The robustness reproducibility of the American Economic Review

Douglas Campbell, Abel Brodeur, Anna Dreber, Magnus Johannesson, Joseph Kopecky, Lester Lusher, Nikita Tsoy

Abstract

We estimate the robustness reproducibility of key results from 17 non-experimental AER papers published in 2013 (8 papers) and 2022/23 (9 papers). We find that many of the results are not robust, with no improvement over time. The fraction of significant robustness tests ($p<0.05$) varies between 17% and 88% across the papers with a mean of 46%. The mean relative t/z-value of the robustness tests varies between 35% and 87% with a mean of 63%, suggesting selective reporting of analytical specifications that exaggerate statistical significance. A sample of economists (n=359) overestimates robustness reproducibility, but predictions are correlated with observed reproducibility.

*Campbell: New Economic School (email: dolcampb@gmail.com); Brodeur: Department of Economics, University of Ottawa (abrodeur@uottawa.ca); Dreber: Department of Economics, Stockholm School of Economics (e-mail: anna.dreber@hhs.se) and Department of Economics, University of Innsbruck, Innsbruck, Austria; Johannesson: Department of Economics, Stockholm School of Economics (e-mail: magnus.johannesson@hhs.se); Kopecky: Department of Economics, Trinity College Dublin (email: jkopecky@tcd.ie); Lusher: Department of Economics, University of Pittsburgh (email: lesterlusher@pitt.edu) and IZA; Tsoy: INSAIT, Sofia University (email: nikita.tsoy@insait.ai)

Acknowledgments: For financial support, we thank the Jan Wallander and Tom Hedelius Foundation (grants P21-0091 and P23-0098 to A.D.), the Knut and Alice Wallenberg Foundation (grant KAW 2018.0134 to A.D.), the Marianne and Marcus Wallenberg Foundation (grant KAW 2019.0434 to A.D.), and Riksbankens Jubileumsfond (grant P21-0168 to M.J.). We thank Elvina Lukmanova, Camilla Puleo, Alexey Seliverstrov, and Kristina Gonchareva for excellent research assistance and seminar participants at George Mason and the New Economic School, and Hozny Zoabi for insightful comments. The reproducibility reports for the 17 included AER papers, the pre-analysis plan of the prediction survey, the robustness plans shown to survey participants, and the prediction survey are posted at OSF.

Introduction

Recent years have seen an increased concern about the credibility of empirical results in the social sciences. Much of this work has focused on lab experiments that are relatively straightforward to attempt to replicate with new data. The reproducibility project psychology (RPP) pioneered systematic work on replicability, in replicating 100 studies published in three top psychology journals (Open Science Collaboration 2015). Only 35 out of the 97 original studies that reported a statistically significant finding replicated, in terms of finding a statistically significant result in the same direction as the original study. RPP has been followed by several other systematic replication studies of lab and online experiments in the social sciences (e.g. Klein et al. 2014, 2018; Camerer et al. 2018; Ebersole et al. 2016), including the experimental economics replication project on lab experiments published in the American Economic Review (AER) and the Quarterly Journal of Economics (Camerer et al. 2016). Taken together, these studies suggest a replication rate of about 50% of lab and online experiments both in terms of the fraction of statistically significant ($p<0.05$) effects in the same direction as the original studies, and in terms of the relative effect sizes of the replications (the replication's effect size divided by the original one), with a point estimate of replicability for experimental economics of about 60% (11 out of 18 studies).

Although it is important to replicate experimental work, the issue of credibility of published findings applies to all empirical work, and it is important to also evaluate studies based on observational data, which constitute the majority of published empirical work in economics. Testing the hypothesis of the original study again in new data using the same research design as the original study is often referred to as a direct replication. Conducting direct replications on observational data studies can be challenging, but the credibility of observational data studies can also be assessed in other ways, such as testing if the posted data and code produce the reported findings and testing if a published result is robust to alternative, equally-reasonable specifications to test the hypothesis. Such tests based on using the same data as in the original study are typically referred to as tests of reproducibility to distinguish such tests from tests of replicability using new data. A further distinction can be made between computational reproducibility based on the same data and code and robustness reproducibility testing alternative

specifications to test the hypothesis (Dreber and Johannesson 2023). Systematic work on computational reproducibility has a relatively long history in economics with the first study by Dewald et al. (1986), published in 1986, and followed by a line of more recent studies (e.g., McCullough et al. 2006, 2008; Glandon 2011; Chang and Li 2017; Gertler et al. 2018; Perignon et al. 2023). Computational reproducibility in economics has been disappointingly low, leading the AER and several other journals to implement a new system with data editors to check that the data and code yield the results in the paper prior to final acceptance.

We first confirm that the AER data editors' updates to the Data and Code Availability Policy (DCAP) and prepublication verification of reproducibility (see Vilhuber et al., 2020) have been effective: in conducting computational reproducibility for all papers in our study, we generally confirmed the original results, with a few exceptions. Our study builds on the data editors' work by detecting several coding errors and potential typesetting errors that are important. For example, in Montero and Yang (2022), we found the authors had incorrectly averaged income data from the Mexican census which had been top-coded as missing, and also made errors when coding which festival dates coincide with optimal planting dates: correcting the errors render a key result (the impact of festival dates on income) just statistically insignificant at the 5% cutoff when robust standard errors are computed.

Systematic work on robustness reproducibility is still at an early stage in economics, despite Leamer (1983) raising concerns about the robustness of published findings based on observational data already in 1983.[1] In this study, we report the results of a systematic robustness reproducibility study on papers published in the AER. We include all empirical papers published within 6 months of the start of our project in 2022/23, and in a 6-month period about 10 years prior in the first half of 2013; that (1) have publicly available data, (2) have working code, (3) make a causal claim, and (4) are not experiments.[2] We identified 8 papers in 2013 and 9 papers in 2022/23 meeting these criteria, and in total, we thus included 17 papers in the study. For each

[1] See also the overview article for economics research by Christensen and Miguel (2018) and the overview of replication and reproducibility studies in economics by Ankel-Peters et al. (2023).

[2] We also made a judgment call to exclude predominantly theory papers, such as calibrated macro models and structural theory papers, regardless of whether they make causal claims, unless they also contain separate regressions that use causal analysis. We include an experimental paper (Bobonis et al., 2022), that also uses observational data, and focused on the latter portion of the paper.

paper we identified 1-3 key results reported as statistically significant ($p<0.05$) and we subject these key results to a series of robustness tests.[3] All of the 17 papers reported having at least one robust statistically significant result as a main finding, without exception.

We report results for two primary indicators of robustness reproducibility. Our first primary indicator is the fraction of robustness tests that are statistically significant at the 5% level. This indicator assesses the robustness of the conclusion of the hypothesis test in the original study. The value on this indicator varies between 17% and 88% across the studies, with a mean of 46%.[4] This result suggests that the key original hypothesis tests are not robust in a substantial fraction of the studies. Our second primary indicator of robustness reproducibility is a relative effect size indicator defined as the average t/z-value of all the robustness tests of that key result divided by the t/z-value in the original study.[5] This indicates whether the original result is systematically biased, with a ratio below one indicating systematic inflation in the reported strength of evidence in original studies.[6] This indicator varies between 35% and 87% across the papers, with a mean of 63%. The mean of this indicator suggests systematically biased results of the original studies on average, in line with selective reporting of regression specifications that yield favorable results for the tested hypotheses. This is sometimes referred to as p-hacking (intentional or unintentional) in the literature resulting from the "researcher's degrees of freedom" in conducting the analysis (Simmons et al. 2011; John et al. 2012; Gelman and Loken 2014). Brodeur et al. (2016, 2020, 2023) have previously found indications of p-hacking in observational data studies in economics. In general, exaggerated effect sizes in original studies can also be due to low statistical power (Button et al. 2013; Ioannidis et al. 2017), testing hypotheses with low priors (Maniadis et al. 2014; Dreber et al. 2015; Johnson et al. 2017), and

[3] See Appendix Table 1 for the 17 included papers.

[4] These statistics are calculated at the paper level. Note that three studies (Pop-Eleches and Urquiola, 2013, Jansen and Zhang 2023, and Bobonis et al. 2022) each had one result, out of three, which was significant at 5% in every robustness check conducted. Two additional results (one from Jansen and Zhang 2023, and one from Carlino et al. 2023) were robust in greater than 90% of robustness checks, which we view to be impressively robust.

[5] This indicator could also be based on the effect sizes rather than t/z-values, but this was not possible in this case without excluding some robustness tests as the units of the effect sizes were not comparable across robustness tests in all robustness tests; e.g. using different functional forms between dependent and independent variables.

[6] This systematic bias can stem from either systematically overestimated effect sizes or systematically underestimated standard errors.

publication bias (Hedges 1992; Stern and Simes 1997; Franco et al. 2014, 2015).[7] Note that in this setting, statistical power and the prior are held constant across robustness tests.

As a secondary reproducibility indicator, we also report the standard deviation in the t/z-value across the robustness tests (the variation indicator), which is an indicator of the variation in results across robustness tests.[8] As t/z-values are scaled in standard error units, a standard deviation in t/z-values across robustness tests of one implies that the variation in results across robustness tests is as large as the sample standard errors. The standard deviation in the t/z-values of the robustness tests varies between 0.6 and 9.5 across the papers, with a mean of 1.8 and a median of 1.2, suggesting large variation in results across robustness tests and thereby substantial researcher degrees of freedom in implementing the analysis and selectively reporting results.

What kinds of robustness tests did we implement that proved influential? Some typical robustness checks included adding fixed effects, control variables, and clustering the standard errors at different levels of aggregation. For example, one paper estimates a difference-in-difference while omitting panel fixed effects; when included, the results are no longer statistically significant at 5%. Several papers include necessary control variables in separate regressions. When we include the controls in the same regression, the results weaken. There are also cases where authors add multiple fixed effects at once, when adding them one at a time leads to insignificant results. Our general guideline was to run robustness checks if we believed the robustness check to be both valid and interesting to run. This leads us to another point: we do not discuss the appropriateness of the identification strategies employed, but rather we take those strategies as given and test robustness to various controls, fixed effects, clustering schemes, and the exclusion of influential outliers. This implies that many of the results are not robust even when we accept the validity of the chosen identification strategies.

We also conducted influential analysis for each study. We omitted outliers with calculated dfbeta statistics, a measure of the influence of each observation, larger than the standard cutoff of $2/\sqrt{N}$

[7] See also the early work by Ioannidis (2005) claiming that most published research findings are false.
[8] The variation indicator can also be defined for effect sizes if the effect sizes are in the same units across robustness tests; and in this case the indicator would be defined as the standard deviation in effect sizes among the robustness tests divided by the standard error of the original study estimate (Dreber and Johannesson 2023). We define this indicator in t/z-value terms so that it is already scaled in standard error units.

(see Belsley, et al. 1980). For the latter exercise, we found that the t/z-values changed by at least 10% in 23 out of 28 regressions where applicable[9] (shrinking in 15 cases and increasing in eight), with relative t/z values ranging between -0.2 and 11.5, and with a median of 0.89 and a mean of 1.25. When we omit outliers in regressions with our added controls and specifications we deemed appropriate, relative t/z-values ranged from -2.2 to 4.2, with a median of 0.65 and a mean of 0.79. This suggests, at a minimum, that influential outliers are a potential issue complicating inference in published economics research. Our findings also demonstrate how easy it could be in practice to generate statistically significant results by including or excluding different categories of observations, or even by excluding a small number of data points. Indeed, we find multiple instances of what appears to be selective data inclusion. For example, an author might implement a necessary robustness check only on an alternative sample, when running the same check on the main sample would lead to p-values above 0.05. We find that excluding groups of data points without adequate explanation to the reader is far too common.[10]

As we included papers from two periods we can also test if there is a trend towards improved robustness during these 10 years. We find no evidence of such a trend for any of our robustness indicators with point estimates of the difference between periods close to zero. However, the low sample size makes it difficult to draw strong conclusions on this as we are only powered to detect large improvements between the periods.

We furthermore implemented a robustness survey where economists were asked to predict the result of the robustness tests of our two primary reproducibility indicators. The forecasters were required to have a Ph.D. in economics or finance or be a Ph.D. student in economics or finance. A sample of 359 forecasters who completed the survey were included in the preregistered

[9] Note that this measure cannot currently be applied in Stata for non-linear regressions, or for two-stage least squares. For two stage least squares, we computed dfbeta statistics for the reduced-form regressions.

[10] Chen (2013) finds that speakers of languages that do not differentiate the present from the future are much more likely to save. He omits South Korea, a country with high levels of savings but with a dominant language that does differentiate the present from the future. Angelucci et al. (2023) find that royal English boroughs which were predicted to become farm grant towns based on their trade-related geography were subsequently more likely to be summoned to Parliament by 1348 due to their administrative independence. A key identifying assumption is that royal and mesne (non-royal) boroughs were otherwise similar. Our review of their sources suggests that 30 non-royal boroughs were excluded from their analyses. Via email, we confirmed that the authors' data collection procedure differed in a subtle way from the description in their posted data appendix. In their appendix, they wrote: "we exclude boroughs that disappeared before 1348", whereas via email from Nico Voigtlander they added that boroughs which had disappeared or for which there was "no confirmation" were excluded.

analyses on the prediction survey.[11] We found a sizable positive association between predictions and outcomes for the predictions of the fraction of statistically significant robustness tests, but only a weak association for the predictions of relative t/z-values. Forecasters also on average overestimated the robustness reproducibility by about 15 percentage points. Forecasters were also asked to rate the credibility of the papers, with the average rating being only six out of ten where ten is "very credible", suggesting that economists harbor doubts about the credibility of published research in the American Economic Review. Forecasters on average rated papers in the more recent period (2022/23) higher on credibility and expected slightly higher robustness of these papers. However, as noted above, no such trend was observed in the actual robustness tests.

We conclude that our study identifies substantive problems with robustness for many of the 17 AER papers studied here. This lack of reproducibility mirrors the problems previously reported on computational reproducibility in economics, where the published results often cannot be reproduced based on the posted data and code (Dewald et al. 1986; McCullough et al. 2006, 2008; Glandon 2011; Chang and Li 2017; Gertler et al. 2018). But, in our context, a lack of computational reproducibility is not driving our results as the computational reproducibility was high in our sample; this is unsurprising given the new Data Editor system implemented by the AER for the 2022/23 sample.[12] Additional policies will be needed to improve robustness reproducibility, and in future work it is also important that published papers openly report the uncertainty in results due to analytical decisions using for instance multiverse or multi-analyst methods (Steegen et al. 2016; Silberzahn et al. 2018; Botvinik-Nezer et al. 2020; Simonsohn et al. 2020; Menkveld et al. forthcoming).[13]

Our paper adds to a large literature on research credibility, in particular studies documenting the extent of selective reporting, p-hacking and publication bias (e.g., Andrews and Kasy 2017; DellaVigna and Linos 2022; Doucouliagos and Stanley 2013; Dreber et al. forthcoming;

[11] A detailed pre-analysis plan (PAP) was posted at Open Science Collaboration (OSF) prior to starting the data collection (https://osf.io/w7vpu/).

[12] We also assessed the computational reproducibility and for all 17 papers the original results could be reproduced almost exactly, with mostly non-material differences (8 papers in 2013 and 9 papers in 2022/23). For several papers, we replicated the coefficients, but estimated slightly larger standard errors than the original papers.

[13] In multiverse analyses (also referred to as specification curve analysis), all theoretically justifiable non-redundant analyses are performed and presented. In multi-analyst studies, many researchers are asked to test the same hypotheses on the same data and these results are then presented, typically highlighting large variation in effect sizes and t-values.

Havranek et al. forthcoming; Milkman et al. 2021; Vivalt 2019), and meta-studies investigating the reproducibility, replicability and robustness of empirical claims (e.g., Camerer et al. 2016, 2018). Christensen and Miguel (2018) and Stanley and Doucouliagos (2014) provide literature reviews. We contribute to this literature by (i) testing the robustness of claims in 17 articles published in a leading economic outlet, in contrast to studies focusing on one method (e.g., Young 2022); (ii) testing whether recent claims are more robust than older ones; and (iii) documenting whether economists can predict the robustness reproducibility of non-experimental work.

I. Methods

A. Computational reproducibility

Our full sample for computational reproducibility includes all empirical papers published in a 6-month period in 2013 (issues 1, 2, and 4; issue 3 was excluded as it was a Papers and Proceedings issue) and a 6-month period about 10 years later in 2022/23 (issues 10-12 in 2022 and issues 1-3 in 2023). We first removed papers that do not seem to make a causal empirical claim, and experiments.[14] For the rest, we then attempted to reproduce all regression tables in the main part of the paper (excluding appendices) using the replication packages provided by the authors via the AER website. We thus exclude papers with non-public data. We identified 8 papers in 2013 (out of 56 papers published in the AER in this period) and 9 papers in 2022/23 (out of 51) meeting these criteria. We recorded the coefficients, standard errors, and t/z-values for each of the key variables of interest, omitting control variables.

B. Robustness reproducibility

[14] Note that on these grounds we have excluded theory papers, descriptive papers, and exclusively macroeconomic calibration exercises, as these are arguably predominantly theory and non-causal, as well as structural identification papers. Also, we did include one paper, Bobonis et al. (2022), which contains an experiment and also contains non-experimental causal claims using observational data. We also include several papers which have calibration exercises, but also run regressions on observational data with causal claims attached. In these cases we focused on the regressions run with causal claims.

We carried out robustness tests on between 1-3 key results for each paper; the included results were those considered most central to the papers and that were statistically significant at the 5% level in the original papers. The robustness tests consisted of a number of alternative analyses to estimate these 1-3 key results using the same data as in the original study (the number of robustness tests could vary between the 1-3 results in a paper and could also vary between the included papers). We refer to the person conducting the robustness tests as the "reproducer" below.

For each key result of the 17 papers, we estimated two primary robustness reproducibility indicators: the statistical significance indicator and the relative t/z-value indicator (Dreber and Johannesson 2023). The statistical significance indicator was defined as the fraction of robustness tests of that key result that was statistically significant at the 5% level in a two-sided test with an effect in the same direction as the effect observed in the original paper. This indicates the strength of support of the hypotheses tested in the original paper. The relative t/z-value indicator was defined as the average t/z-value of all the robustness tests of that key result divided by the t/z-value in the original study. This is an indicator of systematic bias in the original results, with a ratio below one suggesting systematic overestimation of the reported strength of evidence in the original results (due to systematic overestimation of effect sizes and/or systematic underestimation of standard errors). We also included a secondary reproducibility indicator which is the variation indicator. The variation indicator was estimated as the standard deviation in t/z-values among the robustness tests. This is an indicator of the variation among the robustness tests. See Dreber and Johannesson (2023) for a more detailed description and discussion of the robustness indicators used. We report the results of the robustness tests for all the 1-3 key results of each paper as well as aggregated across each paper so that we get one robustness estimate per paper (to estimate the reproducibility for a paper with more than one key result, we take the average of the robustness indicator for each result). We focus on the results at the paper level in the main text, but report results also on the results level in the Appendix.

C. Prediction survey

As part of the study we also carried out a prediction survey to test if economists could predict the outcome of the robustness tests as measured by our two primary reproducibility indicators (see, e.g., Dreber et al. (2015) and DellaVigna and Pope (2018) for earlier work investigating expert predictions of scientific results). In the survey, participants were asked two prediction questions for each key result in four randomly selected papers out of the 17 papers included in the study (two randomly selected AER papers published in 2013 and two randomly selected AER papers published in 2022-2023, with the question order of these papers also randomized on the participant level). We refer to the participants in the prediction survey as forecasters. In the first prediction question forecasters were asked to predict the % of statistically significant robustness tests for each key result in the four randomly selected papers. We refer to this as the "FSP question" below where FSP stands for "fraction significant prediction". In the second prediction question, forecasters were asked to predict the average relative t/z-value of the robustness tests for each key result in the 4 randomly selected papers (where the average relative t/z-value is estimated as the average t/z-value of all the robustness tests of that result divided by the t/z-value of the original result). We refer to this as the "RESP question" below where RESP stands for "relative effect size prediction". They predicted the RESP in % terms as for the first prediction question. Answers to both the FSP and RESP questions were divided by 100 in all analyses and tests below so that they were expressed in fractions rather than % (e.g. a prediction of 80% significant robustness tests in the FSP question was equal to 0.8 in all analyses and tests below).

Before making these predictions, forecasters received information about the following for each key result of the four randomly selected papers: the hypothesis, the coefficient, the standard error, the t/z-value and the p-value. For each paper, they furthermore received a link to the paper and a link to a "robustness plan". The robustness plan included a summary of the planned robustness tests for each key result in that paper (but the exact robustness tests were not detailed in the robustness plan, as the "reproducer" decided on the exact number of robustness tests while conducting the tests as one test may lead to another test and so on, and the exact number of robustness tests were not known at the time of completing the survey). Some robustness tests of a key result were based on two tests needing to be significant at the 5% level, such as the coefficient of a variable and the coefficient of the squared variable for studies testing non-linear relationships, and both a coefficient and an interaction coefficient for studies testing for

interaction effects (in these cases, the relative t/z-value were based on the average for more than one coefficient).[15] This was explained to forecasters. For studies using instrumental variables, only the 2nd stage results were predicted; but for the first prediction question about the % of robustness tests that will be statistically significant, the significance of the 2nd stage robustness tests were based on Anderson-Rubin confidence intervals if the 1st stage Montiel Olea and Pflueger (2013) F-statistic was less than 10.[16] For the second prediction question about the relative t/z score, the t/z score in the 2nd stage was based on the standard error of the regular estimation (as Anderson-Rubin confidence intervals do not yield a standard error or t/z-value).

For each of the four randomly selected papers, forecasters were also asked about the credibility of the results of the paper on a 0 (not at all credible) to 10 (very credible) scale and their familiarity with the methods of this paper on a 0 (not at all familiar) to 10 (very familiar) scale. In addition, forecasters were asked a number of background questions at the end of the survey. The background questions are used to describe the sample of participants, and two of the questions (position/seniority and sub-field of economics) were also included in preregistered exploratory analyses detailed below. The following background questions were included: current position, field, region, and gender. The question about the field was used to code a dummy variable included in the preregistered exploratory analyses for if the forecaster was in the same field as the paper being predicted. The following sub-fields of economics were included in this question: "Economic History/Growth/Macro Development", "Labor/Public/Health/Education", "Development/Political Economy", "Macroeconomics", "International Trade", and "Other". All the 17 AER papers included in our study were coded into one of these sub-fields.

[15] This was the case for two papers; one paper where the test was based on the coefficient of one variable and the coefficient of the squared variable (Ashraf and Galor 2013) and one paper where the test was based on the coefficient of a variable and an interaction coefficient (Boehm and Pandalai-Nayar 2022).

[16] Note that we report the Anderson-Rubin confidence intervals in our robustness reports whenever they can be calculated, regardless of the first stage F-statistic. Also note that Montiel Olea and Pfluegger (2013) F-statistic can not be calculated in the overidentified case with multiple endogenous regressors. Lewis and Mertens (2022) extend the Montiel Olea and Pfluegger (2013) heteroskedasticity-robust F-statistic to the case of multiple endogenous regressors, with computation in Matlab, and where applicable, we implemented their procedure. In practice, the Anderson-Rubin confidence sets carried very similar implications to our second stage t-tests, so that using a cutoff of 10 vs. other critical values does not make a difference for our sample. See Andrews et al. (2019) for an overview of weak instrument issues.

We invited participants to the survey in various ways, sending invitations to economists cited in the included papers, faculty and PhD students at top economics departments, participants in the Institute for Replication replication games (see Brodeur et al. 2023), professional organizations and networks, plus through our personal networks. To participate, the participant had to be a PhD student in economics or finance or have a PhD in economics or finance. Before getting a link to the survey, participants had to fill out a "sign up form" with position and affiliation that we used to check that participants fulfilled our inclusion criteria (PhD student or a PhD in economics or finance). We started sending out invitations to the survey in July, 2023, and the deadline for completing the survey was October 31, 2023. We preregistered to close the survey data collection prior to this deadline if we reached 300 participants who had finished the survey (completing all the questions about the four randomly selected papers included in their survey), and to allow the remaining participants to finish the survey within two weeks (even if these two weeks implied a date after October 31). We reached 300 completed survey responses on October 30 and gave the remaining participants two more weeks to complete the survey (until November 13), reaching 359 completed surveys in total. 51% of the respondents were PhD students, 26% were post-docs or assistant professors, 18% were associate or full professors, 1% had other positions with a PhD in academia, and 3% had a PhD but were working outside of academia. The respondents were allocated across sub-fields in the following way: 7% in "Economic History/Growth/Macro Development", 28% in "Labor/Public/Health/Education", 16% in "Development/Political Economy", 14% in "Macroeconomics", 3% in "International Trade", and 32% in "Other". 66% were working in Europe, 24% in North America, 2% in Central or South America, 4% in Asia, 0% in Africa, and 3% in Australia/New Zealand. The fraction of female respondents was 26% and the fraction of male respondents 71%, and 3% responded "other/prefer not to say" on the gender question.

Only participants who completed the survey, defined as clicking a submit button at the end of the questionnaire ("complete participants"), were included in the tests and analyses of the survey data. The survey was conducted in Qualtrics, and we used "force response" in Qualtrics on all the survey questions about the four randomly selected papers and all "complete participants" therefore by definition responded to all the survey questions about their four randomly selected papers. For the background questions, we did not use "force response" and "complete

participants" were included in the tests and analyses of the survey data even if they did not respond to all the background questions.[17] When we refer to forecasters in the analyses and tests below, we mean "complete participants" who answered all the survey questions about the four randomly selected papers in their survey.

The survey was incentivized using the following quadratic scoring rule:

$$\$30 \; - \; (\overline{Sq.Error} \; \times 120)$$

where $\overline{Sq.Error}$ is the average of the squared errors for all the predictions on both the FSP and RESP questions made by the forecaster. The squared error is the squared difference between the prediction and the outcome with both expressed as fractions rather than %. The forecasters could choose between getting the bonus paid as an Amazon gift card or donating it to charity (with a choice between three charitable organizations).[18] On average forecasters earned $14.63.[19]

The design of the survey data collection and all hypotheses and tests of the survey data were preregistered in a pre-analysis plan (PAP) prior to the start of the survey data collection. This PAP also included the inclusion/exclusion criteria and data collection stopping rule detailed above. The PAP and the prediction survey are posted at OSF (https://osf.io/w7vpu/).

II. Results

A. Computational reproducibility

To a first approximation, we find that the AER data editors do exemplary work – it was possible to reproduce almost all the original regressions exactly, with a few small exceptions.[20] Overall,

[17] With the exception of the preregistered exploratory analyses excluding participants that did not answer the background questions about position/seniority and sub-field of economics included as variables in those exploratory analyses; the sub-field question is used to construct a variable for being in the same sub-field as the paper.
[18] The three charities were GiveWell, GiveDirectly, and Unicef.
[19] Out of the 359 forecasters 125 donated their earnings to charity and 233 opted for an Amazon gift card. For those who donated, 41% opted for Unicef, 38% for GiveWell, and 21% to Give Directly. The payments ranged from $28.92 to zero.
[20] For one paper, Jansen and Zhang (2023), there were some tables created with non-public data which we omitted, but reproduced all the others, including those with the key regressions we tested robustness on.

100% of our reproduction regressions were significant at 5%, and the relative t/z values were 98%. Computational reproducibility slightly improved in 2022/2023 vs. 2013, with relative t/z values increasing from 96% to 100%.

The largest discrepancies were from 2013. In Cloyne (2013), the main result was contained in a figure, and while we could reproduce the figure exactly, we did not find the computed p-values contained in the text in the replication package. In our reproduction, the z-value for one coefficient of interest fell from 3.1 to 2.27 (the z-value for the other variable of interest was essentially unchanged). Also, for Ashraf and Galor (2013), our standard errors differed slightly from the original, with t-values for the two variables of interest falling by about 1.6% – a minor difference.

For 2023, we found what are likely to be several typesetting errors. In Jansen et al. (2023), two robustness regressions were reported as insignificant in the original paper, but our computational reproduction found smaller standard errors and statistically significant results. This is important, as these insignificant robustness checks stood as the only insignificant original robustness checks of the key regressions that we study in our paper. More importantly, in Corno et al. (2023), we found a typesetting error which, if not fixed, would seemingly invalidate the identifying assumptions. The authors write that the fraction of Black students living with a student of a different race in the sample is only a mere 0.023 -- when, given the fraction of Black students in the study, one might have expected exactly half to have had a roommate of a different race, implying that roommate selection was highly non-random. We confirmed with the authors that the true number should be 0.23. This example highlights the difficulty in planning a robustness report, or even in judging empirical research, without access to the original data.

Two papers in our sample (Angelucci et al. 2022 and Montero and Yang 2022) use author-collected data that could be verified with the original sources. We conducted a brief review of both sources, and, as mentioned above, we found significant issues with the data in both papers. This included coding errors and misattributed data points in Montero and Yang (2022). In Angelucci et al. (2022), we found that their data exclusion criteria differed in a subtle

way from what was described in their data appendix. In the latter case, we were able to confirm our interpretation with the authors via email.

B. Robustness reproducibility

In Figure 1, we compare the distribution of t/z values for the 29 included key results from the original papers, with the t/z values of the robustness checks of these key results reported in the original papers, and the t/z values for all the robustness tests run in our study. Values have been normalized so that original findings all have positive t/z-values, and the robustness checks have positive values only when of the same sign. t/z values over five and below -2 were truncated.[21] A striking finding is that for the 29 results we focus on for 17 papers, the original papers report additional 216 regression results we deemed to be close substitutes, and of the combined 245 regressions, there is not a single valid p-value greater than 0.09, and only a small handful are marginally greater than 0.05. Part of the reason for this distribution of p-values was that we focused on benchmark regressions significant at 5%. That said, all 17 of the papers in our study feature at least one central result that the authors argue is very robust. The distribution of t/z-values in our robustness checks, by contrast, is markedly different, with the median (unweighted) absolute t/z-value declining from 2.9 in the benchmark regressions in the original papers to 2.0 in our robustness checks, with nearly half of our robustness checks being statistically insignificant at 5%, and 9% flipping sign. In Appendix Figure 1, we display the distribution of t/z-values for an exercise where we omitted influential observations from the authors' exact specification – roughly one-third of these robustness checks were found to be insignificant.[22]

[21] For regressions with multiple coefficients of interest (such as regressions with linear and quadratic terms), we have averaged the t/z-values.

[22] We also performed an exercise where we conducted influential observations on some of our specifications that included additional controls, and these results are reported in Appendix Figure 2.

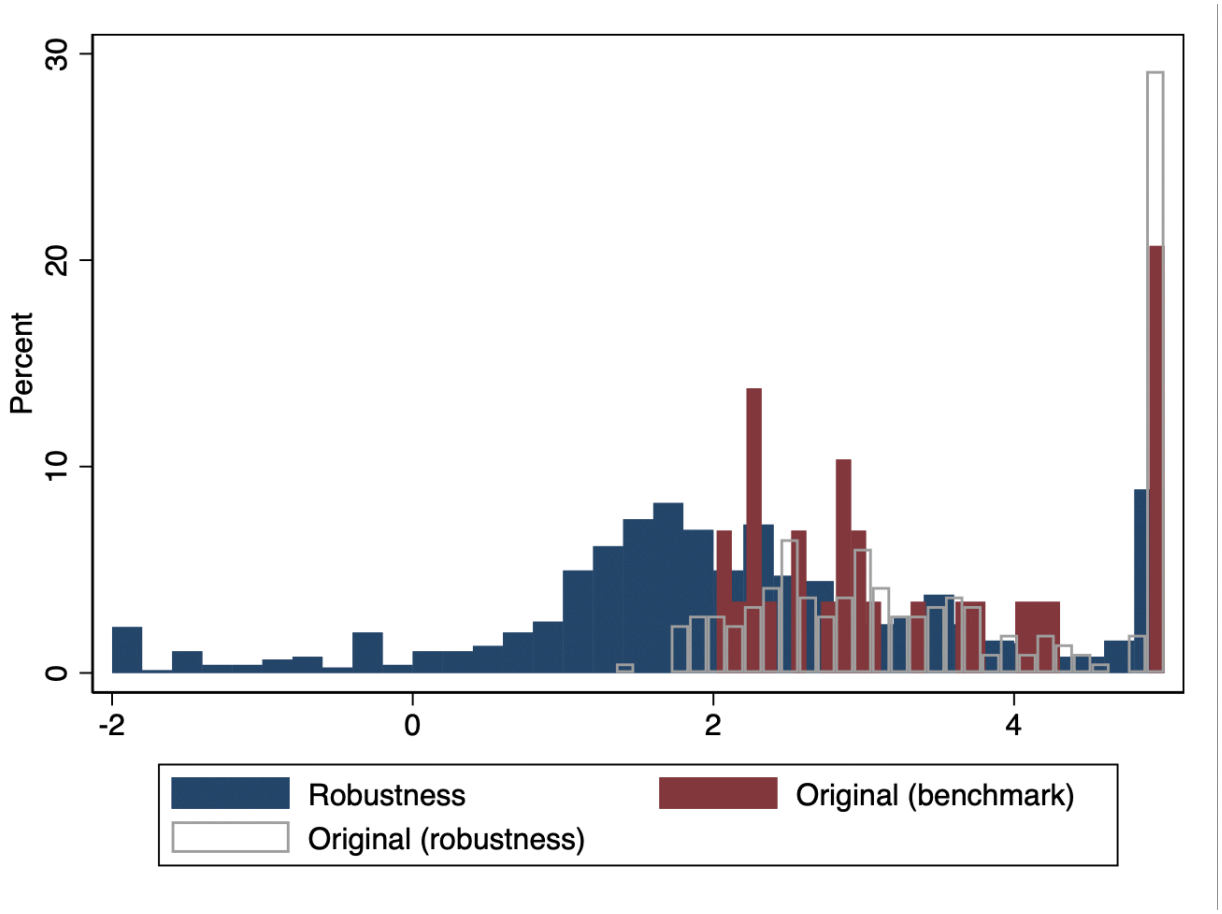


Figure 1: The distribution of absolute t/z-values in the original papers vs. robustness

Notes: The t/z-statistics have been normalized so that all values for the benchmark regressions from the original papers are positive. The robustness t/z-values from the original papers include all regressions which are close substitutes to the benchmark regressions. Our robustness results are normalized so that t/z-values in the same direction as the original study are positive, and those in the opposite direction are negative. There are 29 benchmark results, 216 similar specifications in the original papers, and our 765 robustness checks.

In Figure 2, we show the results for the statistical significance indicator and for the relative effect size indicator at the paper level (the results disaggregated on each key result in a paper are shown in Appendix Figure 3). The fraction of statistically significant robustness tests varies between 17% and 88% across the 17 papers with a mean of 46% and a 95% confidence interval of 35-56%.

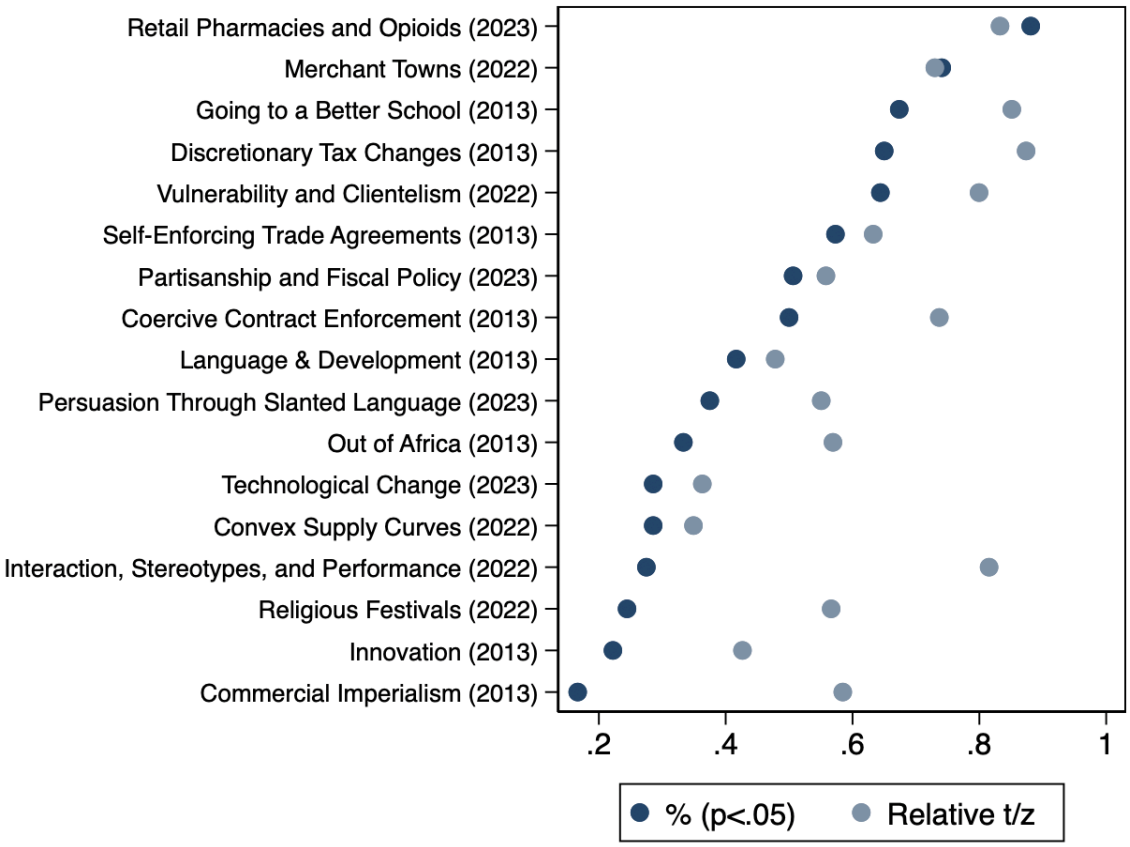


Figure 2: Statistical significance and relative t/z-values by paper

The relative t/z-value varies between 35% and 87% across the papers with a mean of 63% and a 95% confidence interval of 55-72%. The statistical significance and relative t/z-value indicators are associated with each other, but they do not always align (the Pearson correlation between the two indicators at the paper level is 0.67 (p=0.0033) and the correlation at the results level is 0.63, p=0.0002).[23] This makes some sense: studies with initial p-values closer to 0.05 are more likely to have a higher fraction of robustness checks fail at a 95% level of confidence for a given level of relative t/z-values (and a paper can in principle have a low fraction of statistically significant robustness tests and still have a relative t/z-value of one, which would suggest substantial

[23] Regressing relative t/z scores, by result, on the percentage significant indicator yields a coefficient of 0.62 (t=4.25, p=0.0002, R-squared = 0.40); the same regression on the paper level yields a coefficient of 0.83 (t=3.5, p=0.003, R-squared=0.45)

variation in results across robustness tests, but no systematic bias in the original result reported in the paper).

We find that original t/z-values do predict the share of p-values below 0.05, but do not predict the relative t/z-values.[24] Thus, very significant findings are more likely to be robust in terms of the statistical significance indicator, but their t/z-values fall by just as much as significant findings with a p-value close to 0.05. In Figure 3, we plot the relationship between the absolute t/z values reported in the paper vs. the percentage of p-values less than 0.05 from our robustness checks on the left, and vs. relative t/z-values of robustness checks on the right (in Appendix Figure 4 we plot the corresponding relationship disaggregated on each key result).

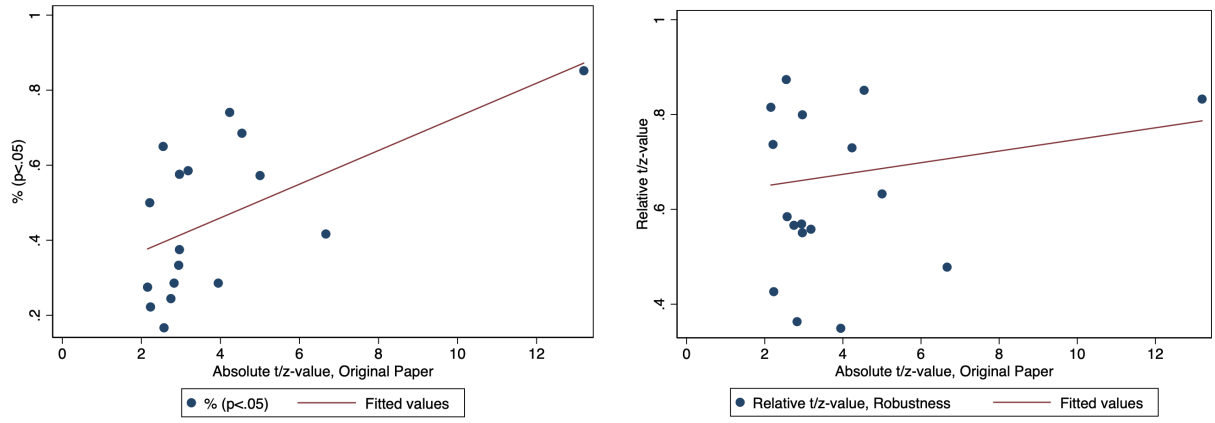


Figure 3: Original absolute t/z values vs. robustness indicators

In Figure 4 we show more detailed information about the relative t/z-value of the different papers including also the median and the 10th and 90th percentiles of the distributions. A value below one suggests that our robustness checks had lower t/z-values than the original study on average, and we test if the mean of 0.63 (63%) is below one using a one-sample t-test. Our results are consistent with systematic bias of the original studies (t-value=9.1; p-value<0.0001; 95% CI 0.54 to 0.71).[25]

[24] The Pearson correlation between original absolute t/z-values and the statistical significance indicator is 0.58 (p=0.014) at the paper level and 0.50 (p=0.006) at the results level. The Pearson correlation between original absolute t/z-values and the relative t/z-value indicator is 0.14 (p=0.60) at the paper level and -0.052 (p=0.78) at the results level.

[25] Interestingly, 85% of the robustness checks we ran (650 out of 765) had relative t/z-values less than one. This indicates that most robustness checks reduce significance.

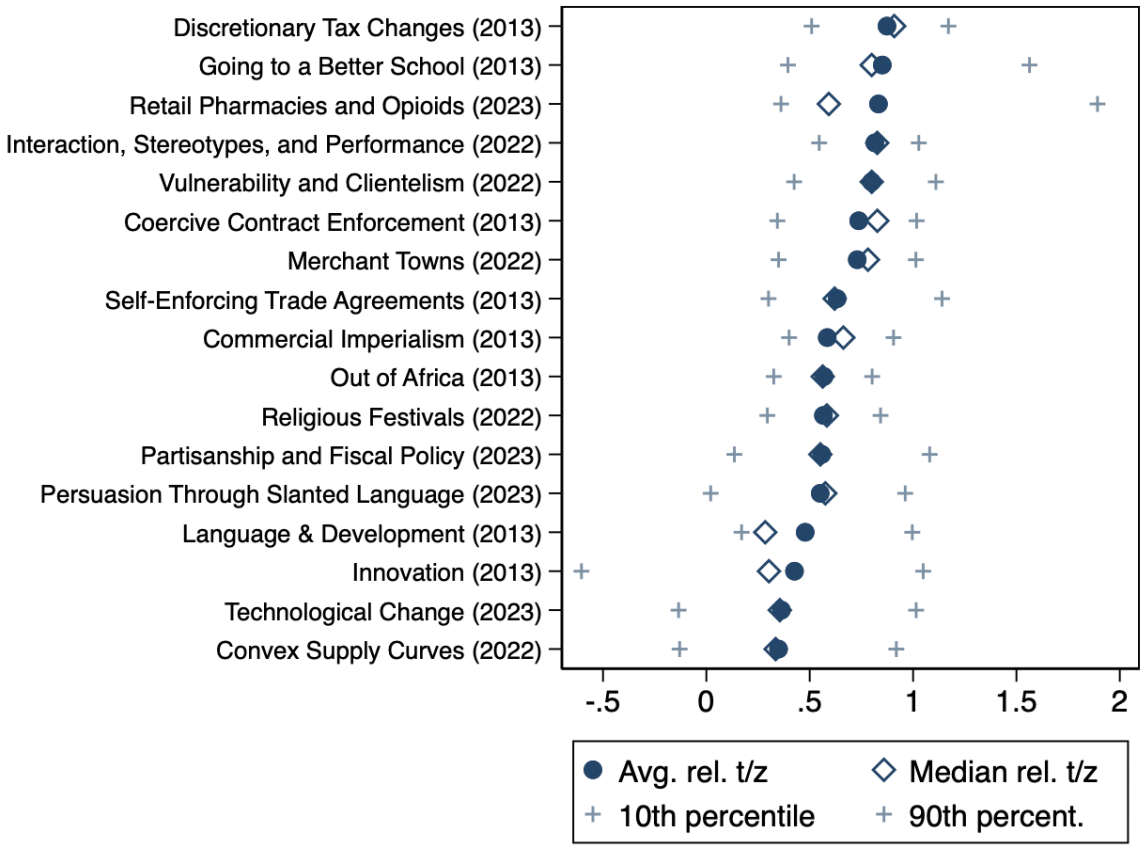


Figure 4: Relative t/z values by paper

Notes: Values for each measure (e.g., 10th percentile) have first been computed at the finding level, and then averaged across findings within each paper.

The results for the variation indicator are reported in Figure 5 and the variation indicator varies between 0.62 and 9.5 across the papers with a mean of 1.8 and a 95% confidence interval of 0.7 to 2.9 (the results disaggregated on each key result in a paper is shown in Appendix Figure 5, and the correlation between initial absolute t/z-values and the variation indicator is plotted in Appendix Figure 6). A variation indicator of one implies variation among the robustness tests corresponding to one standard error of the effect size (i.e. one t/z-value unit), and 10 papers have a variation indicator exceeding one. The variation indicator can be interpreted as a measure of researcher degrees of freedom in the analysis and the scope for p-hacking, and our results suggest that the researchers' degrees of freedom are large on average in these papers with a variation across robustness tests larger than the sampling variation. This also illustrates that

p-hacking or related behaviors may not only imply marginally changing results around the significance threshold, but can be associated with substantial changes in effect sizes and/or standard errors. The large variation in results among robustness tests is also consistent with recent multi-analyst studies where many analysts independently test the same hypothesis in the same data with results varying depending on the analytical choices of the researchers (e.g., Silberzahn et al. 2018; Botvinik-Nezer et al. 2020; Menkveld et al. forthcoming).

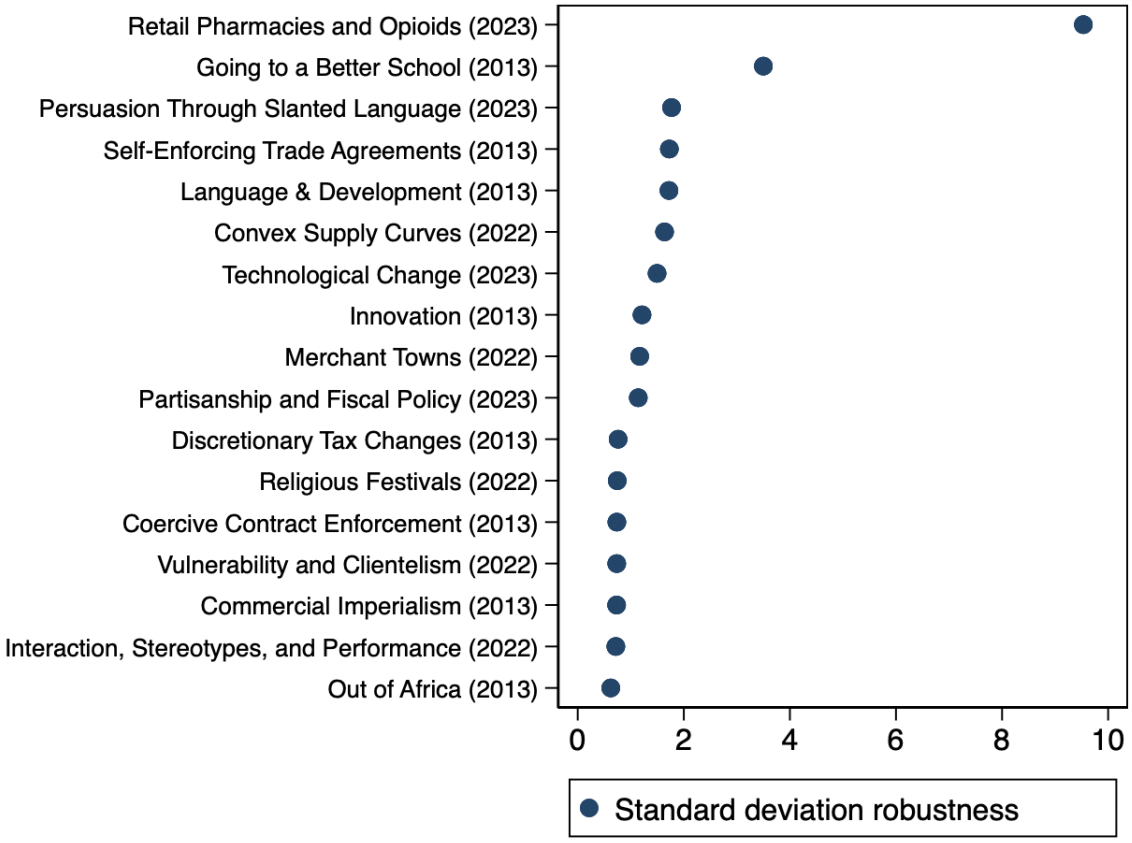


Figure 5: Variation indicator (the standard deviation of the t/z-values of the robustness tests) by paper

It is also interesting to compare robustness results across the two time periods. In Figure 6 we plot the 95% confidence interval for the two primary robustness indicators in the two periods (2013 and 2022/23) on the paper level. The increase in both the statistical significance indicator

and relative t/z-value indicator in this ten year period is close to zero, and we cannot reject the null hypothesis of no difference using an independent samples t-test for any of the robustness indicators at the 5% level.[26] But note that the statistical power of detecting a difference is limited with only 17 observations (papers). The minimum detectable effect size (MDE) to detect a significant difference at the 5% and 0.5% levels are 0.29 and 0.37 for the statistical significance indicator, 0.23 and 0.30 for the relative t/z-value indicator, and 3.6 and 4.7 for the variation indicator.[27] Thus our sample is large enough only to rule out a dramatic increase in robustness reproducibility over this period, and the averages are not indicative of improvement.[28]

[26] For the statistical significance indicator the difference (2022/23 minus 2013) between the two periods is 2.6 percentage units (t-value=0.25; p-value=0.81; 95% CI -0.19 to 0.24); for the relative t/z-value the difference between the two periods is -2.4 percentage units (t-value=-0.29; p-value=0.78; 95% CI -0.20 to 0.15). Results for the variation indicator are not shown in Figure 6, but it has a mean of 2.1 in 2022/23 (95% CI -0.06 to 4.3) and a mean of 1.9 in 2013 (95% -.19 to 3.9), and the difference between the periods is 0.26 (t-value=0.2, p-value=0.8; 95% CI -2.5 to 3.0).

[27] The MDE for 80% statistical power is estimated as 2.8 times the standard error of the difference for tests at the 5% level and 3.65 times the standard error of the difference for tests at the 0.5% level.

[28] That said, for the subsample of robustness tests where we omitted influential outliers with large dfbeta statistics using the authors' own specifications, the later period did slightly better, with a median relative t/z-value of 0.92 vs. 0.66 for 2013. However, note that our influential robustness checks have high variance, and the difference in the medians (or means) is not statistically significant across years.

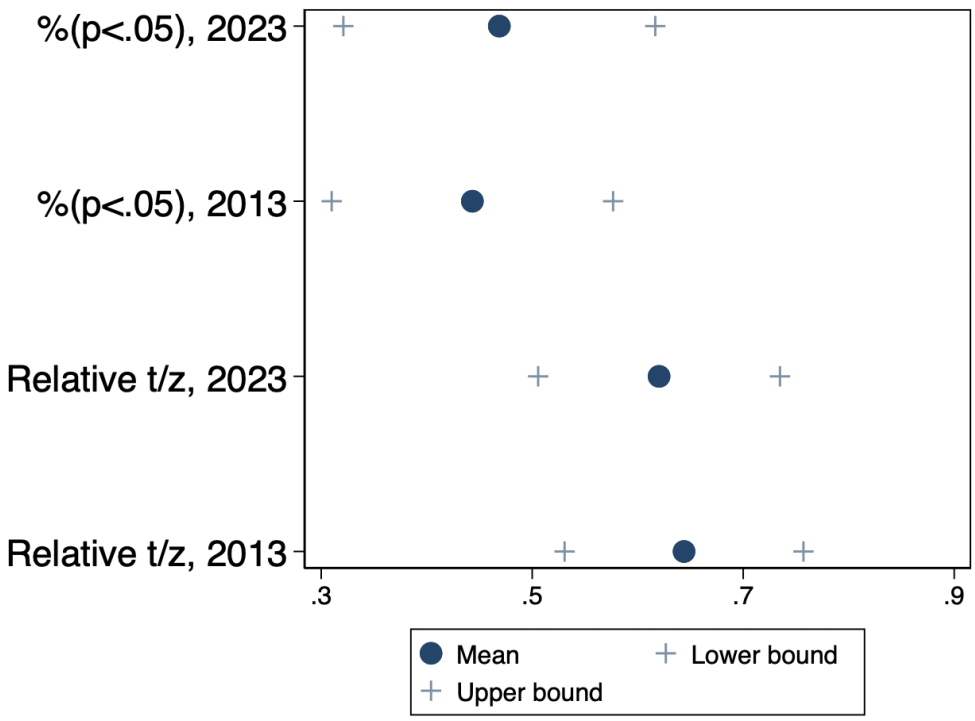


Figure 6: Comparison of 2013 vs. 2022/2023

C. Prediction survey

We pre-registered three primary hypotheses (divided into A and B for the predictions of the fraction of significant and the predictions of relative t/z-values), two secondary hypotheses, and two exploratory analyses (divided into A and B for the predictions of the fraction of significant and the predictions of relative t/z-values). In testing these hypotheses, we will as preregistered interpret a two-sided p-value below 0.05 as "suggestive evidence" and a two-sided p-value below 0.005 as "statistically significant evidence" based on the recommendations of Benjamin et al. (2018).

We furthermore pre-registered to descriptively report the mean, the standard deviation, the standard error and the 95% confidence interval of the FSP question and the RESP question for each of the 1-3 key results of the 17 papers; and also to report these descriptive results on the paper level after averaging the 1-3 key results predicted per paper (these descriptive results are

reported in Appendix Tables A2-A5). Finally, we preregistered to descriptively report the Pearson correlation and the 95% confidence interval of the Pearson correlation between these two sets of aggregated predictions and the observed outcomes that are predicted (four correlations in total; two for the FSP question and two for the RESP question). The tests and analyses below follow the pre-analysis plan exactly with the only deviation that we have added 3 robustness tests that were not pre-registered (and these tests are clearly labeled as not pre-registered below).[29] As preregistered we also report the minimum detectable effect size (MDE) that we had 80% statistical power to detect at the 5% level and the 0.5% level in the primary and secondary hypothesis tests (but not the exploratory analyses).

Figure 7 (FSP question) and Figure 8 (RESP question) descriptively report the 95% confidence intervals of the average predictions for each paper (the corresponding results on the key result level are shown in Appendix Figures 7 and 8). The Pearson correlation between each aggregated prediction and the outcome of the robustness tests is 0.72 ($p<0.001$) on the key results level and 0.54 ($p=0.025$) on the paper level for the FSP question.[30] The corresponding correlations for the RESP question are 0.31 ($p=0.11$) and 0.22 ($p=0.39$).[31] The sizable correlation between the FSPs and outcomes is illustrated in Appendix Figure 9. In primary hypothesis 1 below we more formally test for an association between predictions and robustness tests outcomes.

[29] These not pre-registered robustness tests are the robustness test of the Table 1 results with two-way clustering (reported in Appendix Table A6); the robustness test of the results in Table 2 (reported in the text below); and the robustness test of Table 4 with two-way clustering (reported in Appendix Table A7).

[30] These correlations are of the same magnitude as the correlation between prediction market prices and replication outcomes in previous studies using prediction markets to predict replication outcomes in direct replications of experiments. Gordon et al. (2021) reported a correlation of 0.58 between prediction market prices and replication outcomes pooling data from several studies.

[31] We preregistered that we would descriptively report these correlations with 95% confidence intervals, but they should not be interpreted as hypothesis tests (the hypothesis tests on the association between predictions and outcomes are reported in Table 1). The 95% CIs of these four correlation coefficients are: 0.482-0.860; 0.083-0.811; -0.037-0.624; -0.30-0.63).

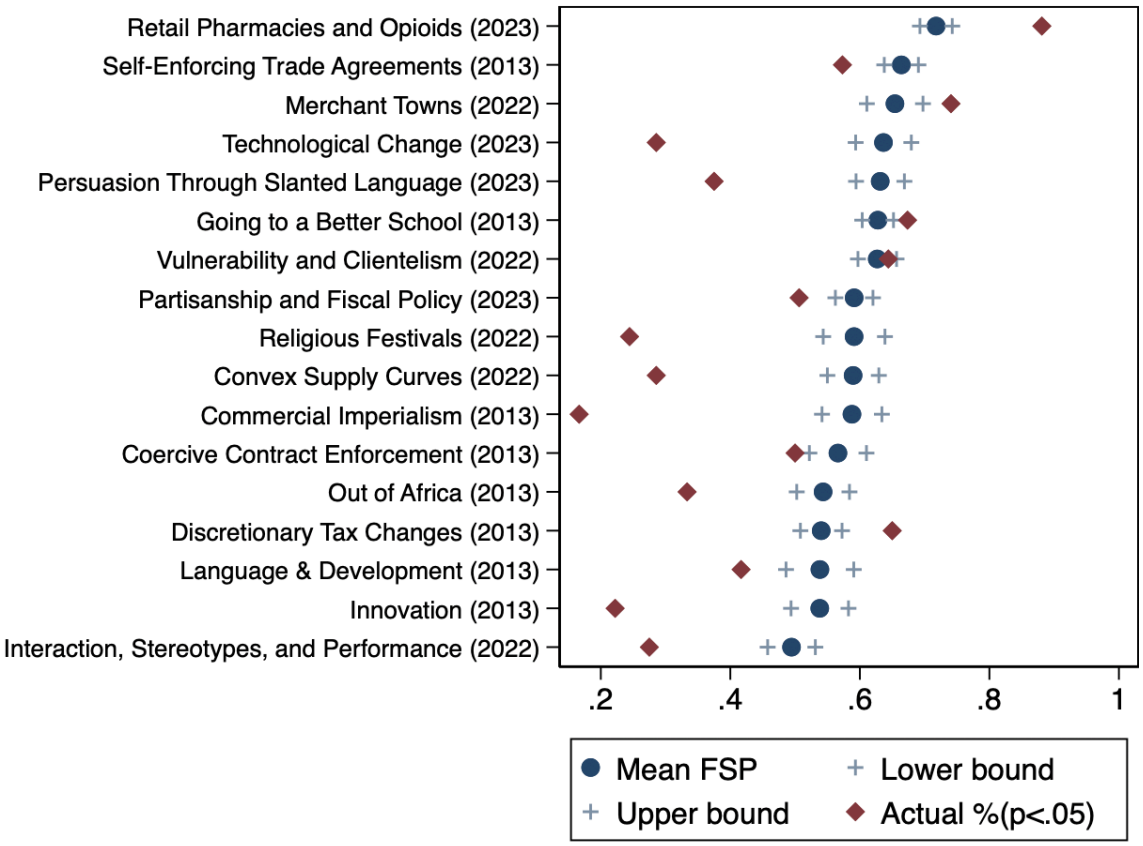


Figure 7: Mean predictions of the fraction of statistically significant robustness tests per paper

Notes: Upper and lower bounds are plotted as 95% confidence intervals.

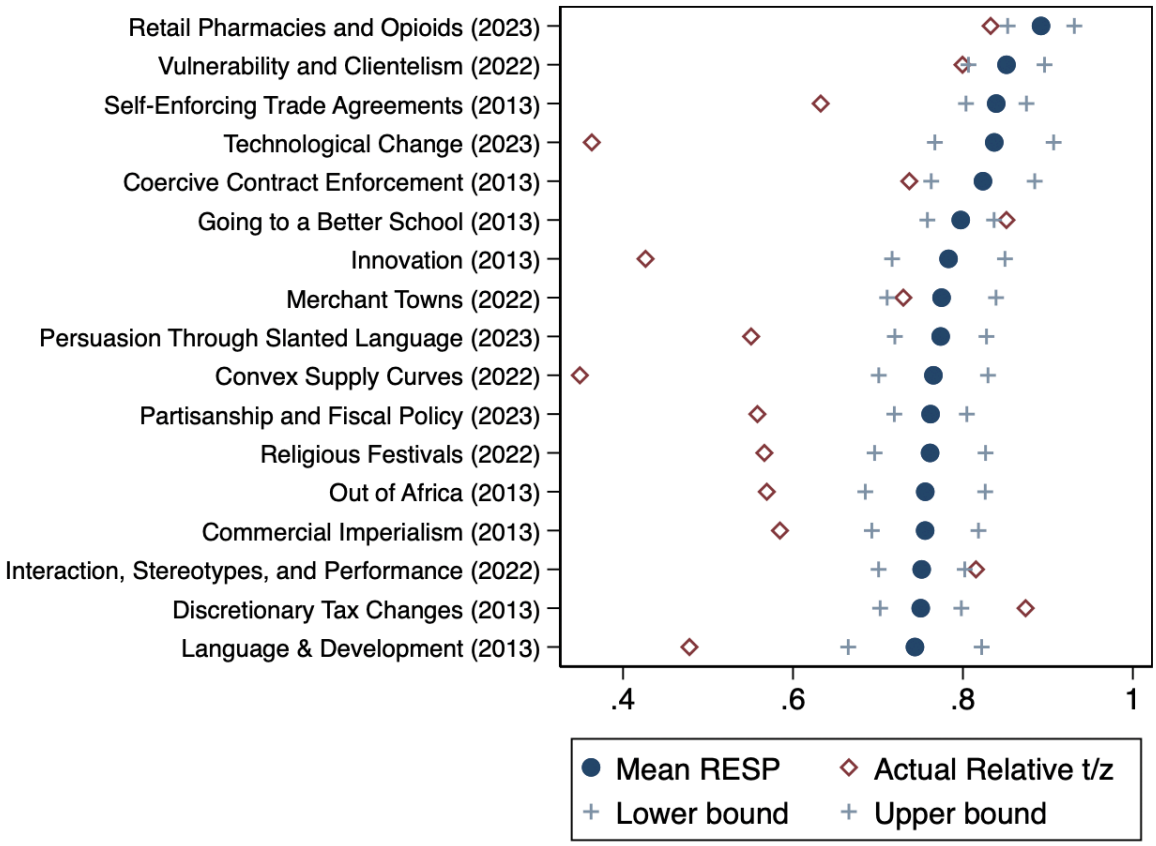


Figure 8: Mean predictions of the relative t/z-values of robustness tests per paper

Notes: Upper and lower bounds are plotted as 95% confidence intervals

Primary hypothesis 1: There is a positive association between predictions and outcomes in predicting the outcomes of robustness tests.

We test this hypothesis in two OLS regressions, one for the FSP question and one for the RESP question, with the observed value of the outcome of the robustness tests as the dependent variable and the individual prediction as the independent variable (with each individual prediction as one observation). We include participant fixed effects and cluster standard errors on the participant level to take into account the multiple observations per participant. We hypothesized a positive sign on the coefficient of the individual prediction variable. These results are reported in Table 1. Consistent with the sizable correlation above on the more aggregated level, we find a statistically significant positive association between predictions and outcomes

for predictions of the fraction of significant robustness tests (t-value=12.5; p-value<0.001). For the predictions of relative t/z-values sizes the association is much weaker, although we find suggestive evidence of a positive association (t-value=2.6; p-value=0.01).[32] It is intuitive that the association is stronger for the FSP question as the fraction of significant robustness tests is correlated with the original p-values making it easier to predict those results. We also carried out a not pre-registered robustness test of the Table 1 regressions (1) and (2); where we use two-way clustering on the respondent and the paper level. These results are reported in Appendix Table A6 and the association between predictions and outcomes remains statistically significant for the FSP question, but there is no longer suggestive evidence of an association for the RESP question. There is thus no robust evidence of an association between the RESP question and the observed relative t/z-values.

[32] The MDE is 0.1 for tests at the 5% level and 0.13 for tests at the 0.5% level for the FSP question, and 0.08 for tests at the 5% level and 0.10 for tests at the 0.5% level for the RESP question.

Table 1: Predicting Replication

	(1) FSP b/se/t/p	(2) RESP b/se/t/p
Survey	0.447	0.0722
	(0.036)	(0.028)
	[12.5]	[2.60]
	{0.000}	{0.010}
Observations	2427	2427
Respondent FEs?	Yes	Yes
Respondent clusters?	Yes	Yes

Notes: The dependent variable in the first column is the fraction of robustness tests significant for up to three findings for 17 papers. The dependent variable in column (2) is the relative effect size indicator. Both regressions include respondent fixed effects, and cluster by survey respondent. Survey sample size = 359, and each participant provided predictions for four papers. Each finding received between 77 and 95 predictions.

Primary hypothesis 2: Forecasters over/under estimate robustness.

Figures 7 and 8 suggest that forecasters on average overestimate robustness reproducibility, which is tested in Primary hypothesis 2. This primary hypothesis is also tested separately for the FSP question and the RESP question. To carry out this test we first estimate the average prediction error for each forecaster for the RESP question and the FSP question, respectively.[33] A positive value on this prediction error variable implies that the forecaster overestimated the robustness and a negative difference that the participant underestimated robustness. We test if the

[33] First averaging predictions on the paper level if more than one key result is predicted for a paper and then averaging across the four predicted papers, so that each paper is weighted equally.

prediction error differs from zero by regressing the prediction error on a constant (equivalent to a one-sample t-test). We had no directional hypothesis for this test. These results are reported in Table 2, and we find statistically significant overestimation of robustness reproducibility for both the fraction of significant robustness tests (mean prediction error 0.140; t-value=15.5; p-value<0.0001) and relative effect sizes (mean prediction error 0.157; t-value=11.2; p-value<0.0001).[34] The magnitude of the overestimation is 14.0 percentage units for the fraction of statistically significant robustness tests and 15.7 percentage units for relative t/z-values.

Table 2: Prediction Errors

	(1) FSP b/se/t/p	(2) RESP b/se/t/p
Constant	0.140	0.157
	(0.0090)	(0.014)
	[15.5]	[11.2]
	{0.000}	{0.000}
Observations	359	359

Notes: The dependent variables are the average prediction errors for the fraction significant prediction (FSP) and the relative effect size prediction (RESP) respectively in columns (1) and (2). Both variables are averaged at the participant level for 359 individuals.

One limitation of this test is that it does not take into account the uncertainty in the variable that is predicted (the observed fraction of significant robustness tests and the observed relative t/z-value). We therefore also added a not pre-registered robustness test where we test if the mean

[34] The MDE is 0.025 for tests at the 5% level and 0.033 for tests at the 0.5% level for the FSP question, and 0.039 for tests at the 5% level and 0.051 for tests at the 0.5% level for the RESP question.

forecast of the 359 forecasters differs from the mean observed reproducibility indicator in an independent samples z-test.[35] For the FSP question the mean forecast is 59.5% compared to the observed mean fraction of significant robustness tests of 45.8% and there is suggestive evidence that these means differ (z-value=2.7, p-value=0.008). For the RESP question the mean forecast is 78.7% compared to the observed mean relative t/z-value of 62.8% and there is statistically significant evidence that these means differ (z-value=3.0, p-value=0.002). This robustness test therefore also provides evidence of overestimation of robustness reproducibility, but the evidence is less strong than for the pre-registered test.

Primary hypothesis 3: Forecasters believe that the robustness of AER papers has increased over time.

As for the first two hypotheses, we test this hypothesis separately for the FSP question and the RESP question. To construct this test we first estimate the average prediction for the two papers predicted in 2013 and the average prediction for the two papers predicted in 2022/23 for each forecaster. This gives us a paired observation for each forecaster and we regress the paired difference on a constant (equivalent to a paired t-test). We hypothesized a positive difference in line with forecasters believing that robustness reproducibility has increased between the two periods. These results are reported in Table 3, and we find statistically significant evidence in support of this hypothesis for the FSP question (mean paired difference=0.032; t-value=3.74; p-value=0.0002), but not for the RESP question (mean paired difference=0.013; t-value=1.26; p-value=0.21).[36] Although the forecasters predict some improvement in robustness over time in terms of the fraction of statistically significant robustness tests, the predicted improvement of 3.2

[35] In estimating the mean forecast, we first estimate the mean forecast of each forecaster by first averaging predictions on the paper level if more than one key result is predicted for a paper and then averaging across the four predicted papers, so that each paper is weighted equally. We then have one mean forecast per forecaster and the mean of this variable is the mean forecast in the study.

[36] The MDE is 0.024 for tests at the 5% level and 0.031 for tests at the 0.5% level for the FSP question, and 0.029 for tests at the 5% level and 0.038 for tests at the 0.5% level for the RESP question.

percentage points is small, and we are not powered to detect such small improvements in actual reproducibility.

Secondary hypothesis 1: Forecasters believe that the credibility of AER papers has increased over time.

Related to primary hypothesis 3, we in secondary hypothesis 2 test if the rated credibility has increased over the investigated 10-year period. To carry out this test, we first estimate the average rated credibility of the papers published in 2013 and the papers published in 2022/23 for each forecaster. As above we then regress the paired difference on a constant, and these results are also reported in Table 3. As hypothesized, we find statistically significant evidence in support of an increase in rated credibility (mean paired difference=0.43; t-value=5.1; p-value<0.0001).[37] This implies an increase in rated credibility of 0.43 on the 0-10 scale used to rate credibility, which can be compared to the average rated credibility of 6.0 and the standard deviation of 2.1.

[37] The MDE is 0.23 for tests at the 5% level and 0.31 for tests at the 0.5% level.

Table 3: Did Survey Participants Expect Improvement Over Time?

	(1) FSP b/se/t/p	(2) RESP b/se/t/p	(3) Credibility b/se/t/p
Constant	0.0316	0.0132	0.432
	(0.0084)	(0.010)	(0.085)
	[3.74]	[1.26]	[5.06]
	{0.000}	{0.207}	{0.000}
Observations	359	359	359

Notes: The dependent variable in column (1) is the differenced average fraction of robustness tests significant prediction (FSP) where we have subtracted the average for 2013 from 2022/2023. In column (2), the dependent variable is the difference in the average of relative effect size predictions (RESP). In column (3), the dependent variable is the differenced average survey responses for how credible a paper is on a zero to ten scale. Each participant (359 total) provided predictions for up to three findings for two papers for 2013, and two for 2022/2023.

Secondary hypothesis 2: Forecasters believe that the robustness tests will result in systematically lower t/z-values than in the original papers.

A relative t/z-value below one implies that the original papers selectively report analytical specifications that exaggerate statistical significance. In this hypothesis we test if forecasters predict such systematic bias by testing if the average predicted relative t-/z-value is below one. To carry out this test, we first estimate the average response of the RESP questions for each forecaster.[38] We test our hypothesis that the mean of this variable differs from one in a one-sample t-test, and we find statistically significant evidence in support of this hypothesis (mean=0.787; t-value=15.6; p-value<0.0001).[39] This predicted inflation in t/z-values of about

[38] First averaging predictions on the paper level if more than one key result is predicted for a paper and then averaging across the four predicted papers, so that each paper is weighted equally.
[39] The MDE is 0.038 for tests at the 5% level and 0.05 for tests at the 0.5% level.

25% is in the observed direction, but it is substantially smaller than the observed inflation of about 60%.

Preregistered exploratory analyses: Are forecasts of more senior economists, economists in the same sub-field as the predicted paper, and economists more familiar with the methods in the predicted paper, different and more accurate?

In the preregistered exploratory analyses we test if being more senior, being in the same sub-field of economics as the predicted paper, and being more familiar with the methods of the predicted paper is associated with predictions and prediction accuracy. We test this in four OLS regressions with the following three independent variables: seniority of the forecaster coded as 1 for associate or full professor and as 0 for all other forecasters, being in the same sub-field as the paper being predicted coded as 1 for yes and 0 for no, and familiarity of the predicted AER paper coded as 0-10 based on the answer to that survey question. The regression is estimated with the FSP (RESP) question answer and the squared prediction error as dependent variables with each individual FSP (RESP) response as one observation. We cluster standard errors on the participant level. We hypothesized that seniority, being in the same sub-field as the paper being predicted and familiarity with the predicted AER paper would be associated with higher prediction accuracy (lower squared prediction errors and thus negative coefficients in the regression), but we did not preregister any hypothesized direction of the effect of these variables on the level of the predictions. These exploratory analyses carry little weight and should mainly be interpreted as hypothesis generating for future studies.

These results are reported in Table 4 and suggest that self-rated familiarity with the methods of the paper is associated with higher predicted robustness reproducibility for the FSP and RESP questions and lower prediction accuracy on the RESP; the lower prediction accuracy on the RESP question is in the opposite direction to our hypothesized effect. The results also suggest that senior economists make different predictions than the other forecasters on the RESP question, but not the FSP question. On the RESP question, senior economists predict lower reproducibility and their predictions are also more accurate (which is in the hypothesized direction). There is no evidence that forecasts or prediction accuracy are associated with being in

the same sub-field as the predicted paper. As for the regression equation in Table 1, we also carried out a not pre-registered robustness test of the Table 4 regressions with two-way clustering on the respondent and the paper level. These results are reported in Appendix Table A7 and the associations for self-rated familiarity and senior economists in Table 4 remain statistically significant in this robustness test.

Table 4: Subgroup Prediction Analysis

	(1) FSP b/se/t/p	(2) FSP Sq.Err b/se/t/p	(3) RESP b/se/t/p	(4) RESP Sq.Err b/se/t/p
Tenured or Associate Professor	-0.0121	0.0000443	-0.0917	-0.0449
	(0.023)	(0.0076)	(0.030)	(0.016)
	[-0.53]	[0.0058]	[-3.04]	[-2.73]
	{0.594}	{0.995}	{0.003}	{0.007}
Familiarity, 0-10 scale	0.0132	0.00176	0.0370	0.0152
	(0.0043)	(0.0016)	(0.0070)	(0.0039)
	[3.05]	[1.08]	[5.30]	[3.94]
	{0.002}	{0.279}	{0.000}	{0.000}
Field match	0.0143	0.000129	0.00593	0.00995
	(0.015)	(0.0079)	(0.025)	(0.018)
	[0.95]	[0.016]	[0.23]	[0.56]
	{0.341}	{0.987}	{0.815}	{0.576}
Constant	0.530	0.0865	0.602	0.0966
	(0.028)	(0.0099)	(0.041)	(0.020)
	[19.2]	[8.73]	[14.6]	[4.81]
	{0.000}	{0.000}	{0.000}	{0.000}
Observations	2427	2427	2427	2427

Notes: The dependent variable in column (1) is the fraction of robustness tests significant prediction (FSP) for up to three findings for 17 papers, with one observation for each prediction. The dependent variable in column (2) is the squared prediction error for the FSP question. Column (3) is the relative effect size prediction (RESP). Column (4) is the squared error for the RESP. The survey included 359 participants, and each participant provided predictions for four papers. Each finding within a paper received between 77 and 95 predictions. Standard errors are clustered at the participant level.

III. Concluding remarks

The primary aim of this study is to document robustness reproducibility rates for several studies published in the leading journal of the American Economic Association. We find that slightly more than half of our robustness tests lead to non-significant results ($p>0.05$) with an average mean relative t/z-value of 63%. While our distribution of p-values is certainly different from the original papers, is it more indicative of the truth? Note that (1) the fact that the 29 results in our study were essentially robust in 100% of cases in the original papers' direct robustness checks is likely indicative of selective reporting, (2) a relatively standard influential analysis exercise yielded insignificant results in one-third of cases, despite using the authors' own preferred specifications, (3) survey participants could to some extent intuitively predict which studies we would find to be robust.[40] Anecdotal evidence for our selective reporting thesis is that one of the author teams described to us via email that they had initially tried an alternative specification implemented in our robustness checks, but dropped it because it was not statistically significant.[41]

These findings suggest selective reporting of analytical specifications that exaggerate effect sizes and statistical significance in our sample of studied articles. Our robustness indicator results are based on weighting all the robustness tests equally, and in reality some robustness tests may be perceived as more important than others but it is non-trivial to introduce a weighting scheme that deviates from equal weighting. The robustness indicators should thus only be viewed as proxies for robustness reproducibility and could also be complemented by additional information such as subjectively rated robustness reproducibility also taking into account potential additional issues not addressed by the robustness tests. Further work is needed on this and our reproducibility reports posted at our project repository at OSF (https://osf.io/w7vpu/) also provide more detailed information of the robustness tests of each paper.

[40] Exceptions here are several insignificant robustness checks from Jansen and Zhang (2023) which we believe were typesetting errors, and several other p-values that were just above the 5% statistical significance threshold. The smallest valid absolute t/z-value in a direct robustness check for the 29 results in our study was 1.75 with an implied p-value of 0.08.

[41] Note that there are also reasons to believe that we may be overestimating the robustness of the results. For example, our methodology assumes that the identification approaches in the original papers are perfectly valid, whereas in some of our robustness reports we discuss reasons and present evidence why some assumptions underlying the identification approaches used in the original papers may not be valid. We are also only testing results in-sample: out-of-sample tests on new data may yield different results. This can also be true of conceptual replications.

Both our primary robustness indicators are continuous rather than binary indicators and they will not automatically classify key results for papers as robust or non-robust but such a classification would require defining cut-offs for each indicator. If we for instance define high robustness reproducibility as a value over 75% on these indicators, intermediate robustness reproducibility as 50-75%, and low robustness reproducibility as below 50%; then four out of the 29 key results are classified as having high robustness reproducibility on both indicators and one of the 17 papers when the indicators are aggregated across key results. Eleven key results and seven papers are classified as having intermediate or high robustness reproducibility on both indicators, and 18 key results and 10 papers are classified as having low robustness reproducibility on at least one of the two indicators.

We conduct two additional exercises. We first compare studies across the two time periods, finding no improvement in robustness reproducibility over time. This result is consistent with no changes in p-hacking and publication bias over time documented in Brodeur et al. (2020). Second, we conduct a survey in which we ask PhD students and economists with a PhD to predict the outcomes of robustness tests. There is a strong positive association between predictions and the fraction of statistically significant robustness tests, but only a weak positive association between predictions and the relative t/z-value of the robustness tests. It is intuitive that it is easier to predict the fraction of significant robustness tests as that robustness indicator is correlated with the original p-value. Forecasters on average overestimated the robustness reproducibility by about 15 percentage units, and were thus overly optimistic about the robustness of the included papers. Forecasters expected slightly higher credibility and robustness for papers published more recently, although no such trend towards increased robustness reproducibility over time was observed in the actual robustness tests. Forecasters on average rated the credibility of the included 17 AER papers as 6.0 on a 0 (not at all credible) to 10 (very credible) scale, which seems disappointingly low for a flagship journal.

Though our study is limited to a single journal, our results are likely generalizable to other leading outlets publishing empirical papers for at least two reasons. First, the editorial board at the AER has completely changed from 2013 to 2022 and several editors handled those 17 papers.

Second, we included all empirical papers published in a 6-month period in 2013 and in a 6-month period in 2022/23. Our choice for the 17 papers was thus not based on potential determinants of robustness reproducibility, but rather on methods and data availability. By necessity we only included papers with available data and working code, and we cannot rule out that the robustness reproducibility differs for papers without posted data and code; but we would be surprised if the robustness reproducibility is higher in this sub-group of papers. Brodeur et al. (forthcoming) found no evidence of a difference in p-hacking depending on data-sharing policy.

Further investigation is needed to illuminate the issues discussed here by delving into the reproducibility and replicability of research published in lower ranked journals and journals without a data editor. Additionally, our research does not deal with experiments, structural estimation and calibration exercises, which could be the focus of future research. Nevertheless, our findings indicate that, among the papers published in a leading economic journal, the findings are often not robust to sensitivity analysis and reproducibility rates are not increasing over time.

References

Aghion P, Van Reenen J, Zingales L. Innovation and Institutional Ownership. American Economic Review 2013; 103(1):277-304.

Andrews I, Kasy M. Identification of and Correction for Publication Bias. American Economic Review 2019; 109(8):2766-94.

Andrews I, Stock JH, Sun L. Weak Instruments in Instrumental Variables Regression: Theory and practice. Annual Review of Economics. 2019; 11:727-53.

Angelucci C, Meraglia S, Voigtländer N. How Merchant Towns Shaped Parliaments: From the Norman Conquest of England to the Great Reform Act. American Economic Review 2022; 112(10):3441-3487.

Ankel-Peters J, Fiala N, Neubauer F. Do Economists Replicate? Journal of Economic Behavior & Organization 2023; 212:219-232.

Ashraf Q, Galor O. The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development. American Economic Review 2013; 103 (1):1-46.

Athey S, Imbens G. A Measure of Robustness to Misspecification. American Economic Review: Papers & Proceedings 2015; 105:476-480.

Belsley, D.A., E. Kuh and R.E. Welsch (1980), Regression Diagnostics (Wiley: New York).

Benjamin DJ, et al. Redefine statistical significance. Nature Human Behaviour 2018; 2:6-10.

Berger D, Easterly W, Nunn N, Satyanath S. Commercial Imperialism? Political Influence and Trade during the Cold War. American Economic Review 2013; 103(2):863-896.

Bobonis GJ, Gertler PJ, Gonzalez-Navarro M, Nichter S. Vulnerability and Clientelism. American Economic Review 2022; 112(11):3627-3659.

Boehm CE, Pandalai-Nayar N. Convex Supply Curves. American Economic Review 2022; 112(12):3941-3969.

Botvinik-Nezer R, et al. Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams. *Nature* 2020; 582:84–88.

Bown CP, Crowley MA. Self-Enforcing Trade Agreements: Evidence from Time-Varying Trade Policy. American Economic Review 2013; 103(2):1071-1090.

Brodeur A, Carrell S, Figlio D, Lusher L. Unpacking p-Hacking and Publication Bias. American Economic Review. 2023; 113(11):2974-3002.

Brodeur A, Cook N, Heyes A. Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. American Economic Review 2020; 110:3634-3660.

Brodeur A, Cook N, Neisser C. P-Hacking, Data Type, and Data-Sharing Policy. Economic Journal, forthcoming.

Brodeur A, Dreber A, Hoces de la Guardia F, Miguel E. Replication Games: How to Make Reproducibility Research More Systematic. Nature 2023; 621:684-686.

Brodeur A, Lé M, Sangnier M, Zylberberg Y. Star Wars: the Empirics Strikes Back. American Economic Journal: Applied 2016: 8:1-32.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafo MR. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. Nature Reviews Neuroscience 2013; 14:365-376.

Camerer CF, et al. Evaluating Replicability of Laboratory Experiments in Economics. Science 2016; 351:1433-1436.

Camerer CF, et al. Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015. Nature Human Behaviour 2018;2:637-644.

Carlino G, Drautzburg T, Inman R, & Zarra N. Partisanship and Fiscal Policy in Economic Unions: Evidence from US States. American Economic Review 2023; 113(3):701-737.

Carter BJ, Taska B. Technological Change and the Consequences of Job Loss. American Economic Review 2023; 113(2): 279-316.

Chang AC, Li P. A Preanalysis Plan to Replicate Sixty Economics Research Papers that Worked Half of the Time. American Economic Review Papers and Proceedings 2017; 107:60-64.

Chen MK. The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets. American Economic Review 2013; 103(2):690-731.

Christensen G, Miguel E. Transparency, Reproducibility, and the Credibility of Economics Research. Journal of Economic Literature 2018; 56:920-980.

Cloyne J. Discretionary Tax Changes and the Macroeconomy: New Narrative Evidence from the United Kingdom. American Economic Review 2013; 103(4):1507-1528.

Corno L, La Ferrare E, Burns J. Interaction, Stereotypes, and Performance: Evidence from South Africa. American Economic Review 2022; 112(12):3848-3875.

DellaVigna S, Linos E. RCTs to Scale: Comprehensive Evidence from Two Nudge Units. Econometrica 2022; 90(1):81-116.

DellaVigna S, Pope D. Predicting Experimental Results: Who Knows What? Journal of Political Economy 2018; 126(6):2410-2456.

Dewald WG, Thursby J, Anderson R. Replication in Empirical Economics: Journal of Money, Credit and Banking Project. American Economic Review 1986; 76:587-603.

Djourelova M. Persuasion through Slanted Language: Evidence from the Media Coverage of Immigration. American Economic Review 2023; 113(3):800-835.

Doucouliagos C, Stanley TD. Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity. Journal of Economic Surveys 2013; 27(2):316-39.

Dreber A, Johannesson M. A Framework for Evaluating Reproducibility and Replicability in Economics. SSRN working paper, 2023.

Dreber A, Johannesson M, Yang Y. Selective Reporting of Placebo Tests in Top Economics Journals. Economic Inquiry, forthcoming, 2024.

Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, Nosek BA, Johannesson M. Using Prediction Markets to Estimate the Reproducibility of Scientific Research. PNAS 2015; 112:15343-15347.

Ebersole CR, et al. Many Labs 3: Evaluating Participant Pool Quality Across the Academic Semester via Replication. Journal of Experimental Social Psychology 2016; 67:68-82.

Franco A., Malhotra N, Simonovits G. Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science* 2014; 345:1502-1505.

Franco A, Malhotra N & Simonovits G. Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results. Political Analysis *2015;*23:306-312.

Gelman A & Loken E. The Statistical Crisis in Science. American Scientist 2014; 102:460-465.

Gertler P, Galiani S, Romero M. How to make replication the norm. Nature 2018; 554:417-419.

Glandon PJ. Appendix to the Report of the Editor: Report on the American Economic Review data availability compliance project. American Economic Review Papers & Proceedings 2011;101:695-699.

Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T. Predicting Replicability: Analyses of Survey and Prediction Market Data from Large-Scale Forecasting Projects. PLoS ONE 2021; 16;e0248780.

Havranek T, Irsova Z, Laslopova L, Zeynalova O. Publication and Attenuation Biases in Measuring Skill Substitution. Review of Economics and Statistics, forthcoming.

Hedges LV. Modeling Publication Selection Effects in Meta-Analysis. Statistical Science 1992; 7:246-255.

Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Medicine 2005; 2:e124.

Ioannidis JPA, Stanley TD, Doucouliagos H. The Power of Bias in Economics Research. Economic Journal 2017; 127:F236-F265.

Jansen A, Zhang X. Retail Pharmacies and Drug Diversion During the Opioid Epidemic. American Economic Review 2023; 11(1):1-33.

John LK, Loewenstein G & Prelec D. Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling. Psychological Science 2012; 23:524-532.

Johnson VE, Payne RD, Wang T, Asher A, Mandal S. On the Reproducibility of Psychological Science. Journal of the American Statistical Association 2017:112:1-10.

Klein RA, et al. Investigating Variation in Replicability: A “Many Labs” Replication Project. Social Psychology 2014: 45:142-152.

Klein RA, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances in Methods and Practices in Psychological Science 2018; 1:443-490.

Leamer EE. Let’s Take the Con Out of Econometrics. American Economic Review 1983; 73:31-43.

Lewis, Daniel J., and Karel Mertens. "A Robust Test for Weak Instruments with Multiple Endogenous Regressors," 2022.

Maniadis Z, Tufano F, List JA. One Swallow Doesn’t Make a Summer: New Evidence of Anchoring Effects. American Economic Review 2014; 104(1):277-290.

McCullough BD, McGeary KA, Harrison TD. Lessons from the JMCB archive. Journal of Money, Credit and Banking 2006; 38:1093-1107.

McCullough BD, McGeary KA, Harrison TD. Do Economics Journal Archives Promote Replicable Research? Canadian Journal of Economics 2008; 41:1406-1420.

Menkveld, A. *et al*. Non-Standard Errors. Journal of Finance, forthcoming 2024.

Milkman KL, Gromet D, Ho H, Kay JS, Lee TW, Pandiloski P, Park Y, Rai A, Bazerman M, Beshears J, Bonacorsi L. Megastudies Improve the Impact of Applied Behavioural Science. Nature. 2021; 600(7889):478-83.

Montero E, Yang D. Religious Festivals and Economic Development: Evidence from the Timing of Mexican Saint Day Festivals. American Economic Review 2022; 112(10):3176-3214.

Naidu S, Yuchtman N. Coercive Contract Enforcement: Law and the Labor Market in Nineteenth Century Industrial Britain. American Economic Review 2013; 103(1):107-144.

Open Science Collaboration. Estimating the Reproducibility of Psychological Science. Science 2015; 349:aac4716.

Perignon C, Akmansoy O, Hurlin C, Dreber A, Holzmeister F, Huber J, Johannesson M, Kirchler M, Menkveld AJ, Razen M, Weitzel U. Computational Reproducibility in Finance: Evidence from 1,000 tests. SSRN Working paper, 2023.

Pop-Eleches C, Ugquiola M. Going to a Better School: Effects and Behavioral Responses. American Economic Review 2013:103(4):1289-1324.

Silberzahn et al. Many Analysts, One Dataset: Making Transparent How Variation in analytical choices affect results. Advances in Methods and Practices in Psychological Science 2018;1:337-356.

Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 2011; 22:1359-1366.

Simonsohn U, Simmons JP, Nelson LD. Specification Curve Analysis. Nature Human Behaviour 2020; 4:1208-1214.

Stanley TD, Doucouliagos H. Meta‐Regression Approximations to Reduce Publication Selection Bias. Research Synthesis Methods 2014; 5(1):60-78.

Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency through a Multiverse Analysis. Perspectives on Psychological Science 2016; 11:702-712.

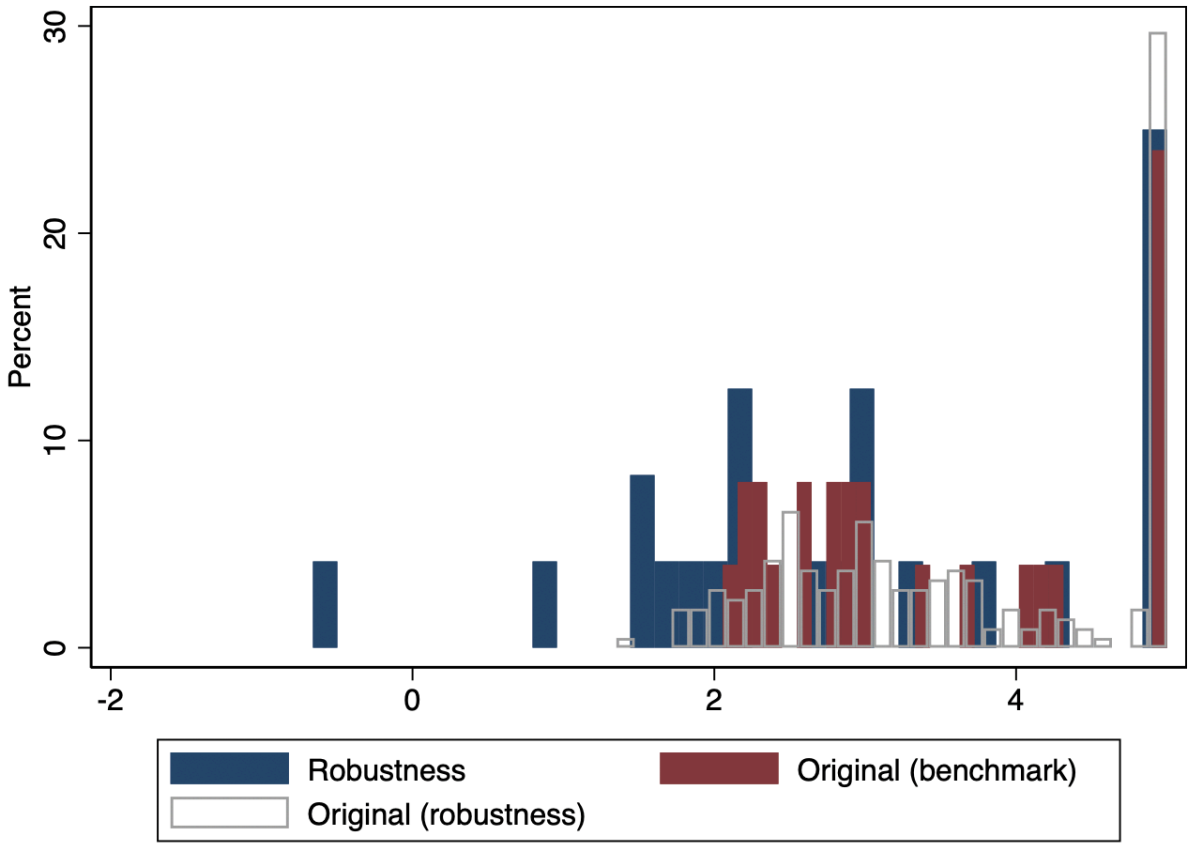
Stern JM, Simes RJ. Publication Bias: Evidence of Delayed Publication in a Cohort of Clinical Research Projects. *BMJ* 1997; 315:640-645.

Vilhuber, Lars, James Turrito, and Keesler Welch. 2020. Report by the AEA Data Editor. AEA Papers and Proceedings, 110; 764-75.

Vivalt E. Specification Searching and Significance Inflation Across Time, Methods and Disciplines. Oxford Bulletin of Economics and Statistics 2019; 81(4):797-816.

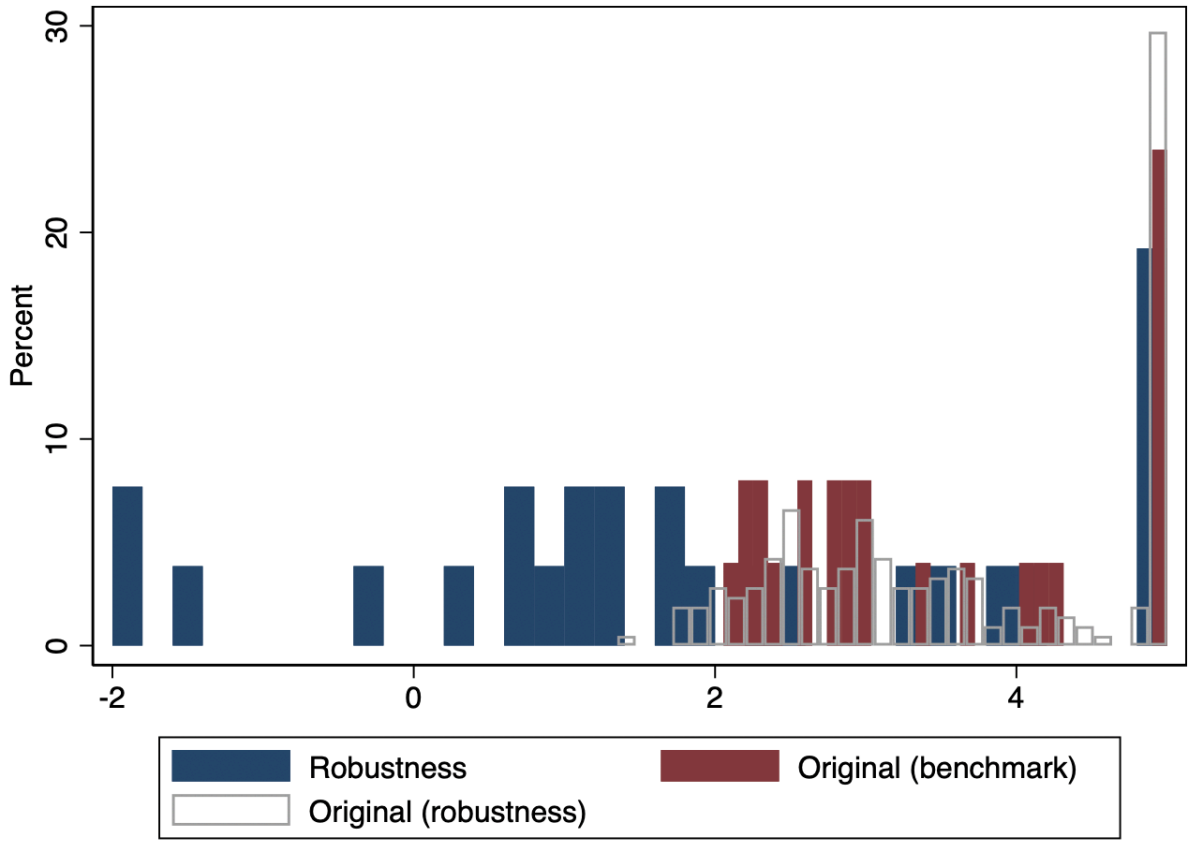
Young A. Consistency Without Inference: Instrumental Variables in Practical Application. European Economic Review 2022; 147:104112.

Appendix:



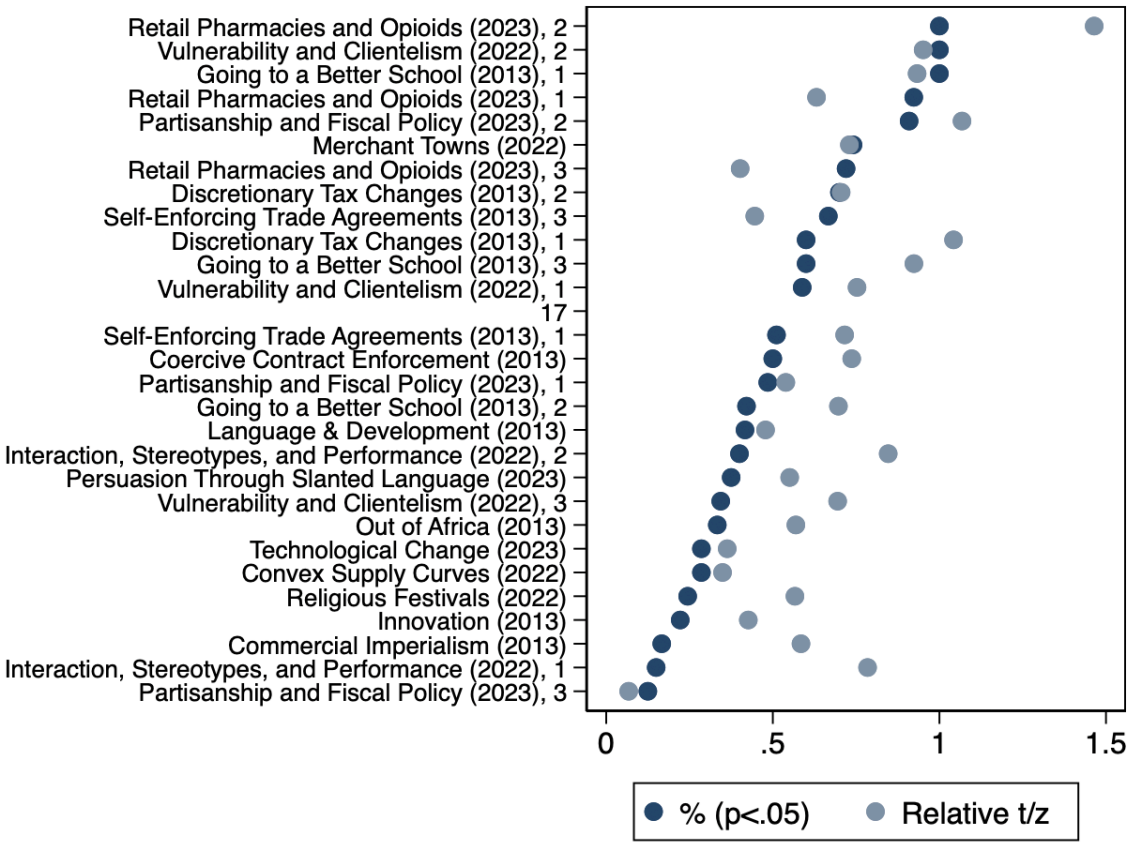
Appendix Figure 1: T/z-value distribution from influential analysis on benchmark specifications

Notes: The robustness checks in blue use the authors' benchmark regression specifications exactly, only remove influential observations with larger absolute dfbeta statistics than the standard cutoff.

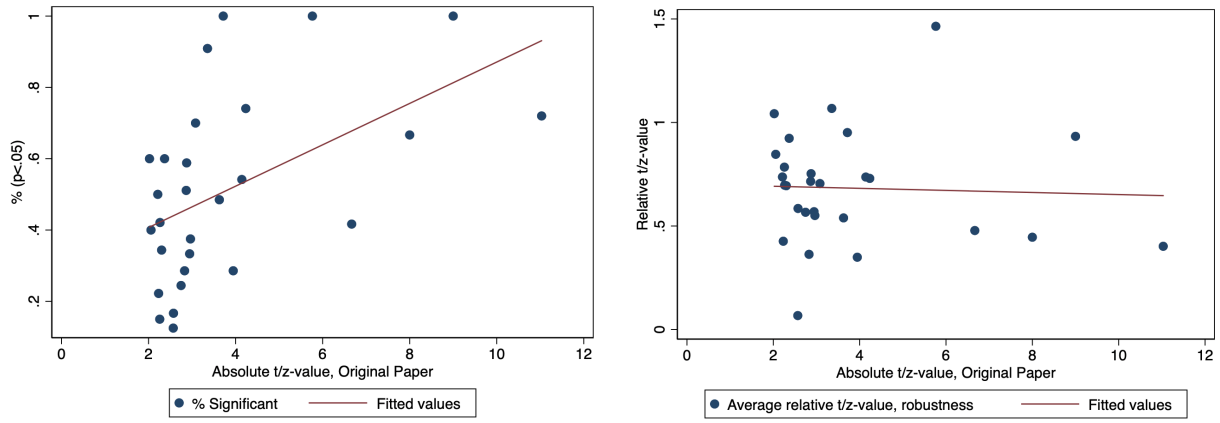


Appendix Figure 2: T/z-value distribution from influential analysis on alternative specifications

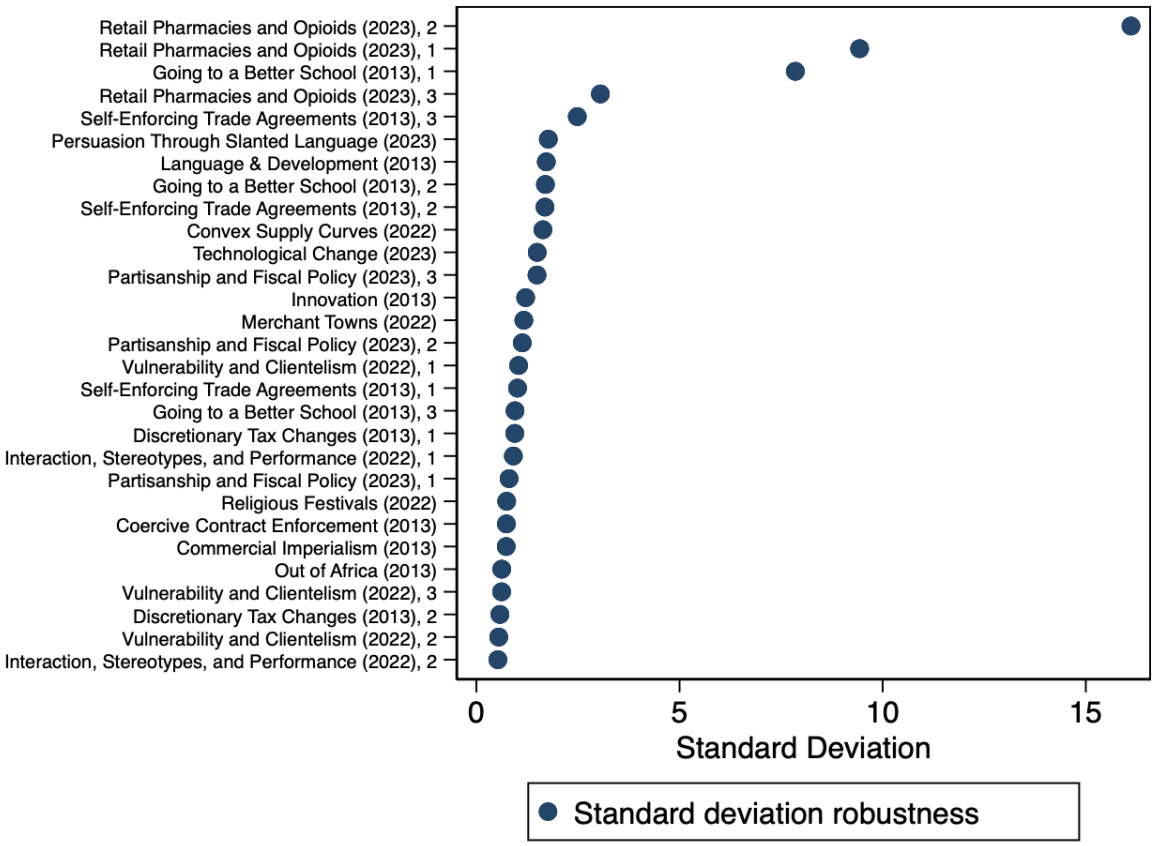
Notes: The robustness checks in blue use our alternative regression specifications, and remove influential observations with larger absolute dfbeta statistics than the standard cutoff.



Appendix Figure 3: Statistical significance and relative t/z-values by result

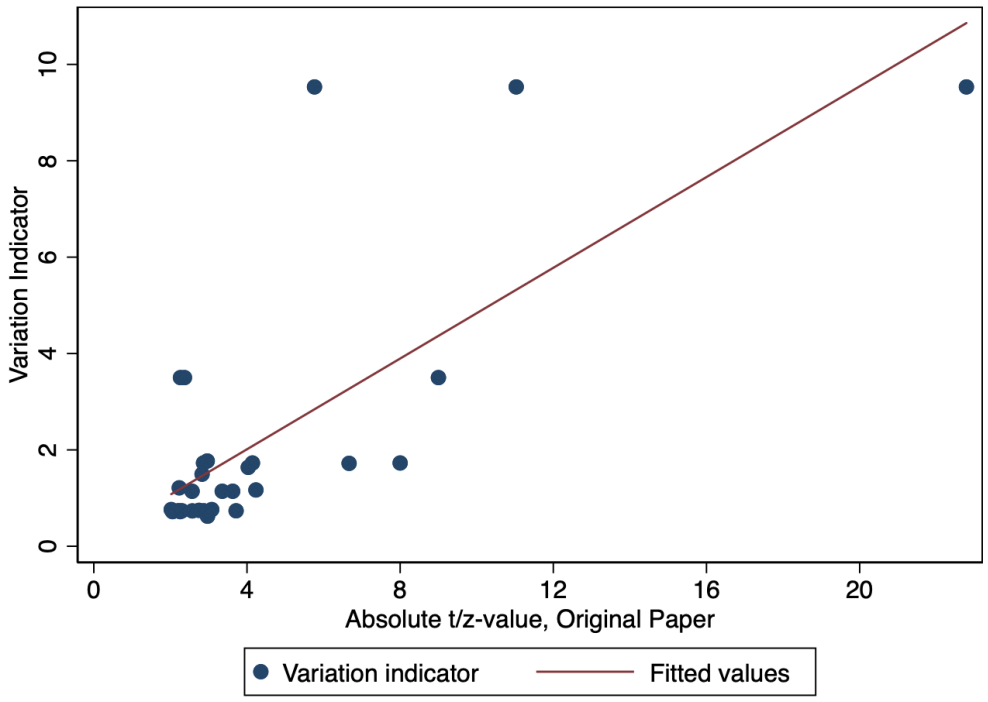


Appendix Figure 4: Original absolute t/z values vs. robustness indicators, results level

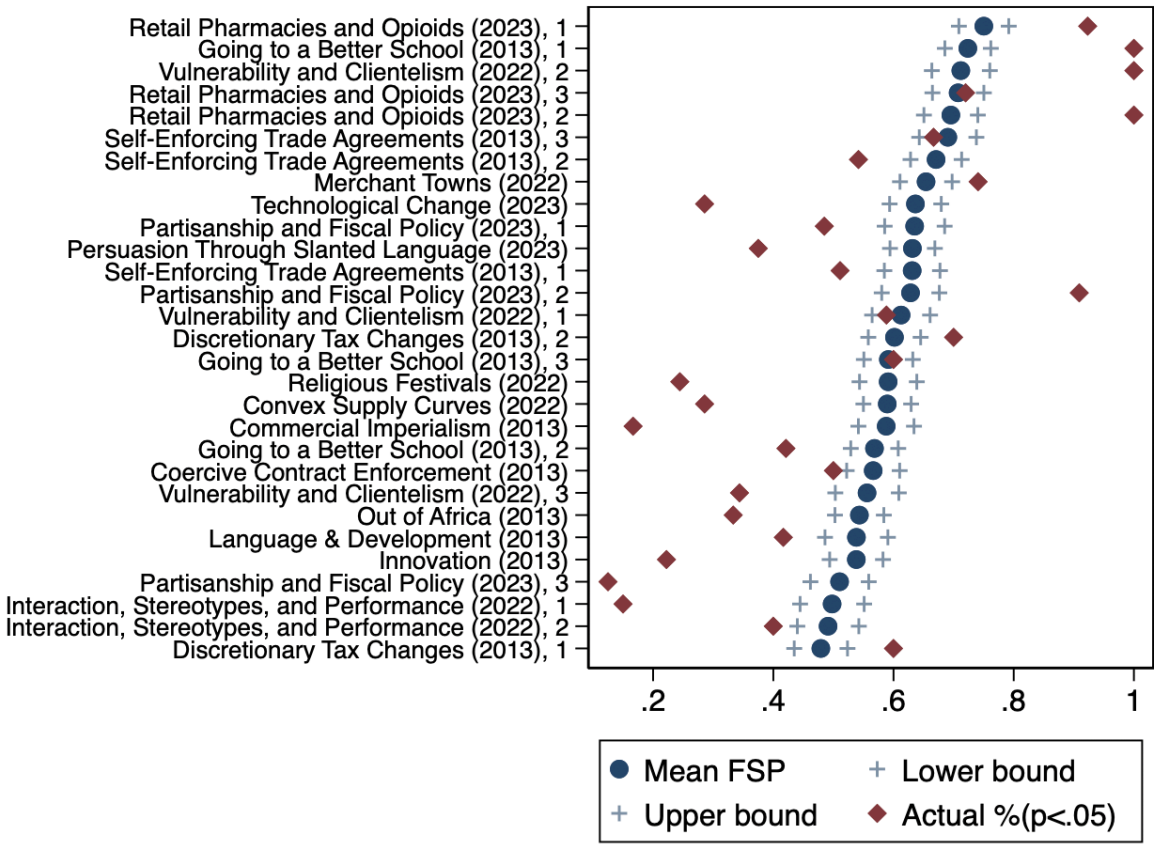


Appendix Figure 5: Variation indicator by finding

Notes: This plots the standard deviation of the t/z-values of the robustness tests for up to three findings for each paper.

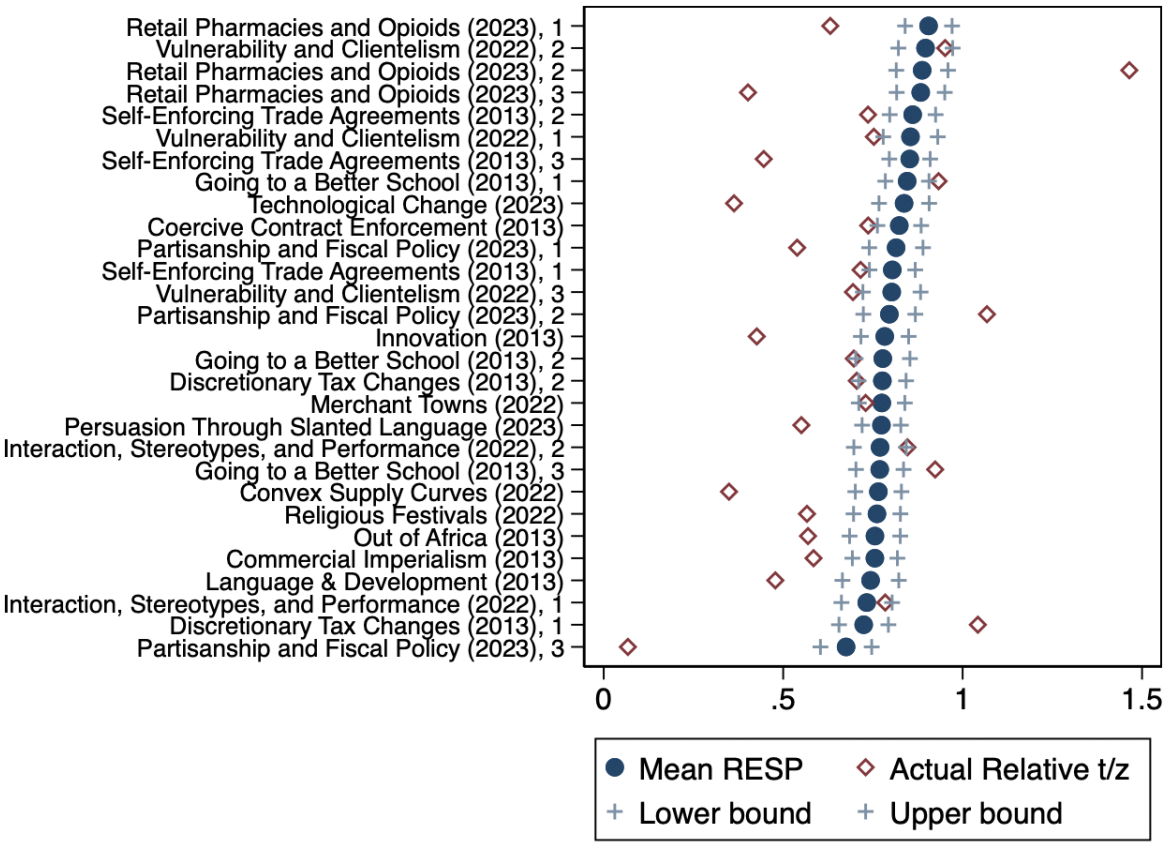


Appendix Figure 6: Relation variation indicator and initial t/z-values



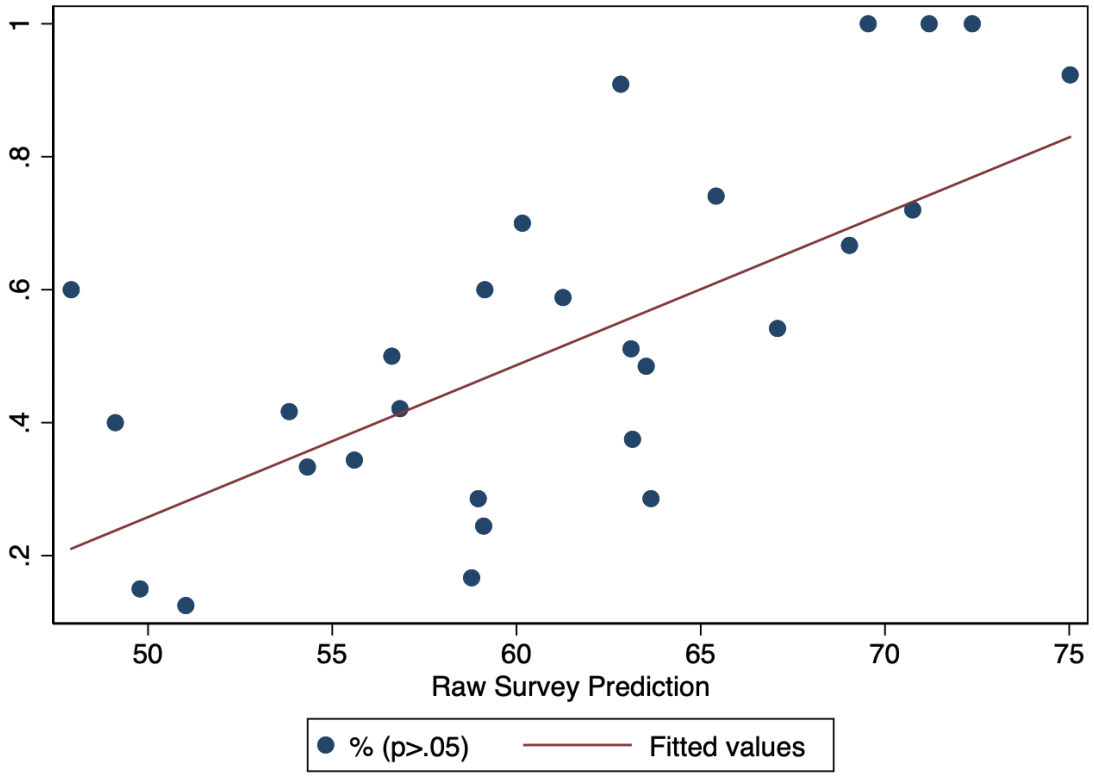
Appendix Figure 7: Mean predictions of the fraction of statistically significant robustness tests per predicted key result

Notes: Upper and lower bounds are plotted as 95% confidence intervals.



Appendix Figure 8: Mean predictions of the fraction of the relative t/z-value of robustness tests per predicted key result

Notes: Upper and lower bounds are plotted as 95% confidence intervals



Appendix Figure 9: Predicting replication

Note: This graph displays survey predictions averaged at the finding level vs. the fraction significant indicator.

Table A1: List of Included Studies

Paper	Year
1. The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development	2013
2. Coercive Contract Enforcement: Law and the Labor Market in Nineteenth Century Industrial Britain	2013
3. Innovation and Institutional Ownership	2013
4. The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets	2013
5. Commercial Imperialism? Political Influence and Trade during the Cold War	2013
6. Self-Enforcing Trade Agreements: Evidence from Time-Varying Trade Policy	2013
7. Going to a Better School: Effects and Behavioral Responses	2013
8. Discretionary Tax Changes and the Macroeconomy: New Narrative Evidence from the United Kingdom	2013
9. Interaction, Stereotypes, and Performance: Evidence from South Africa	2022
10. Convex Supply Curves	2022
11. Vulnerability and Clientelism	2022
12. Religious Festivals and Economic Development: Evidence from the Timing of Mexican Saint Day Festivals	2022
13. How Merchant Towns Shaped Parliaments: From the Norman Conquest of England to the Great Reform Act	2022
14. Technological Change and the Consequences of Job Loss	2023
15. Persuasion through Slanted Language: Evidence from the Media Coverage of Immigration	2023
16. Retail Pharmacies and Drug Diversion during the Opioid Epidemic	2023
17. Partisanship and Fiscal Policy in Economic Unions: Evidence from US States	2023

Table A2: Fraction Significant Prediction Summary Statistics, by Paper

Paper	Mean	Std.Error	Std.Dev.	UB	LB
Retail Pharmacies and Opioids (2023)	0.72	0.01	0.19	0.74	0.69
Self-Enforcing Trade Agreements (2013)	0.66	0.01	0.21	0.69	0.64
Merchant Towns (2022)	0.65	0.02	0.2	0.7	0.61
Technological Change (2023)	0.64	0.02	0.2	0.68	0.59
Persuasion Through Slanted Language (2023)	0.63	0.02	0.17	0.67	0.59
Going to a Better School (2013)	0.63	0.01	0.2	0.65	0.6
Vulnerability and Clientelism (2022)	0.63	0.02	0.24	0.66	0.6
Partisanship and Fiscal Policy (2023)	0.59	0.01	0.22	0.62	0.56
Religious Festivals (2022)	0.59	0.02	0.22	0.64	0.54
Convex Supply Curves (2022)	0.59	0.02	0.18	0.63	0.55
Commercial Imperialism (2013)	0.59	0.02	0.23	0.63	0.54
Coercive Contract Enforcement (2013)	0.57	0.02	0.2	0.61	0.52
Out of Africa (2013)	0.54	0.02	0.2	0.58	0.5
Discretionary Tax Changes (2013)	0.54	0.02	0.23	0.57	0.51
Language & Development (2013)	0.54	0.03	0.25	0.59	0.49
Innovation (2013)	0.54	0.02	0.22	0.58	0.49
Interaction, Stereotypes... (2022)	0.49	0.02	0.24	0.53	0.46

Notes: This table lists the mean, standard error, standard deviation, and 95% confidence interval upper and lower bounds for the fraction significant predictions for all 17 papers and 359 forecasters.

Table A3: Relative Effect Size Prediction Summary Statistics, by Paper

Paper	Mean	Std.Error	Std.Dev.	UB	LB
Retail Pharmacies and Opioids (2023)	0.89	0.02	0.3	0.93	0.85
Vulnerability and Clientelism (2022)	0.85	0.02	0.35	0.9	0.81
Self-Enforcing Trade Agreements (2013)	0.84	0.02	0.29	0.87	0.8
Technological Change (2023)	0.84	0.04	0.32	0.91	0.77
Coercive Contract Enforcement (2013)	0.82	0.03	0.28	0.88	0.76
Going to a Better School (2013)	0.8	0.02	0.33	0.84	0.76
Innovation (2013)	0.78	0.03	0.32	0.85	0.72
Merchant Towns (2022)	0.77	0.03	0.3	0.84	0.71
Persuasion Through Slanted Language (2023)	0.77	0.03	0.25	0.83	0.72
Convex Supply Curves (2022)	0.77	0.03	0.29	0.83	0.7
Partisanship and Fiscal Policy (2023)	0.76	0.02	0.33	0.8	0.72
Religious Festivals (2022)	0.76	0.03	0.3	0.83	0.7
Out of Africa (2013)	0.76	0.04	0.35	0.83	0.69
Commercial Imperialism (2013)	0.76	0.03	0.31	0.82	0.69
Interaction, Stereotypes... (2022)	0.75	0.03	0.33	0.8	0.7
Discretionary Tax Changes (2013)	0.75	0.02	0.33	0.8	0.7
Language & Development (2013)	0.74	0.04	0.38	0.82	0.67

Notes: This table lists the mean, standard error, standard deviation, and 95% confidence interval upper and lower bounds for the fraction significant predictions for all 17 papers and 359 forecasters.

Table A4: Fraction Significant Predictions, by Result

Paper and Result	Mean	Std.Error	Std.Dev.	UB	LB
Retail Pharmacies and Opioids (2023), 1	0.91	0.03	0.19	0.97	0.84
Vulnerability and Clientelism (2022), 2	0.9	0.04	0.22	0.97	0.82
Retail Pharmacies and Opioids (2023), 2	0.89	0.04	0.2	0.96	0.82
Retail Pharmacies and Opioids (2023), 3	0.88	0.03	0.19	0.95	0.82
Self-Enforcing Trade Agreements (2013), 2	0.86	0.03	0.2	0.92	0.8
Vulnerability and Clientelism (2022), 1	0.85	0.04	0.22	0.93	0.78
Self-Enforcing Trade Agreements (2013), 3	0.85	0.03	0.22	0.91	0.8
Going to a Better School (2013), 1	0.85	0.03	0.19	0.91	0.78
Technological Change (2023)	0.84	0.04	0.2	0.91	0.77
Coercive Contract Enforcement (2013)	0.82	0.03	0.2	0.88	0.76
Partisanship and Fiscal Policy (2023), 1	0.81	0.04	0.22	0.89	0.74
Self-Enforcing Trade Agreements (2013), 1	0.8	0.03	0.22	0.87	0.74
Vulnerability and Clientelism (2022), 3	0.8	0.04	0.24	0.88	0.72
Partisanship and Fiscal Policy (2023), 2	0.8	0.04	0.21	0.87	0.72
Innovation (2013)	0.78	0.03	0.22	0.85	0.72
Going to a Better School (2013), 2	0.78	0.04	0.19	0.85	0.7
Merchant Towns (2022)	0.77	0.03	0.2	0.84	0.71
Persuasion Through Slanted Language (2023)	0.77	0.03	0.17	0.83	0.72
Interaction, Stereotypes... (2022), 2	0.77	0.04	0.23	0.84	0.7
Going to a Better School (2013), 3	0.77	0.03	0.2	0.84	0.7
Convex Supply Curves (2022)	0.77	0.03	0.18	0.83	0.7
Religious Festivals (2022)	0.76	0.03	0.22	0.83	0.7
Out of Africa (2013)	0.76	0.04	0.2	0.83	0.69
Commercial Imperialism (2013)	0.76	0.03	0.23	0.82	0.69
Language & Development (2013)	0.74	0.04	0.25	0.82	0.67
Interaction, Stereotypes... (2022), 1	0.73	0.04	0.24	0.8	0.66
Discretionary Tax Changes (2013), 2	0.72	0.04	0.22	0.79	0.66
Partisanship and Fiscal Policy (2023), 3	0.68	0.04	0.22	0.75	0.6

Notes: This table lists the mean, standard error, standard deviation, and 95% confidence interval upper and lower bounds for the fraction significant predictions for all 29 results and 359 forecasters.

Table A5: Relative Effect Size Predictions, by Result

Paper and Result	Mean	Std.Error	Std.Dev.	UB	LB
Retail Pharmacies and Opioids (2023), 1	0.91	0.03	0.29	0.97	0.84
Vulnerability and Clientelism (2022), 2	0.9	0.04	0.35	0.97	0.82
Retail Pharmacies and Opioids (2023), 2	0.89	0.04	0.32	0.96	0.82
Retail Pharmacies and Opioids (2023), 3	0.88	0.03	0.3	0.95	0.82
Self-Enforcing Trade Agreements (2013), 2	0.86	0.03	0.3	0.92	0.8
Vulnerability and Clientelism (2022), 1	0.85	0.04	0.35	0.93	0.78
Self-Enforcing Trade Agreements (2013), 3	0.85	0.03	0.26	0.91	0.8
Going to a Better School (2013), 1	0.85	0.03	0.3	0.91	0.78
Technological Change (2023)	0.84	0.04	0.32	0.91	0.77
Coercive Contract Enforcement (2013)	0.82	0.03	0.28	0.88	0.76
Partisanship and Fiscal Policy (2023), 1	0.81	0.04	0.34	0.89	0.74
Self-Enforcing Trade Agreements (2013), 1	0.8	0.03	0.3	0.87	0.74
Vulnerability and Clientelism (2022), 3	0.8	0.04	0.37	0.88	0.72
Partisanship and Fiscal Policy (2023), 2	0.8	0.04	0.32	0.87	0.72
Innovation (2013)	0.78	0.03	0.32	0.85	0.72
Going to a Better School (2013), 2	0.78	0.04	0.37	0.85	0.7
Discretionary Tax Changes (2013), 2	0.78	0.03	0.33	0.84	0.71
Merchant Towns (2022)	0.77	0.03	0.3	0.84	0.71
Persuasion Through Slanted Language (2023)	0.77	0.03	0.25	0.83	0.72
Interaction, Stereotypes... (2022), 2	0.77	0.04	0.33	0.84	0.7
Going to a Better School (2013), 3	0.77	0.03	0.32	0.84	0.7
Convex Supply Curves (2022)	0.77	0.03	0.29	0.83	0.7
Religious Festivals (2022)	0.76	0.03	0.3	0.83	0.7
Out of Africa (2013)	0.76	0.04	0.35	0.83	0.69
Commercial Imperialism (2013)	0.76	0.03	0.31	0.82	0.69
Language & Development (2013)	0.74	0.04	0.38	0.82	0.67
Interaction, Stereotypes... (2022), 1	0.73	0.04	0.32	0.8	0.66
Partisanship and Fiscal Policy (2023), 3	0.68	0.04	0.32	0.75	0.6

Notes: This table lists the mean, standard error, standard deviation, and 95% confidence interval upper and lower bounds for the fraction significant predictions for all 29 results and 359 forecasters.

Table A6: Predicting Replication: Two-way Clustering

	(1) FSP b/se/t/p	(2) RESP b/se/t/p
Survey	0.447	0.0722
	(0.099)	(0.062)
	[4.50]	[1.17]
	{0.000}	{0.261}
Observations	2427	2427
Respondent FEs?	Yes	Yes
Respondent clusters?	Yes	Yes

Notes: The dependent variable in the first column is the fraction of robustness tests significant for up to three findings for 17 papers (29 findings total). The dependent variable in column (2) is the relative effect size indicator. Both regressions include respondent fixed effects, and cluster by survey respondent and by paper. Survey sample size = 359, and each participant provided predictions for four papers. Each finding received between 77 and 95 predictions.

Table A7: Subgroup Prediction Analysis: Two-way Clustering

	(1) FSP b/se/t/p	(2) FSP Sq.Err b/se/t/p	(3) RESP b/se/t/p	(4) RESP Sq.Err b/se/t/p
Tenured or Associate Professor	-0.0121	0.0000443	-0.0917	-0.0449
	(0.022)	(0.0070)	(0.029)	(0.015)
	[-0.54]	[0.0063]	[-3.17]	[-3.08]
	{0.596}	{0.995}	{0.006}	{0.007}
Familiarity, 0-10 scale	0.0132	0.00176	0.0370	0.0152
	(0.0044)	(0.0018)	(0.0063)	(0.0049)
	[3.03]	[0.98]	[5.87]	[3.10]
	{0.008}	{0.344}	{0.000}	{0.007}
Field match	0.0143	0.000129	0.00593	0.00995
	(0.021)	(0.011)	(0.026)	(0.027)
	[0.70]	[0.011]	[0.23]	[0.37]
	{0.497}	{0.991}	{0.821}	{0.719}
Constant	0.530	0.0865	0.602	0.0966
	(0.037)	(0.011)	(0.042)	(0.027)
	[14.3]	[7.69]	[14.5]	[3.58]
	{0.000}	{0.000}	{0.000}	{0.002}
Observations	2427	2427	2427	2427

Notes: The dependent variable in column (1) is the fraction of robustness tests significant prediction (FSP) for up to three findings for 17 papers, with one observation for each prediction. The dependent variable in column (2) is the squared prediction error for the FSP question. Column (3) is the relative effect size prediction (RESP). Column (4) is the squared error for the RESP. The survey included 359 participants, and each participant provided predictions for four papers. Each finding within a paper received between 77 and 95 predictions. Standard errors are multi-way clustered at the participant and paper levels.