.017]**	[0.004]***	[0.001]***	[0.001]***
Null result x P-value framing	-5.214**	-0.296**	-0.286**	0.140	0.088
	(2.430)	(0.136)	(0.141)	(0.124)	(0.135)
	[0.080]*	[0.068]*	[0.113]	[0.917]	[1.000]
P-value framing	-2.824	0.022	-0.032	-0.066	-0.104
	(2.091)	(0.114)	(0.118)	(0.114)	(0.120)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250

Note: This table presents estimates that are exactly analogous to Table 4. Im addition, Romano and Wolf (2005) p-values adjusted for multiple hypothesis testing are shown in square brackets with corresponding significance stars for the key coefficients of interest (Clarke, 2021), excluding interactants.

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

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

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Main treatment:					
Null result	-11.072***	-0.029	-0.219	-0.330**	-0.390***
	(2.681)	(0.151)	(0.160)	(0.132)	(0.135)
	[0.001]***	[1.000]	[0.755]	[0.026]**	[0.006]***
Interaction effects:					
Null result x Low expert forecast	-1.862	-0.169	0.130	0.030	0.058
	(2.470)	(0.162)	(0.159)	(0.120)	(0.117)
	[0.997]	[0.958]	[0.994]	[1.000]	[1.000]
Null result x High expert forecast	-6.251**	-0.083	0.033	0.048	-0.025
	(2.632)	(0.165)	(0.152)	(0.124)	(0.127)
	[0.042]**	[1.000]	[1.000]	[1.000]	[1.000]
Null result x Field journal	0.871	0.003	-0.027	0.038	0.006
	(1.966)	(0.131)	(0.121)	(0.101)	(0.103)
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Null result x PhD student	1.707	-0.165	-0.196*	-0.047	-0.101
	(2.054)	(0.121)	(0.108)	(0.102)	(0.098)
	[0.994]	[0.755]	[0.305]	[1.000]	[0.960]
Null result x Low-ranked university	0.408	0.021	0.028	-0.011	-0.018
	(1.965)	(0.121)	(0.124)	(0.108)	(0.106)
	[1.000]	[1.000]	[1.000]	[1.000]	[1.000]
Null result x P-value framing	-3.652*	-0.344***	-0.362***	-0.021	0.049
	(2.164)	(0.122)	(0.120)	(0.109)	(0.112)
	[0.407]	[0.006]***	[0.004]***	[1.000]	[1.000]
Interactants:					
Low expert forecast	-0.876	0.200*	-0.007	-0.076	-0.041
	(1.666)	(0.108)	(0.107)	(0.092)	(0.081)
High expert forecast	1.977	0.108	0.110	-0.049	-0.019
	(1.789)	(0.113)	(0.096)	(0.085)	(0.088)
Field journal	12.204***	0.120	0.097	0.108	0.100
	(1.396)	(0.097)	(0.090)	(0.073)	(0.069)
PhD student	-4.600***	-0.036	-0.056	0.068	0.016
	(1.407)	(0.094)	(0.082)	(0.071)	(0.069)
Low-ranked university	-3.986***	-0.105	-0.235***	0.006	-0.046
	(1.363)	(0.081)	(0.076)	(0.077)	(0.073)
N	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250

Note: This table shows regression estimates of our treatment effects on our key outcomes of interest from a specification that includes the full interactions between the null treatment indicator and indicators for all cross-randomized features. The data set is at the vignette-respondent level and contains four observations for each respondent. We include individual and vignette fixed effects in all regressions. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. "Low expert forecast" and "High expert forecast" are treatment indicators taking the value one if the group of experts predicted, respectively, a low or high treatment effect estimate (and zero otherwise). "Field journal" is a treatment indicator taking the value one if the vignette included a field journal and zero if it included a general interest journal. "PhD student" is a treatment indicator taking the value one if the team behind the vignette research study included PhD students and zero if it included professors. "Low-ranked university" is a treatment indicator taking the value one if the team behind the vignette research study was affiliated with a lower-ranked university and zero if it was affiliated with a higher-ranked university. "P-value framing" is a treatment indicator taking the value one if the vignette treatment effect had an associated p-value and zero if it had an associated standard error estimate. Romano and Wolf (2005) p-values adjusted for multiple hypothesis testing are shown in square brackets with corresponding significance stars for the key coefficients of interest (Clarke, 2021), excluding interactants.

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

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

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Panel A					
Null result treatment	-13.202***	-0.389***	-0.454***	-0.329***	-0.430***
	(1.253)	(0.072)	(0.074)	(0.060)	(0.068)
Matching field	0.883	0.041	0.053	0.185*	0.025
	(1.922)	(0.118)	(0.114)	(0.105)	(0.102)
Null result treatment x Matching field	-2.764	0.042	-0.019	0.018	0.048
	(2.192)	(0.130)	(0.125)	(0.121)	(0.114)
Observations	1,920	920	920	1,000	1,000
Respondents	480	230	230	250	250
Panel B: Non-matching fields					
Null result treatment	-13.206***	-0.366***	-0.493***	-0.332***	-0.510***
	(1.457)	(0.080)	(0.082)	(0.070)	(0.081)
Observations	988	468	468	520	520
Respondents	247	117	117	130	130
Panel C: Matching fields					
Null result treatment	-18.067***	-0.414***	-0.493***	-0.395***	-0.416***
	(2.117)	(0.119)	(0.114)	(0.126)	(0.112)
Observations	566	307	307	259	259
Respondents	183	98	98	85	85

Note: This table shows regression estimates of our treatment effects on our key outcomes of interest. The data set is at the vignette-respondent level. "Matching field" is a binary indicator taking value one if the vignette presented to the respondent belongs to a research field that overlaps with the respondent's fields of specialization, and zero otherwise. Panel A includes all observations, while Panel B includes respondents that are specialized in research fields that are unrelated to the studies they encountered in the survey. Panel C uses only observations where respondents are familiar with the field of specialization that the study presented in the hypothetical vignette belongs to. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. We include vignette fixed effects and respondent fixed effects in all regressions and control for all other cross-randomized vignette features.

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

Table A.10: Descriptive statistics on time use

	count	p25	p50	p75	min	max	mean	sd
Duration in seconds	480	339	450.5	694	78.0	93322.0	1306.8	7218.4
Duration (winsorized)	480	339	450.5	694	78.0	1333.0	552.2	300.7
Page time (in seconds)	1920	65.49	93.648	140.1	7.7	92747.4	177.3	2120.5
Page time (winsorized)	1920	65.49	93.648	140.1	7.7	301.5	113.6	69.1

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

Table A.11: Time spent on vignettes

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

Note: The table shows OLS regression estimates of our treatment effects on the time that respondents spent on a vignette in the main experiment. The data is at the respondent-vignette level. The dependent variables in columns 3 and 4 are winsorized at the 5^{th} and 95^{th} percentile. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. We include treatment indicators for the other cross-randomized conditions in addition to vignette fixed effects in all regressions. The omitted category for the vignettes is the "Female empowerment" vignette with a 116 word description of the study. For each vignette, the word count difference relative to the female empowerment vignette is indicated.

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

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

	Publishability	Quality (z-scored)		Importance (z-scored)	
	(1) Beliefs in percent	(2) First-order beliefs	(3) Second-order beliefs	(4) First-order beliefs	(5) Second-order beliefs
Panel A: 4+ minutes					
Null result treatment	-14.685***	-0.420***	-0.498***	-0.344***	-0.419***
	(1.126)	(0.063)	(0.062)	(0.056)	(0.057)
Observations	1,788	864	864	924	924
Respondents	447	216	216	231	231
Panel B: 6+ minutes					
Null result treatment	-15.518***	-0.407***	-0.502***	-0.356***	-0.460***
	(1.320)	(0.071)	(0.072)	(0.064)	(0.066)
Observations	1,360	656	656	704	704
Respondents	340	164	164	176	176
Panel C: 8+ minutes					
Null result treatment	-16.654***	-0.482***	-0.556***	-0.310***	-0.413***
	(1.575)	(0.086)	(0.085)	(0.075)	(0.079)
Observations	884	412	412	472	472
Respondents	221	103	103	118	118
Panel D: 10+ minutes					
Null result treatment	-16.986***	-0.581***	-0.569***	-0.285***	-0.449***
	(1.853)	(0.109)	(0.106)	(0.085)	(0.089)
Observations	640	284	284	356	356
Respondents	160	71	71	89	89

Note: This table shows regression estimates of our treatment effects on our key outcomes of interest. The data set is at the vignette-respondent level and each panel includes only observations where the respondent spent at least k minutes on the overall survey. The panel header indicates the value of k. "Null result treatment" is a treatment indicator taking the value one if the vignette included a low treatment effect estimate (a null result) and zero if it included a high treatment effect estimate. We include vignette fixed effects and respondent fixed effects in all regressions and control for all other cross-randomized vignette features.

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

B Expert characteristics

B.1 Survey sample

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

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

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

B.2 Sampling population

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

- *Female*: Binary indicator for female experts, derived from their first and last name using the Gender API algorithm.
- *Year of first publication*: This is the number of years since the first publication, as recorded in the Scopus database.
- *Region of institution*: Regional indicators are based on the country of the institution a researcher is affiliated with. The information is obtained from the Scopus database.
- *H-index*: The Scopus h-index, which is derived from Scopus data on the citations of all publications of an author as of December 2019.

- *Number of top 5 publications*: This variable is the number of publications in five highly cited general-interest economics journals (the American Economic Review, the Quarterly Journal of Economics, the Journal of Political Economy, Econometrica, and the Review of Economic Studies).

- *Repeated top 5 referee*: Binary indicator for whether the researcher has repeatedly refereed for the five highly cited general-interest economics journals (the American Economic Review, the Quarterly Journal of Economics, the Journal of Political Economy, Econometrica, and the Review of Economic Studies) in the years from 2015-2020. This variable is derived from the list of referees that have refereed for these five journals. These lists are published on an annual basis. For more details, see Appendix A.4 of Andre and Falk (2021).

- *Current editor*: Binary indicator for having been an editor at one of the top 100 journals in economics at some point in the years from 2015-2020. For more details, see Appendix A.4 of Andre and Falk (2021).

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

B.3 Weights

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

- Gender (two groups)

- Region of institution (four groups)

- Repeated top 5 referee (two groups)

- Current editor (two groups)

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

C Numerical features used in the vignettes

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

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

- Standard error: We conducted a simulation exercise to obtain an estimate of the standard error that one would obtain based on the number of observations, the control group mean, the assignment to treatment and control groups, and the main effect from the original studies on which we based our vignettes. This ensures that our reported standard errors and p-values are internally consistent with the description of the sample and the empirical strategy.

- Number of experts: This is an integer drawn uniformly from the interval $[20, 35]$.

- Standard deviation of the expert prior: We multiplied the standard error (see above) with a number drawn uniformly from the interval $[1, 2]$. This ensures that a Bayesian with the experts' prior should put a weight of at least 0.5 on the study's findings when updating his belief about the underlying "true" effect. This implies that the study findings are informative relative to the experts' prior, irrespective of whether the main effect is statistically significant or not.

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

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

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

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

- p-values (high and low): The p-values are obtained from the hypothetical t-statistic used to generate the statistically (non-)significant main effect.

- Expert prior (high mean): This number is equal to $\mu_{\text{high}} + 0.25x$. Here, μ_{high} is the statistically significant main effect and $x \sim N(0, S)$ where S is the standard error (see above).

- Expert prior (low mean): This number is equal to the high expert prior minus the absolute difference between the statistically significant and the statistically non-significant main effect. This ensures that the absolute difference between the high and low expert mean is equal to the absolute difference between the statistically significant and non-significant main effects.

D Updating about quality

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

D.1 Discussion

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

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

Empirically, we find that respondents that do not receive an expert forecast update negatively about the quality of research studies with null results (column 2 of Table 3). This would be consistent with the mechanism outlined above if respondents were sufficiently certain that there is a true causal effect. Next, we exploit the exogenous variation in prior beliefs induced by the high vs low expert forecasts. Note that the expert

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

forecasts affect respondents' assessment of the publishability of null results (column 1 of Table 4, Panel A), suggesting that participants—at least to a certain degree—internalize the information about the expert predictions. Column 2 in Panel A of Table 4 shows that respondents update negatively about the quality of research studies independent of whether they were informed that experts predict a large or a small effect. If anything, participants update somewhat more negatively about the quality of research studies with unsurprising nulls (relative to the experts' prior) compared to surprising nulls. While these effects are somewhat noisily estimated, they are directionally at odds with rational inference about the quality of a study.

D.2 Formalization

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

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

We will examine how a Bayesian researcher will rationally revise his belief about the unobserved quality θ of a study depending on whether it yielded a statistically significant main effect or not. The thought experiment we are interested in is one where there are two studies, A and B, with the same observable characteristics θ_{obs}, but study A yielded a null result while study B yielded a statistically significant result. This thought experiment mirrors our experimental design, where we fix observable characteristics such as the sample size, and then exogenously vary the statistical significance of the main finding. How will a Bayesian observer revise his beliefs about the quality of study A compared to study B? In the following, we suppress θ_{obs} from the notation to simplify the exposition as we are only interested in the observers' inferences about the unobserved study features.

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

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

$$\pi^0(\theta) = P(s = 0 \mid \omega = 0, \theta) \tag{D.2}$$

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

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

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

PROPOSITION 1. *Let $\hat{\rho}(s)$ denote the posterior probability of a researcher with prior ρ that a study is of high-quality after observing the results $s \in \{0, 1\}$ of a study. Then*

$$\hat{\rho}(s = 0) \leq \rho \leq \hat{\rho}(s = 1) \tag{D.3}$$

if the following inequality holds

$$\frac{(\pi_H^0 - \pi_L^0)}{(\pi_H^0 - \pi_L^0) + (\pi_H^1 - \pi_L^1)} \leq p \tag{D.4}$$

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

$$\frac{\partial \hat{\rho}(s=0)}{\partial p} < 0 < \frac{\partial \hat{\rho}(s=1)}{\partial p} \tag{D.5}$$

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

$$\hat{\rho}(0) = \frac{P(H)P(s=0|H)}{P(L)P(s=0|L) + P(H)P(s=0|H)} \tag{D.6}$$

$$= \frac{P(H)\left(P(s=0|\omega=0,H)P(\omega=0) + P(s=0|\omega=1,H)P(\omega=1)\right)}{\left(\begin{array}{c} P(H)\left(P(s=0|\omega=0,H)P(\omega=0) + P(s=0|\omega=1,H)P(\omega=1)\right) + \\ P(L)\left(P(s=0|\omega=0,L)P(\omega=0) + P(s=0|\omega=1,L)P(\omega=1)\right) \end{array}\right)} \tag{D.7}$$

$$= \frac{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right)}{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) + (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right)} \tag{D.8}$$

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

$$\frac{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right)}{\rho\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) + (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right)} \leq \rho \tag{D.9}$$

$$\iff (1-\rho)\left(\pi_H^0(1-p) + (1-\pi_H^1)p\right) \leq (1-\rho)\left(\pi_L^0(1-p) + (1-\pi_L^1)p\right) \tag{D.10}$$

$$\iff \pi_H^0(1-p) + (1-\pi_H^1)p \leq \pi_L^0(1-p) + (1-\pi_L^1)p \tag{D.11}$$

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

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

$$\frac{(\pi_H^0 - \pi_L^0)}{(\pi_H^0 - \pi_L^0) + (\pi_H^1 - \pi_L^1)} \leq p, \tag{D.13}$$

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

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

$$\hat{\rho}(1) = \frac{P(H)P(s=1|H)}{P(L)P(s=1|L)+P(H)P(s=1|H)} \tag{D.14}$$

$$= \frac{P(H)\left(P(s=1|\omega=0,H)P(\omega=0)+P(s=1|\omega=1,H)P(\omega=1)\right)}{\left(\begin{array}{c} P(H)\left(P(s=1|\omega=0,H)P(\omega=0)+P(s=1|\omega=1,H)P(\omega=1)\right)+ \\ P(L)\left(P(s=1|\omega=0,L)P(\omega=0)+P(s=1|\omega=1,L)P(\omega=1)\right)\end{array}\right)} \tag{D.15}$$

$$= \frac{\rho\left((1-\pi_H^0)(1-p)+\pi_H^1 p\right)}{\rho\left((1-\pi_H^0)(1-p)+\pi_H^1 p\right)+(1-\rho)\left((1-\pi_L^0)(1-p)+\pi_L^1 p\right)} \tag{D.16}$$

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

$$\frac{\rho\left(\pi_H^0(1-p)+(1-\pi_H^1)p\right)}{\rho\left(\pi_H^0(1-p)+(1-\pi_H^1)p\right)+(1-\rho)\left(\pi_L^0(1-p)+(1-\pi_L^1)p\right)} \geq \rho \tag{D.17}$$

$$\iff (1-\rho)\left(\pi_H^0(1-p)+(1-\pi_H^1)p\right) \geq (1-\rho)\left(\pi_L^0(1-p)+(1-\pi_L^1)p\right) \tag{D.18}$$

$$\iff \pi_H^0(1-p)+(1-\pi_H^1)p \geq \pi_L^0(1-p)+(1-\pi_L^1)p \tag{D.19}$$

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

The above condition is equivalent to equation D.12. The same argument as above then shows that $\rho \leq \hat{\rho}(1)$ if and only if equation D.13 holds.

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

$$f \equiv \rho\left(\pi_H^0(1-p)+(1-\pi_H^1)p\right) \tag{D.21}$$

$$g \equiv \rho\left(\pi_H^0(1-p)+(1-\pi_H^1)p\right)+(1-\rho)\left(\pi_L^0(1-p)+(1-\pi_L^1)p\right) \tag{D.22}$$

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

$$sgn\left(\frac{\partial\hat{\rho}(s=0)}{\partial p}\right) = sgn\left(\frac{\partial(f/g)}{\partial p}\right) = sgn\left(\frac{\partial f}{\partial p}g - f\frac{\partial g}{\partial p}\right) \tag{D.23}$$

$$=sgn(\rho(1-\pi_H^0-\pi_H^1)(1-\rho)(\pi_L^0(1-p)+(1-\pi_L^1)p)- \tag{D.24}$$

$$(1-\rho)(1-\pi_L^0-\pi_L^1)\rho(\pi_H^0(1-p)+p(1-\pi_H^1))) \tag{D.25}$$

$$= sgn\left((1-\pi_H^0-\pi_H^1)\pi_L^0-(1-\pi_L^0-\pi_L^1)\pi_H^0\right) = -1 \tag{D.26}$$

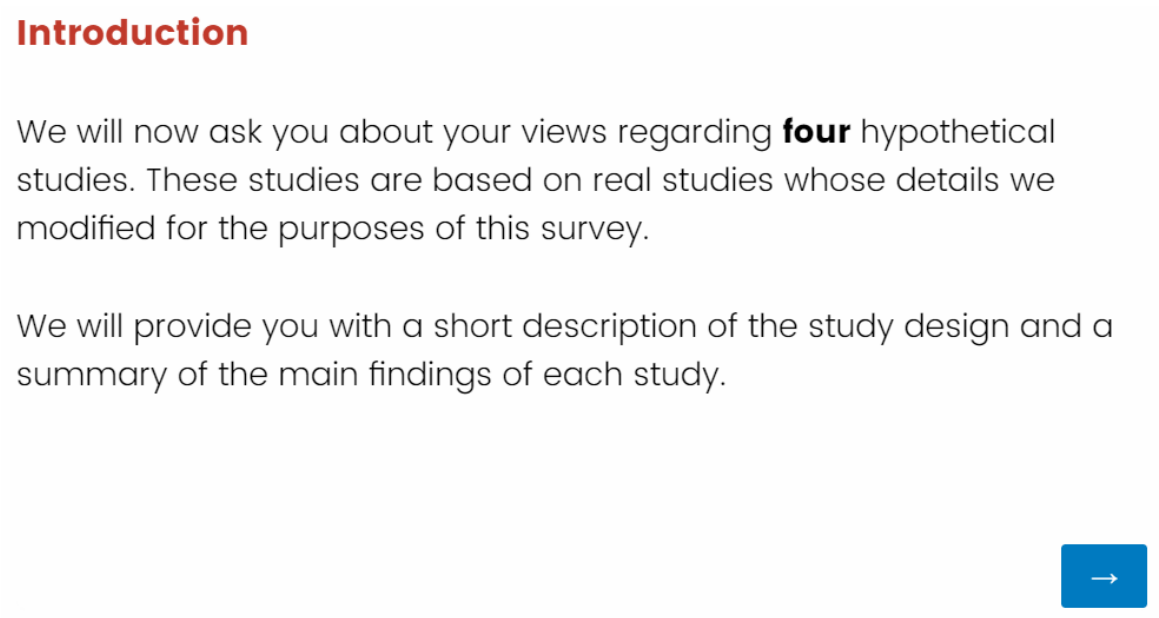
where the last equality follows from $\pi_L^0 \leq \pi_H^0$ and $\pi_H^0+\pi_H^1 > \pi_L^0+\pi_L^1$. An analogous calculation establishes that $sgn\left(\frac{\partial\hat{\rho}(s=1)}{\partial p}\right) = 1$, which concludes the proof. □

E Screenshots

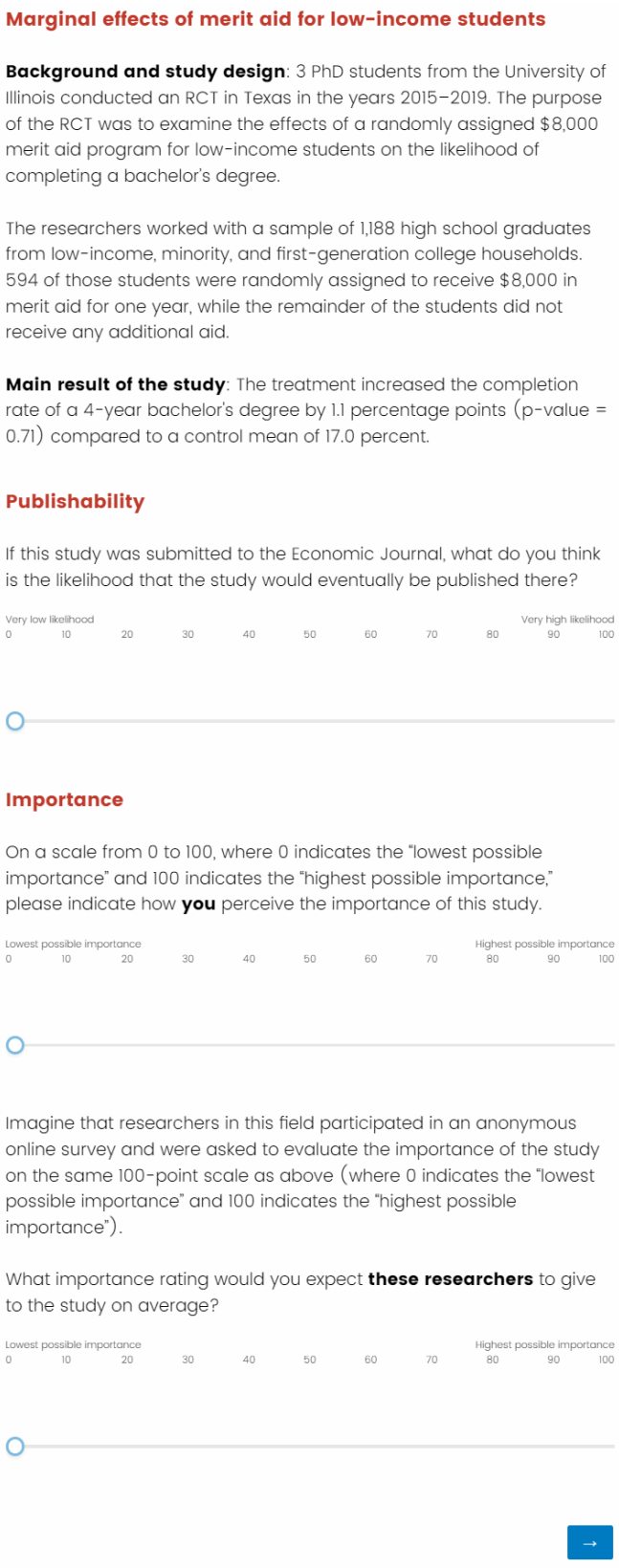
E.1 Main experiment

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

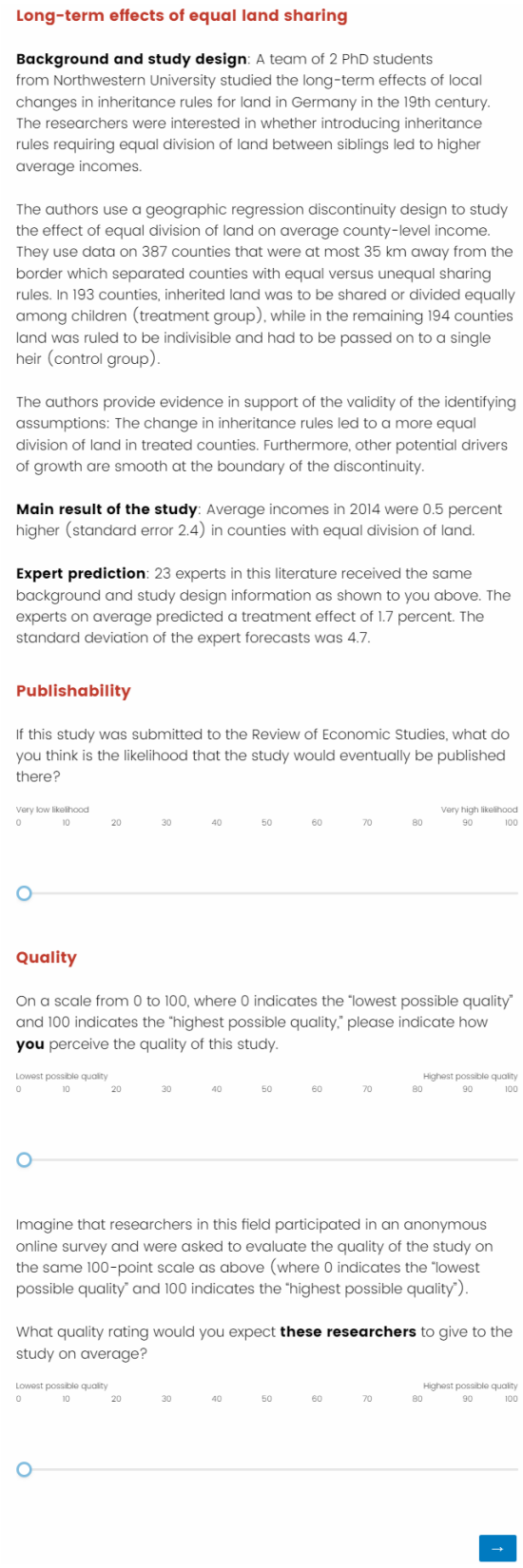
E.1.1 Pre-treatment information



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



E.1.3 Long-term effects of equal land sharing



Long-term effects of equal land sharing

Background and study design: A team of 2 PhD students from Northwestern University studied the long-term effects of local changes in inheritance rules for land in Germany in the 19th century. The researchers were interested in whether introducing inheritance rules requiring equal division of land between siblings led to higher average incomes.

The authors use a geographic regression discontinuity design to study the effect of equal division of land on average county-level income. They use data on 387 counties that were at most 35 km away from the border which separated counties with equal versus unequal sharing rules. In 193 counties, inherited land was to be shared or divided equally among children (treatment group), while in the remaining 194 counties land was ruled to be indivisible and had to be passed on to a single heir (control group).

The authors provide evidence in support of the validity of the identifying assumptions: The change in inheritance rules led to a more equal division of land in treated counties. Furthermore, other potential drivers of growth are smooth at the boundary of the discontinuity.

Main result of the study: Average incomes in 2014 were 0.5 percent higher (standard error 2.4) in counties with equal division of land.

Expert prediction: 23 experts in this literature received the same background and study design information as shown to you above. The experts on average predicted a treatment effect of 1.7 percent. The standard deviation of the expert forecasts was 4.7.

Publishability

If this study was submitted to the Review of Economic Studies, what do you think is the likelihood that the study would eventually be published there?

Very low likelihood — Very high likelihood
0 10 20 30 40 50 60 70 80 90 100

Quality

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

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

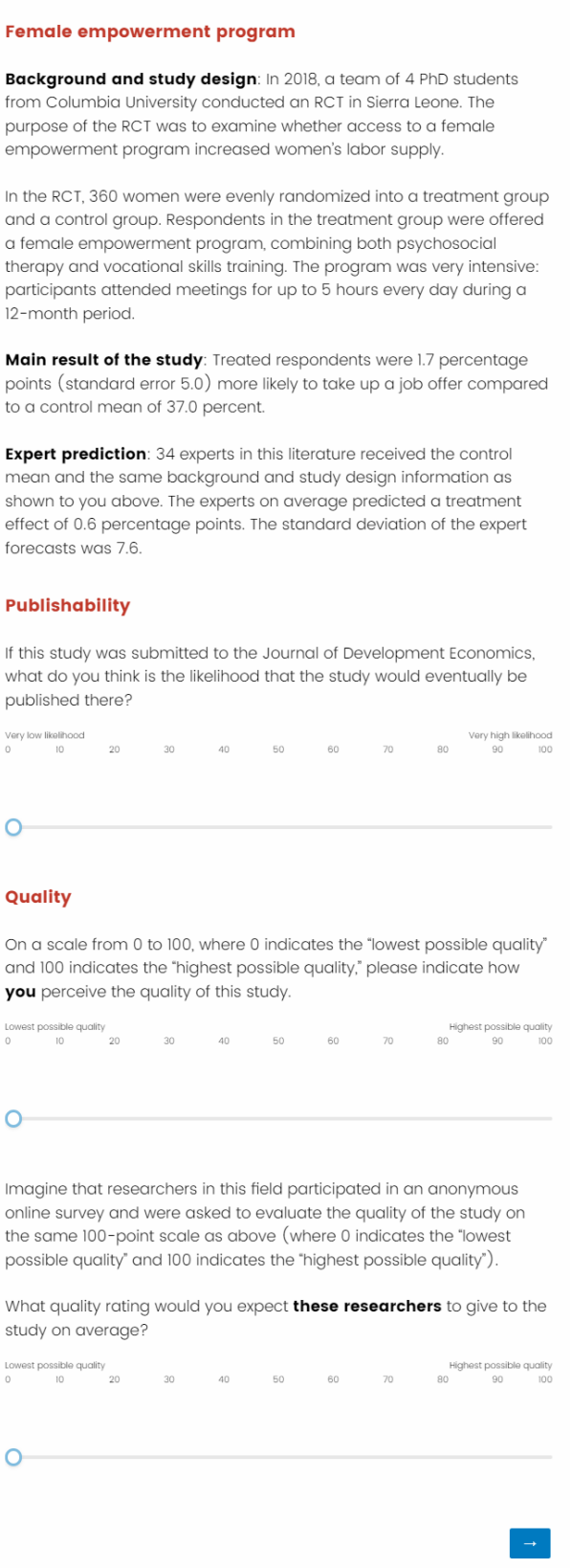
Imagine that researchers in this field participated in an anonymous online survey and were asked to evaluate the quality of the study on the same 100-point scale as above (where 0 indicates the "lowest possible quality" and 100 indicates the "highest possible quality").

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

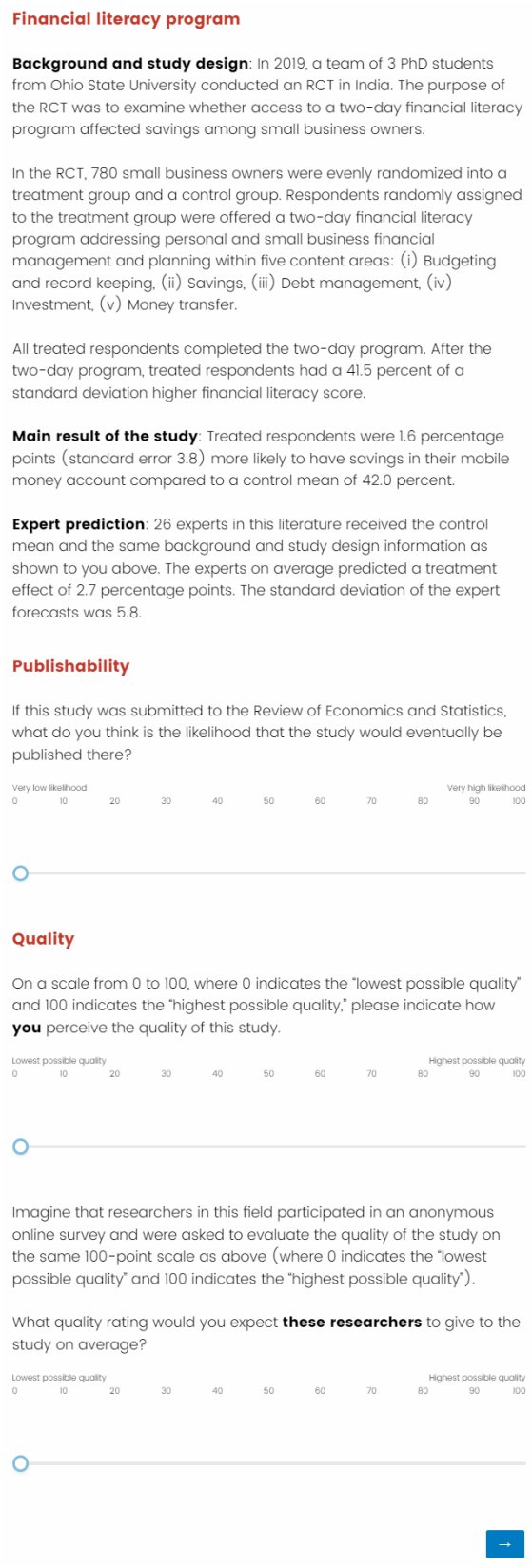
Lowest possible quality — Highest possible quality
0 10 20 30 40 50 60 70 80 90 100

→

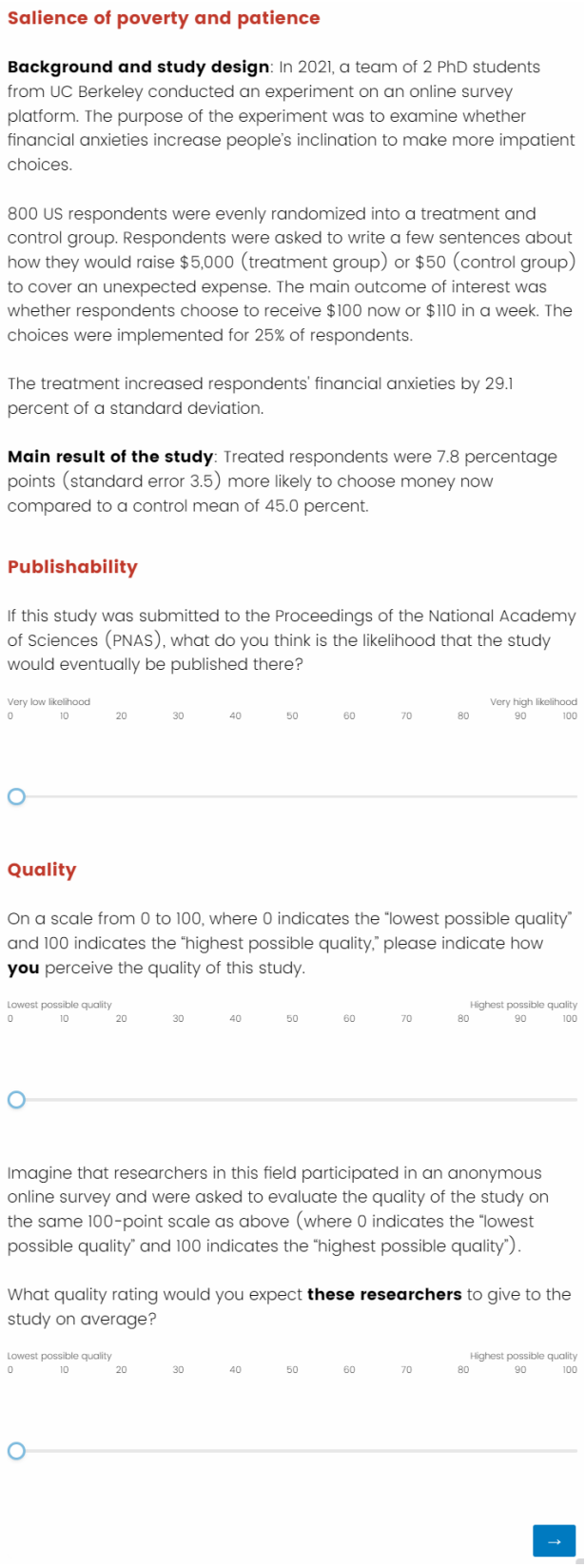
E.1.4 Female empowerment program



E.1.5 Financial literacy program



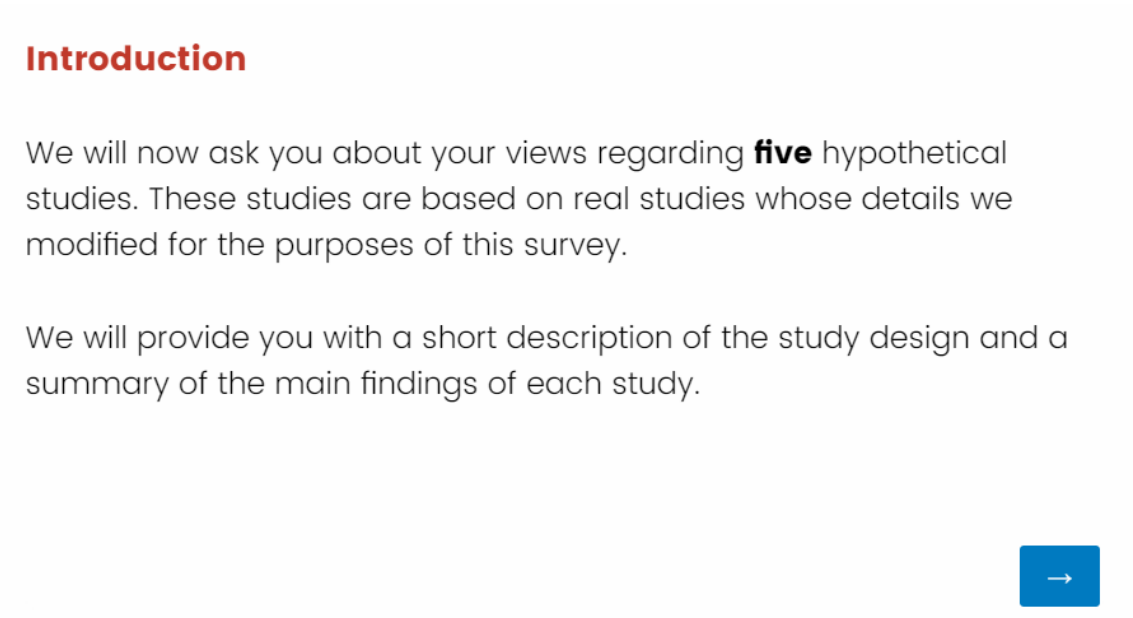
E.1.6 Salience of poverty and patience



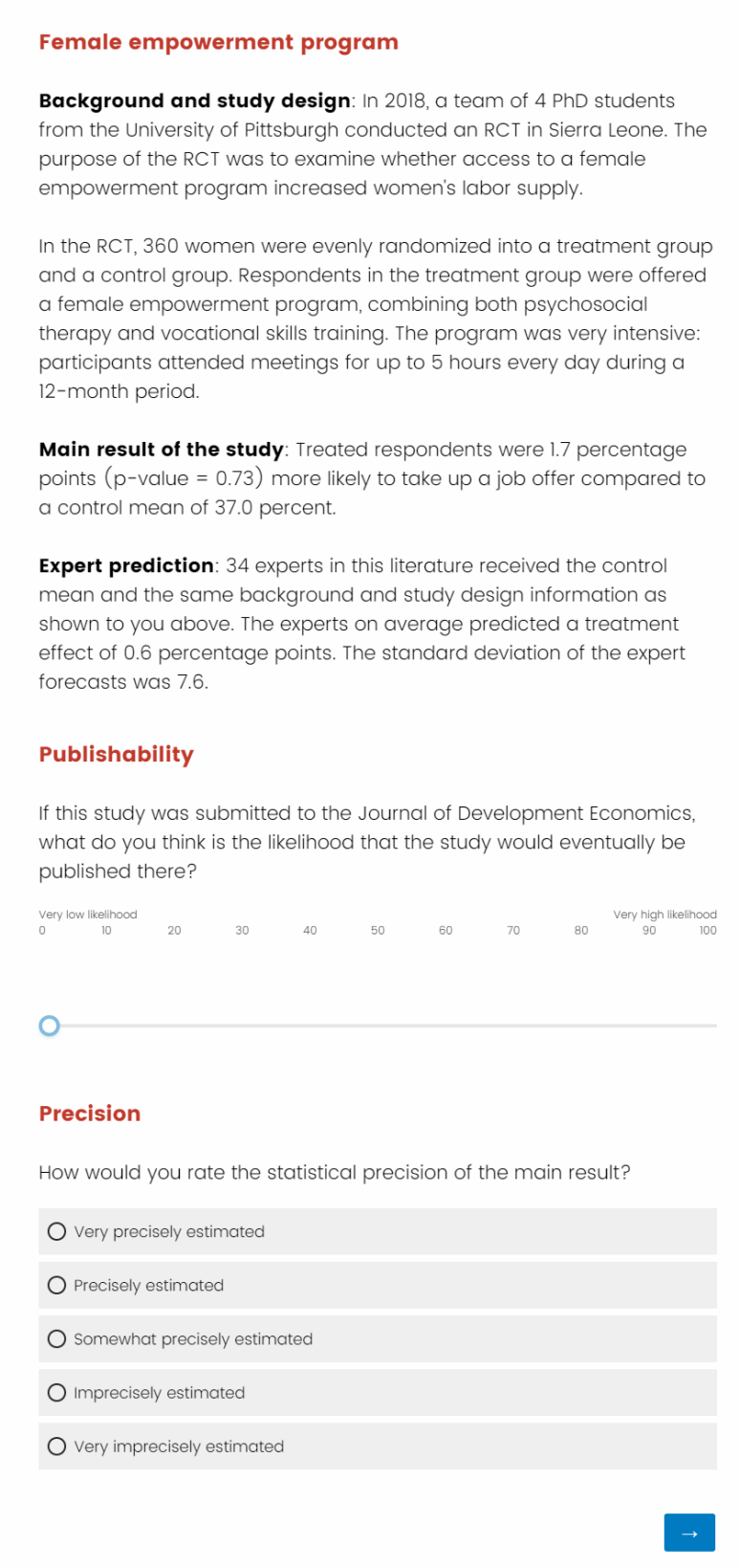
E.2 Mechanism experiment

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

E.2.1 Pre-treatment information



E.2.2 Female empowerment program



F Pre-analysis plans

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

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

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

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

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

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

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

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

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

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

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

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

In addition, we measure four secondary outcomes:

Half of our respondents receive the following two secondary outcomes:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In addition, we measure the following secondary outcome:

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

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

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

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

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

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

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

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

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

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

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

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

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