Dreber, Anna; Johannesson, Magnus; Yang, Yifan

Working Paper

Selective Reporting of Placebo Tests in Top Economics Journals

I4R Discussion Paper Series, No. 31

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Dreber, Anna; Johannesson, Magnus; Yang, Yifan (2023) : Selective Reporting of Placebo Tests in Top Economics Journals, I4R Discussion Paper Series, No. 31, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/271197

INSTITUTE for REPLICATION

No. 31
I4R DISCUSSION PAPER SERIES

Selective Reporting of Placebo Tests in Top Economics Journals

Anna Dreber

Magnus Johannesson

Yifan Yang

May 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 31

Selective Reporting of Placebo Tests in Top Economics Journals

Anna Dreber[1], Magnus Johannesson[2], Yifan Yang[2]

[1]*Stockholm School of Economics, Stockholm/Sweden and University of Innsbruck/ Austria*
[2]*Stockholm School of Economics, Stockholm/Sweden*

MAY 2023

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Selective reporting of placebo tests in top economics journals

Anna Dreber, Magnus Johannesson, Yifan Yang*

Abstract

Placebo tests, where a null result is used to support the validity of the research design, is common in economics. Such tests provide an incentive to underreport statistically significant tests, a form of reversed p-hacking. Based on a pre-registered analysis plan, we test for such underreporting in all papers meeting our inclusion criteria (n=377) published in 11 top economics journals between 2009-2021. If the null hypothesis is true in all tests, 2.5% of them should be statistically significant at the 5% level with an effect in the same direction as the main test (and 5% in total). The actual fraction of statistically significant placebo tests with an effect in the same direction is 1.29% (95% CI [0.83, 1.63]), and the overall fraction of statistically significant placebo tests is 3.10% (95% CI [2.2, 4.0]). Our results provide strong evidence of selective underreporting of statistically significant placebo tests in top economics journals.

* 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); Yang: Department of Economics, Stockholm School of Economics (e-mail: yifan.yang@phdstudent.hhs.se). For financial support, we thank the Jan Wallander and Tom Hedelius Foundation (grant P21-0091 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 Jeffrey Clark for excellent research assistance and are grateful for insightful comments from John List and Roberto Weber.

Introduction

In observational data studies trying to estimate causal effects, typically using instrumental variables, difference-in-differences, or regression discontinuity methods, it has become standard to carry out so-called placebo tests.[1] There are different types of placebo tests, but in the most common variant the main hypothesis test is carried out on a time period or situation where the estimated effect is expected to be zero (i.e. the null hypothesis is expected to be true). An example can be using an outcome where there should be no effect or applying a regression discontinuity test on another time period than that for the studied discontinuity. A failure to reject the null hypothesis in the placebo test is interpreted as supporting the validity of the research design to identify causal effects.

In conducting placebo tests researchers face different incentives than in regular hypothesis tests. In regular hypothesis tests researchers have an incentive to engage in "p-hacking" and selectively report statistically significant findings (Simmons et al. 2011, John et al. 2012, Gelman and Loken 2014, Elliott et al. 2022). Brodeur et al. (2016, 2020) provide evidence for this type of selective reporting in regular hypothesis tests in top economics journals, while Vivalt (2019) finds evidence of this for impact evaluations of development programmes.[2] In placebo tests, researchers have an incentive to report null results and thus have an incentive to engage in a form of "reverse p-hacking" (selectively only reporting placebo tests that cannot reject the null hypothesis). Protzko (2018) referred to such behavior as "null hacking" in a setting where researchers have an incentive to report null results, and Eggers et al. (2023) recently mentioned this possibility in the context of placebo tests. But we are not aware of any previous empirical work testing for "null hacking" of placebo tests. We fill this gap in this study and test if statistically significant placebo tests are selectively underreported in top economics journals.

[1] We are not sure who invented the placebo test and when it was first used in general or in economics and there does not seem to be a standard reference for placebo tests (typically when these tests are introduced in a paper no reference is provided, but it is considered common knowledge). But it is a relatively recent thing and at the start of our data collection period (2009) they are relatively uncommon in our sample of journals, and then increases rapidly. Eggers et al. (2023) found a similar pattern in political sciences in searching seven top political sciences journals for the term "placebo test" between 2005 and 2021 and found no papers before 2009 and then an increasing trend to over 50 papers in 2021.

[2] See also the related work in sociology and political sciences by Gerber and Malhotra (2008,a,b) using similar methods to assess the distribution of test statistics around the significance threshold but referring to this as tests of publication bias.

If the null hypothesis is true in placebo tests, the false positive probability of these tests should equal the significance threshold used (i.e. if the tests are carried out at the 5% significance level, 5% of the placebo tests should report a statistically significant finding). This implies that 5% of published placebo tests should be significant at the 5% level if there is no selective reporting. If less than 5% of placebo tests in published papers have a two-sided p-value below 0.05, this provides evidence that placebo tests are selectively reported. As placebo tests that yield a significant effect in the opposite direction of the main results are sometimes interpreted as supporting the validity of the research design (see e.g. Ananyev and Guriev (2019) and Bahar and Rapoport (2018) from our pilot data collection discussed below); the incentives to underreport statistically significant placebo tests is strongest for placebo tests that yield a significant effect in the same direction as the main results. In our primary hypothesis test below we therefore test if the fraction of significant placebo tests with an effect in the same direction as the main results differ from 2.5% (the expected fraction if true null hypotheses are tested). We only include placebo tests where the authors argue that they expect the null hypothesis to be true and an eventual null result is used to support the validity of the research design (ruling out for instance placebo tests comparing if the effect size is larger in the main test than in the placebo test). Our test is conservative as it is unlikely that the null hypothesis is true in all placebo tests and without selective reporting the fraction of placebo tests that are significant at the 5% level can thus be expected to exceed 5% (and exceed 2.5% in our primary hypothesis test). This is likely to bias our results against our hypothesis of selective reporting of placebo tests.

To carry out our data collection an algorithm was developed to search published papers for the term “placebo tests(s)” to get a sample of potential papers to include. We first carried out a pilot study collecting data on placebo tests in Economic Journal between 2009-2021. This was to test the feasibility of the study in terms of the number of expected papers reporting placebo tests and to determine inclusion and exclusion criteria for the study. After the pilot study, we pre-registered an analysis plan detailing the data collection, inclusion and exclusion criteria, and all hypotheses and tests. After posting the pre-analysis plan a research assistant used the algorithm to search for potential papers from 11 top economics journals between 2009 and 2021. The identified potential papers were then manually searched for inclusion and data collected on placebo tests for included papers. Out of 540 papers in the potential sample, 377 (70%) were included in the study.

The mean fraction of significant placebo tests (at the 5% level) in these papers with an effect in the same direction as in the main analysis is 1.29%, which is statistically significantly below 2.5% (t-value=-5.41; p-value<0.00001). We thus find strong evidence of selective underreporting of statistically significant placebo tests. We also carry out 4 additional pre-registered secondary hypothesis tests. We find that the fraction of statistically significant placebo tests, irrespective of direction, of 3.10% is significantly below 5% (t-value=-3.95; p-value=0.00009). This provides further evidence of selective underreporting. The fraction of placebo tests that are statistically significant at the 5% level with an effect in the opposite direction of 1.82% is not statistically significantly different from 2.5% (t-value=-1.72; p-value=0.087). This is in line with weaker incentives for underreporting placebo tests with an effect in the opposite direction of the main results, but this fraction is not statistically significantly different from the fraction of significant placebo tests in the same direction as in the main analysis (t-value=1.22; p-value=0.223). We therefore cannot draw strong conclusions about if the underreporting of significant placebo tests differ for tests in the same and opposite direction of the main results. Finally, we coded each paper depending on if the authors concluded that the placebo test results supported the validity of the research design (yes/no) and compared this to the conservative benchmark of 97.5%. There was some ambiguity in coding this variable for 4 papers and depending on the coding of this variable between 98.9% and 100% of the papers concluded that the placebo tests supported the validity of the research design. We conclude that our results provide strong evidence of selective underreporting of placebo tests in top economics journals.

I. Methods

A. Data collection and inclusion/exclusion criteria

An algorithm to search for papers reporting placebo tests was developed for the project by a research assistant. Papers identified by the algorithm were then manually searched for inclusion into the study and data on placebo tests were collected for the included papers. The algorithm was first applied to the Economic Journal in a pilot study, to determine the feasibility of the data collection and inclusion/exclusion criteria. After conducting the pilot study, described in more detail in the Online Appendix, we pre-registered an analysis plan at OSF (https://osf.io/hfa9d/)

detailing the inclusion/exclusion criteria for the study, the outcome measures and the exact hypotheses and statistical tests to be conducted in the study. Thereafter a research assistant applied the algorithm to the 11 economics journals included in the main data collection and provided us with a list of the potential sample, i.e. all papers in these 11 journals identified by the algorithm where the term "placebo test(s)" had been used.[3]

We collected data on placebo tests for papers published in 12 top journals in economics that reported at least one placebo test (several placebo tests are often reported in papers reporting placebo tests). Data was collected for papers published between 2009-2021 (the motivation for starting the data collection in 2009 was that two of the included journals started in 2009). Data was collected for the following 12 journals: American Economic Journal: Applied Economics; American Economic Journal: Economic Policy; American Economic Review; Econometrica; Economic Journal; Journal of Development Economics; Journal of Labor Economics; Journal of Political Economy; Journal of the European Economic Association; Review of Economics and Statistics; Review of Economic Studies; Quarterly Journal of Economics. As mentioned above one of these 12 journals (Economic Journal) was included in a pilot study conducted prior to posting the pre-analysis plan, and papers from this journal are therefore not included in any hypotheses tests in the study (but we report descriptive results for this journal as well in Figure 1-4 and Online Appendix Table 1 below). These journals were chosen since they are highly influential journals in economics and were likely to publish studies with placebo tests (the time period of the data collection and the list of journals were pre-registered).

In the Online Appendix, the 10 inclusion/exclusion criteria used for the data collection are listed. There are different types of placebo tests in the literature and we used the following definition of placebo tests for inclusion: "A test where the authors argue that they expect the null hypothesis to be true and an eventual failure to reject the null hypothesis would be interpreted by the authors as support of the validity of their research design." This rules out placebo tests testing for a difference in the main treatment effect and the placebo effect (e.g. Card et al (2012)), where a significant result is used to support the validity of the research design. We also excluded placebo tests in

[3] The pre-analysis plan also included a signed statement by the research assistant that the list of articles in the potential sample would only be provided after the pre-analysis plan had been posted.

studies using the synthetic control method (typically simulating a placebo distribution that is compared to the main treatment effect; e.g. Abadie et al. (2010)), and in studies using randomized experiments (where tests labeled as placebo tests are typically balance tests). Papers reporting more than 100 placebo tests were also excluded to simplify the data collection. When we use the term "placebo test(s)" below we refer only to placebo tests covered by our inclusion/exclusion criteria.

Out of the 540 papers that the algorithm identified as mentioning the term "placebo test(s)", 163 were excluded and 377 were included (and out of the 65 pilot observations identified by the algorithm, 15 were excluded). There was ambiguity about the coding of several papers and this is discussed more in the Online Appendix. Due to this ambiguity we report results for two robustness tests that were not pre-registered (in addition to a pre-registered robustness test); see more on this below.

B. Outcome measures

For the 377 included papers we collected data on the following 4 outcome measures:

(1) the fraction of placebo tests in the paper reporting a two-sided p-value <0.05 and an effect in the same direction as the main hypothesis test.

(2) the fraction of placebo tests in the paper reporting a two-sided p-value <0.05 and an effect in the opposite direction of the main hypothesis test.

(3) the fraction of placebo tests in the paper reporting a two-sided p-value <0.05 (the sum of (1) and (2) above).

(4) a binary variable for if the authors of the paper in the text interpret the results of the placebo tests as supporting their research design and findings or not (yes=1 and no=0).

For all the outcome measures we use each paper as one observation rather than each placebo test as one observation as the placebo tests in a paper are not independent observations. We also collected information about the total number of placebo tests when that information was available (some papers only report that all placebo tests were non-significant without reporting the number of placebo tests conducted and these were still used to collect data about outcome variables (1) to (4) above; this was the case for 5 papers). The average (median) number of placebo tests per paper for the 372 papers where this information was available was 12.22 (8). To determine if the direction of significant placebo tests were in the same direction as the main hypothesis test, the closest estimation to the placebo test in the main results was used to determine the direction of the main hypothesis test irrespective of if this test was statistically significant or not.

C. Statistical power: Minimum detectable effect size

Based on the pilot study we carried out an ex ante estimation of the expected minimum detectable effect size (MDE) we had 80% power to detect for our primary outcome measure. This resulted in an MDE of 0.36 percentage units for tests at the 5% level ("suggestive evidence"; see below) and 0.47 percentage units for tests at the 0.5% level ("statistically significant evidence"; see below). See the Online Appendix for more details. We also pre-registered to report the MDE based on the observed standard errors in our study for primary hypothesis 1 and secondary hypothesis 1-4.

II. Results

We pre-registered one primary hypothesis and four secondary hypotheses so that we carry out five hypothesis tests in total (one for each of the four outcome measures described above; plus one test for the difference in two of these outcome measures). We also pre-registered to report the 95% confidence intervals for the outcome variables used in the five hypotheses tests, including separate confidence intervals for each journal including also Economic Journal used in the pilot data collection. But, as pre-registered, the confidence intervals for each journal should not be interpreted as hypothesis tests as they are likely to be underpowered. We also pre-registered one robustness test reported below. As pre-registered we will interpret a two-sided p-value <0.05 in the hypothesis tests as "suggestive evidence" and a two-sided p-value <0.005 "as statistically

significant evidence" in line with the recommendation of Benjamin et al. (2018). Each of the 377 papers meeting the inclusion criteria is one observation in all hypothesis tests below, except for secondary hypothesis 3 excluding one paper (see below) and the robustness tests based on excluding some of these observations as detailed below. The description of the hypotheses and test below exactly follow the pre-analysis plan unless otherwise noted. The results of the hypothesis tests are reported in Table 1.

Table 1: Results for the tests of primary hypothesis 1 and secondary hypothesis 1-4. Baseline results.

	(1)	(2)	(3)	(4)	(5)
	Primary	Secondary			
	Hypothesis 1	Hypothesis 1	Hypothesis 2	Hypothesis 3	Hypothesis 4
Mean Fraction	0.0129	0.0182	0.0310	0.9894	
Standard Deviation	0.0432	0.0772	0.0932	0.1026	
Standard Error	0.0022	0.0040	0.0048	0.0053	
95% CI	[0.0086, 0.0173]	[0.0104, 0.0260]	[0.0216, 0.0405]	[0.9790, 0.9997]	
Benchmark of Test	0.025	0.025	0.05	0.975	
Difference	-0.0121	-0.0068	-0.01895	0.0144	0.0052
Standard Error of Difference	0.0022	0.0039	0.0048	0.0053	0.0043
t/z-value	-5.4122	-1.7152	-3.9491	2.7145	1.2195
p-value	< 0.00001	0.08713	0.00009	0.00664	0.22343
95% CI of Difference					[-0.0032, 0.0137]
DF	376	376	376		376
Observations	377	377	377	376	377

A. Pre-registered hypotheses tests

Primary hypothesis 1: The mean fraction of placebo tests with a two-sided p-value <0.05 and an effect in the same direction as the main hypothesis test is less than 2.5%.

As our primary hypothesis test, we test if the mean fraction of placebo tests that are statistically significant in the same direction as the main results is below 2.5% (the expected fraction if true null hypotheses are tested in all placebo tests). This test is based on the first outcome measure above. We pre-registered this as the primary hypothesis test as placebo tests that are significant with an effect in the opposite direction of the main finding is sometimes interpreted as supporting the validity of the research design and findings (see e.g. Ananyev and Guriev (2019) and Bahar

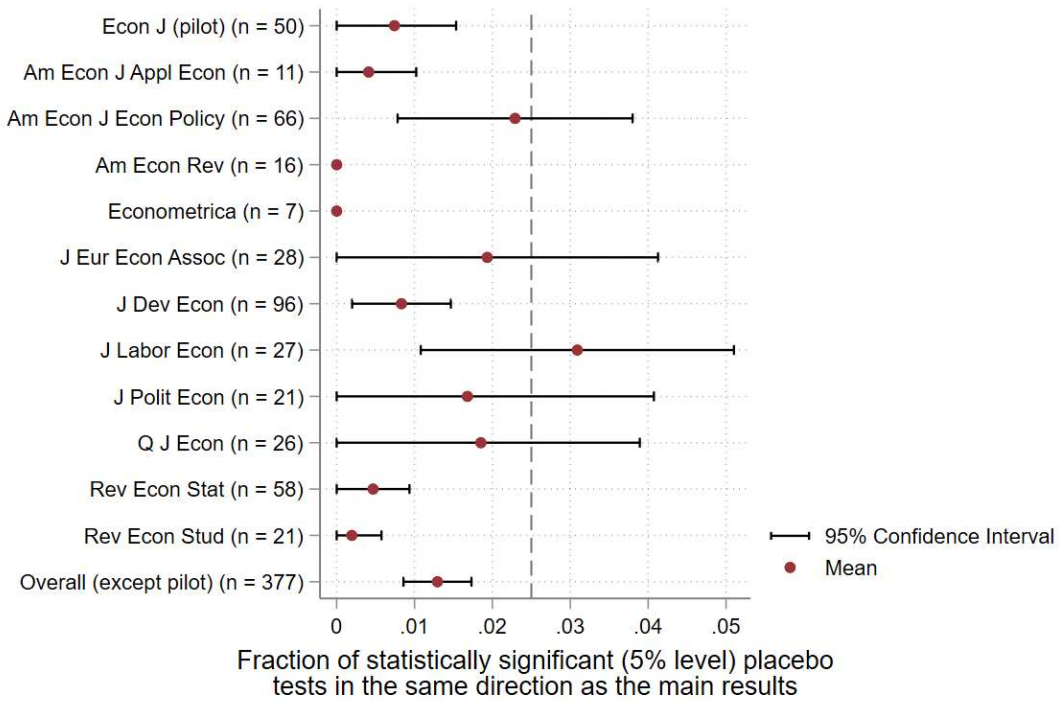

Figure 1. The mean fraction of significant placebo tests (at the 5% level) with an effect in the same direction as the main hypothesis test for each journal and "Overall" (the "Overall" result is the test of primary hypothesis test 1). The results are also shown for Economic Journal used in a pilot study, but these observations are not included in the "Overall" result used in the hypothesis test. The line at 0.025 (2.5%) shows the expected fraction if true null hypotheses are tested in all placebo tests and there is no selective reporting. 95% confidence intervals that overlap 0 are bounded at 0 in the Figure (as the fraction cannot be negative).

and Rapoport (2018) from the pilot data collection). There may thus be less incentives to underreport such significant tests. This is a conservative test of selective reporting as the true effect size is unlikely to be zero in all placebo tests.

We test primary hypothesis 1 in a one-sample t-test. The mean fraction of significant placebo tests with an effect in the same direction is 1.29%, which is statistically significantly lower than 2.5% (t-value=-5.41; p-value<0.00001). We thus confirm hypothesis 1. In Figure 1 we show the confidence intervals for the overall results and for each journal separately.

For 80% statistical power, the minimum detectable effect size (MDE) is 0.62 percentage units for tests at the 5% level and 0.81 percentage units for tests at the 0.5% level. This is higher than the ex ante power estimations of an MDE of 0.36 percentage units for tests at the 5% level and 0.47 percentage units for tests at the 0.5% level. This difference is due to a somewhat higher standard deviation of the primary outcome measure of 0.043 in the main data collection versus 0.030 in the pilot data collection and a lower number of included papers than the prediction based on the pilot data collection (377 versus 495).

Secondary hypothesis 1: The mean fraction of placebo tests with a two-sided p-value <0.05 and an effect in the opposite direction of the main hypothesis test differs from 2.5%.

In our first secondary hypothesis test, we test if the mean fraction of placebo tests that are statistically significant in the opposite direction of the main results differs from 2.5% (the expected fraction if true null hypotheses are tested in all placebo tests). This test is based on the second outcome measure above. We had no a priori hypothesized direction of this hypothesis test. The incentives to underreport significant placebo tests in the opposite direction of the main hypothesis test is less strong. Even if there is some underreporting of these tests, this may also be counteracted by placebo tests that do not test true null hypotheses so that the fraction of significant placebo tests exceeds 2.5%.

We test secondary hypothesis 1 in a one-sample t-test. The mean fraction of significant placebo tests with an effect in the opposite direction is 1.82%, which is not statistically significantly

different from 2.5% (t-value=-1.72; p-value=0.087). We therefore cannot reject the null hypothesis, for secondary hypothesis 1. In Figure 2 we show the confidence intervals for the overall results and for each journal separately. For 80% statistical power, the minimum detectable effect size (MDE) is 1.11 percentage units for tests at the 5% level and 1.45 percentage units for tests at the 0.5% level.

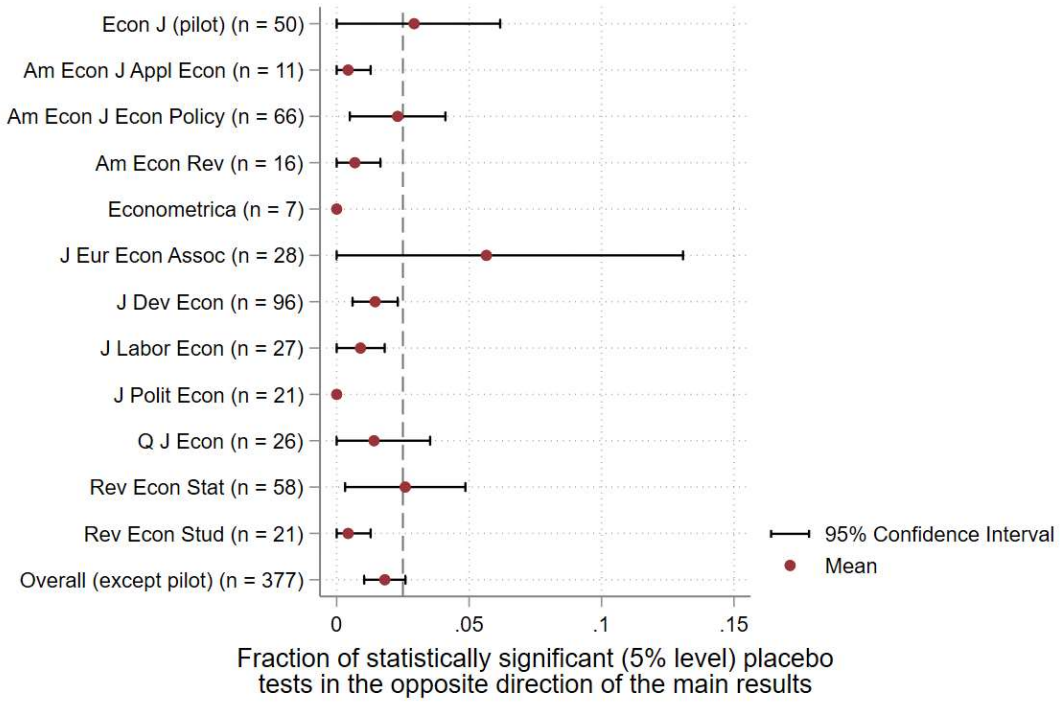

Figure 2. The mean fraction of significant placebo tests (at the 5% level) with an effect in the opposite direction of the main hypothesis test for each journal and "Overall" (the "Overall" result is the test of secondary hypothesis test 1). The results are also shown for Economic Journal used in a pilot study, but these observations are not included in the "Overall" result used in the hypothesis test. The line at 0.025 (2.5%) shows the expected fraction if true null hypotheses are tested in all placebo tests and there is no selective reporting. 95% confidence intervals that overlap 0 are bounded at 0 in the Figure (as the fraction cannot be negative).

Secondary hypothesis 2: The mean fraction of placebo tests with a two-sided p-value <0.05 differs from 5%.

In secondary hypothesis 2, we test if the mean fraction of placebo tests that are statistically significant at the 5% level differs from 5% (the expected fraction if true null hypotheses are tested

in all placebo tests). This test is based on the third outcome measure above. We had no a priori hypothesized direction of this hypothesis test. We expect selective underreporting of significant placebo tests with an effect in the same direction as the main hypothesis test, but this may be counteracted by the fraction of significant placebo tests with an effect in the opposite direction of the main hypothesis test exceeding 2.5% due to placebo tests not testing true null hypotheses (and the total fraction of significant placebo tests could exceed 5% even if primary hypothesis 1 is supported).

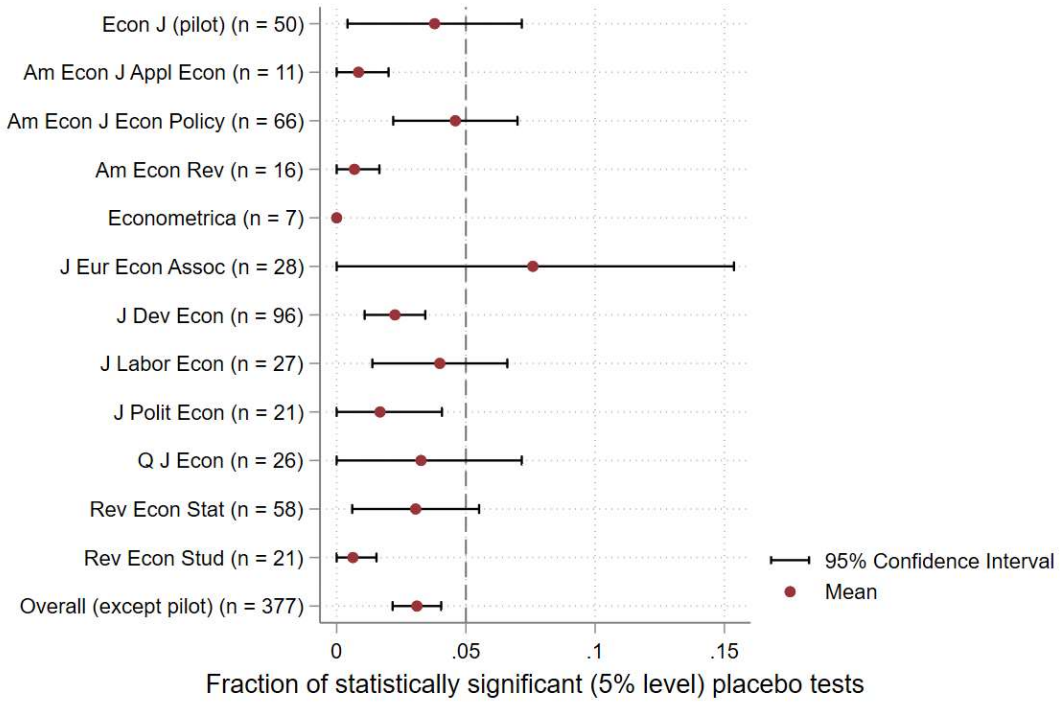

Figure 3. The mean fraction of significant placebo tests (at the 5% level) for each journal and "Overall" (the "Overall" result is the test of secondary hypothesis test 2). The results are also shown for Economic Journal used in a pilot study, but these observations are not included in the "Overall" result used in the hypothesis test. The line at 0.05 (5%) shows the expected fraction if true null hypotheses are tested in all placebo tests and there is no selective reporting. 95% confidence intervals that overlap 0 are bounded at 0 in the Figure (as the fraction cannot be negative).

We test secondary hypotheses 2 in a one-sample t-test. The mean fraction of significant placebo tests is 3.10%, which is statistically significantly different from 5% (t-value=-3.95; p-value=0.00009). We thus reject the null hypothesis, and find evidence of selective underreporting

of statistically significant placebo tests. In Figure 3 we show the confidence intervals for the overall results and for each journal separately. For 80% statistical power, the minimum detectable effect size (MDE) is 1.34 percentage units for tests at the 5% level and 1.75 percentage units for tests at the 0.5% level.

Secondary hypothesis 3: The fraction of papers where the authors conclude that the placebo test results support the validity of the research design and findings exceeds 97.5%.

Secondary hypothesis 3 is based on the fourth outcome measure described above, the subjective evaluation of if the authors conclude that the placebo test results support the validity of the research design (coded as a binary yes/no variable). This mean fraction will be between 0 and 100%. It is not obvious whether to compare this fraction to 95% or 97.5% depending on how statistically significant placebo tests with an effect in the opposite direction of the main results are interpreted by authors. We conservatively chose to compare the fraction to 97.5% based on assuming that authors only interpret statistically significant placebo tests with an effect in the same direction as their main results as a threat to validity.

Coding this variable turned out to be difficult for some papers. We excluded one paper from this test that explicitly concluded that the test results could be used to both conclude that the research design was valid as the placebo tests were not significant and that it could be used to conclude that it was not valid due to the wide confidence intervals of the placebo tests (Stevens et al. 2015). Excluding this paper is a deviation from the pre-analysis plan in terms of this being a situation we did not foresee would arise. For four additional papers we were unable to code this variable as a yes or a no as the writing of the authors can be used to support both interpretations (Martins 2009; Hoynes et al. 2015; Bollinger at al. 2020; MacPherson and Sterck 2021). These papers express some concerns about the significant placebo tests, but still conclude that their overall results and research design are valid. We think all these observations are more “yes” than “no”, but we report results for both codings for full transparency. In the baseline results we conservatively code these observations as “no”, and then we have added a robustness test reported below where we code these four observations as “yes”. The added robustness test is a deviation from the pre-analysis plan as we did not anticipate that we would be unable to code this variable for some papers.

We test secondary hypotheses 3 in a one-sample z-test (as it is a binary variable). The only papers coded as "yes" on this outcome variable are the four ambiguous observations discussed above. The mean fraction of papers concluding that the placebo tests support the validity of the research design is 98.94% and there is suggestive evidence that this is higher than 97.5% (z-value=2.714; p-value=0.0066). The ambiguity about the coding of the four ambiguous papers implies less clear cut evidence for this test, but the evidence is in line with selective underreporting of statistically significant placebo tests. In Online Appendix Table 1 we report the results separately for each journal reporting 95% confidence intervals for those cases where the fraction is not equal to 100%. The statistical power is lower on this test as it is based on a binary variable. For 80% statistical power, the minimum detectable effect size (MDE) is 1.48 percentage units for tests at the 5% level and 1.93 percentage units for tests at the 0.5% level.

Secondary hypothesis 4: The mean fraction of placebo tests with a two-sided p-value <0.05 and an effect in the opposite direction of the main hypothesis test exceeds the mean fraction of placebo tests with a two-sided p-value <0.05 and an effect in the same direction as the main hypothesis test.

In secondary hypothesis 4, we test if the mean fraction of placebo tests that are statistically significant at the 5% level in the opposite direction exceeds the mean fraction of placebo tests in the same direction as the main hypothesis test. This test is based on the difference between the second and the first outcome measure above. We hypothesized that the mean value of the paired difference would be positive, based on stronger incentives to underreport statistically significant placebo tests with an effect in the same direction as the main test. A positive difference will imply either selective reporting of placebo tests or that authors selectively implement placebo tests where they expect the true effect to go in the opposite direction of the effect in the main hypothesis test. We test secondary hypotheses 4 in a paired t-test where each paper is one paired observation with the value of outcome variable (2) and the value of outcome variable (1) for that paper. The mean fraction of significant placebo tests with an effect in the opposite direction of the main test is 1.82% and the mean fraction with an effect in the same direction is 1.29%, and the difference between these two means is 0.52 percentage units. This difference is in the hypothesized direction, but the difference is not statistically significant and we cannot reject the null hypothesis (t-value=1.22; p-

value=0.223). In Figure 4 we show the confidence intervals for the overall results and for each journal separately.

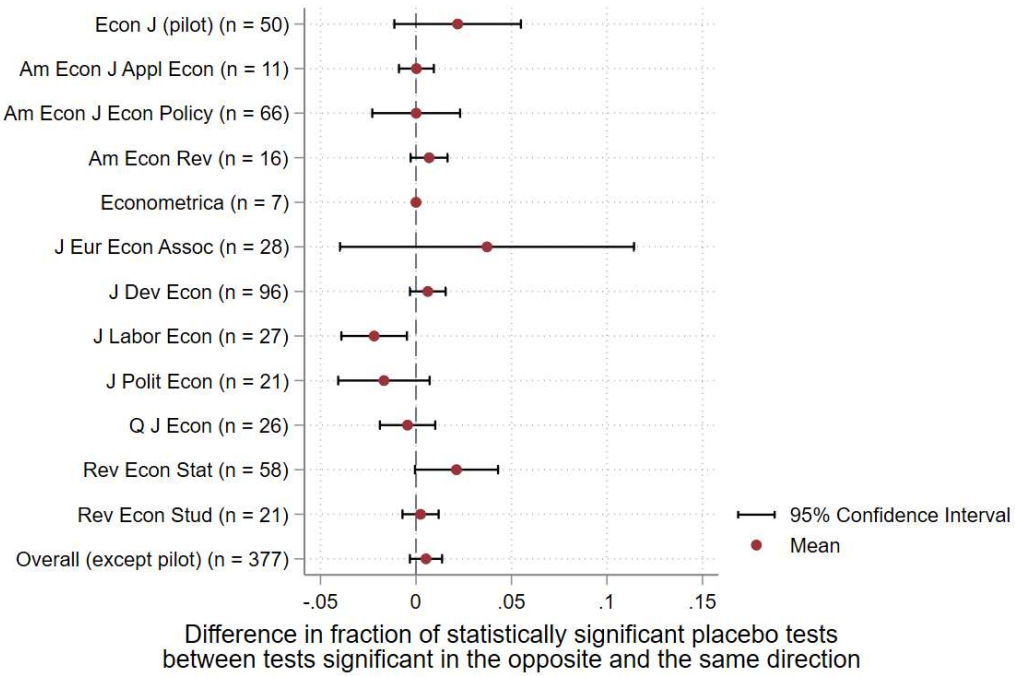

Figure 4. The difference in the mean fraction of significant placebo tests with an effect in the same direction as the main hypothesis test and the mean fraction of significant placebo tests with an effect in the opposite direction of the main hypothesis test; results shown for each journal and "Overall" (the "Overall" result is the test of secondary hypothesis test 4). The results are also shown for Economic Journal used in a pilot study, but these observations are not included in the "Overall" result used in the hypothesis test.

The statistical power is limited on this test as the standard error of the difference is substantially larger than for primary hypothesis 1 (testing the results against the same benchmark). For 80% statistical power, the minimum detectable effect size (MDE) is 1.20 percentage units for tests at the 5% level and 1.57 percentage units for tests at the 0.5% level.

B. Pre-registered Robustness test

In a pre-registered robustness test, we exclude all observations where we lack information about the number of placebo tests in the paper; this is the case for papers only reporting in the text that none of the placebo tests were significant without reporting the number of tests. This was the case

for 5 papers that are excluded in this robustness test. This does not change the conclusions in any of the hypothesis tests above. These results are reported in Online Appendix Table 2.

C. Not pre-registered robustness test: Ambiguous observations

We also carried out one robustness test of five included papers with strong ambiguity about inclusion (Fafchamps et al. 2010; Neal and Schanzenbach 2010; Saez et al. 2012; Niehaus et al. 2013; Borcan et al; 2017). We included these papers in the baseline results as that was most conservative for testing our hypotheses. However, we also report results in this section based on excluding these papers (or for one paper excluding a subset of the results). See the online Appendix for more details about the ambiguity of these five papers. After these exclusions the evidence of selective reporting in primary hypothesis 1 and secondary hypothesis 2 is even stronger; see Online Appendix Table 3 for these results. For primary hypothesis 1, the mean fraction of statistically significant placebo tests with an effect in the same direction as the main results is now 1.14% (t-value=-6.82 and p-value<0.00001 compared to the 2.5% benchmark). For secondary hypothesis 2 the mean fraction of statistically significant placebo tests, irrespective of direction, is now 2.37% (t-value=-8.00 and p-value<0.00001 compared to the 5% benchmark). There is now also statistically significant evidence of selective reporting of placebo tests in secondary hypothesis 1; the mean fraction of statistically significant placebo tests with an effect in the opposite direction of the main results is now 1.23% (t-value=-5.51 and p-value<0.00001 compared to the 2.5% benchmark).

D. Not pre-registered robustness test: Ambiguous coding of outcome variable 4

As mentioned above we found it difficult to code the variable for if the authors concluded that the placebo tests supported the validity of their research design or not for four papers (Martins 2009; Hoynes et al. 2015; Bollinger at al. 2020; MacPherson and Sterck 2021). These observations are discussed in more detail in the Online Appendix. In the baseline results reported above for secondary hypothesis 3 these four papers were coded as “no”, and we added a robustness test that was not pre-registered where we code these four papers as “yes” on this outcome variable. If the four ambiguous papers are coded as “yes” on this outcome variable the mean fraction of papers

concluding that the placebo tests support the validity of the research design is 100% and this is statistically significantly different from 97.5% (although the z-test cannot formally be conducted when the fraction is 100% as the standard error is 0).

III. Concluding remarks

The so-called "replication crisis" fuelled by low reported replicability and reproducibility of scientific findings has increased the interest in assessing and improving scientific practices (see, e.g., Ioannidis 2005, Maniadis, Tufano and List 2014, Open Science Collaboration 2015, Camerer et al 2016, 2018, Christensen and Miguel 2018, Andrews and Kasy 2019, Menkveld et al. 2023). Early pioneering work in this tradition in economics was made by Leamer (1983), with the classic article title "Let's take the con out of econometrics" and Dewald et al. (1986) testing if published findings in macroeconomics could be reproduced based on posted data and code. Our study testing for selective reporting of placebo tests is within this growing body of work now often referred to as metascience.

Our study provides strong evidence of selective underreporting of statistically significant placebo tests. We pre-registered our analysis after conducting a pilot study (not included in the main results) and before starting the main data collections. To ensure the credibility of the pre-registration a research assistant signed a statement as part of the pre-analysis plan where he ensured that he would not provide us with the list of potential papers for the main data collection until after we had posted the pre-analysis plan. We strictly followed our pre-registered analysis plan, with the exception of some added robustness tests for papers with ambiguity about the coding. However, we would still argue that our pre-analysis plan is weaker in this project than for a proper pre-analysis plan for an experimental study or other prospective data collections.[4] The ambiguity involved in the data collection in this type of project still offers some researcher degrees of freedom in collecting and coding the observations, and this is difficult to fully address with a pre-analysis

[4] By a proper pre-analysis plan we mean a detailed analysis plan that is subsequently followed in the published paper. Several studies suggest that economists often deviate from their pre-registrations or pre-analysis plans in intransparent ways (Abrams et al. 2020; Ofosu and Posner 2021; Brodeur et al. 2022).

plan. We added some additional robustness tests due to this ambiguity, but it still implies some caution in interpreting our results (and the results of other studies based on similar data collections).

Our study cannot cleanly answer what mechanism is driving the underreporting of statistically significant placebo tests. It could be a form of publication bias such that papers with statistically significant placebo tests are less likely to get published, or it could be due to authors deciding to underreport statistically significant placebo tests to increase the publication potential of their papers. Both these mechanisms are driven by the incentive system in academic publishing. One way to combat this problem would be pre-registering the placebo tests as part of a pre-analysis plan or publishing Registered Reports publications where the acceptance decision is made prior to collecting the data (Nosek et al. 2018; Arpinon and Espinosa 2023).[5] Unfortunately, it is not straightforward to implement pre-analysis plans and Registered Reports for the type of observational data studies included in this project as it is often difficult to verify that the researchers did not already have access to the data prior to posting the pre-analysis plan. Data complexity may also make it difficult to specify placebo tests prior to having access to the data. We may therefore need to come up with new innovative systems to resolve selective reporting issues in retrospective observational data studies, and this applies to both selective reporting of null results and statistically significant results. There is also substantive confusion about what a placebo test is even if we limit it to tests where a null result is used to support the validity of the research design. Tests labeled as placebo tests can be almost anything from tests where it is plausible that no effect should be observed such as replacing the key independent variable with lead values of the same variable, to tests where there is little a priori reason that the null hypothesis should be true such as testing the equal trends assumption in difference-in-difference studies or even testing for pure trends. Tests of for instance the equal trends assumption are better labeled as such rather than referring to them as placebo tests. That the null hypothesis cannot be rejected does also not mean that the null hypothesis is true and if placebo tests are poorly powered the fraction of false negative results may be high. A null result furthermore does not imply that the effect size differs from the main effect size, but some placebo tests instead test for a difference in effect sizes between the main result and the placebo treatment using a significant difference to support the validity of the research design (Card et al. 2012). This is a more demanding test, but to claim support for the

[5] Journal of Development Economics was the first adopter of Registered Reports in economics in 2018.

validity of the research design for designs that report statistically significant effect sizes it seems reasonable that both these conditions should be fulfilled (i.e. the placebo effect size should not differ significantly from the null and it should differ significantly from the main treatment effect size).

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105: 493-505.

Abrams, Eliot, Jonathan Libgober, and John A. List. 2020. "Research Registries: Facts, Myths, and Possible Improvements." NBER Working Paper 27250.

Ananyev, Maxim, and Sergei Guriev. 2019. "Effect of Income on Trust: Evidence from the 2009 Economic Crisis in Russia." *Economic Journal* 129: 1082-1118.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766-2794.

Arpinon, Thibaut, and Romain Espinosa. 2023. "A Practical Guide to Registered Reports for Economists." *Journal of the Economic Science Association* https://doi.org/10.1007/s40881-022-00123-1

Bahar, Dany, and Hillel Rapoport. 2018. "Migration, Knowledge Diffusion and the Comparative Advantage of Nations." *Economic Journal* 128: F273-F305.

Benjamin, Daniel J., et al. 2018. "Redefine Statistical Significance." *Nature Human Behavior* 2: 6-10.

Bollinger, Bryan, Jesse Burkhardt, Kenneth T. Gillingham. 2020. "Peer Effects in Residential Water Conservation: Evidence from Migration." *American Economic Journal: Economic Policy* 12 (3): 107-133.

Borcan, Oana, Mikael Lindahl, and Andreea Mitrut. 2017. "Fighting Corruption in Education: What Works and Who Benefits?" *American Economic Journal: Economic Policy* 9 (1): 180-209.

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

Brodeur, Abel, Nikolai Cook, Jonathan Hartley, and Anthony Heyes. 2022. "Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias?" Working paper.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110: 3634-3660.

Camerer, C. F., et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351: 1433-1436.

Camerer, C. F., et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2: 637-644.

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez. 2012. "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction." *American Economic Review* 102: 2981-3003.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920-980.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review* 76: 587-603.

Eggers, Andrew C., Guadalupe Tunón, and Allan Dafoe. 2023. "Placebo Tests for Causal Inference." Working paper.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wuthrich. 2022. "Detecting p-Hacking." *Econometrica* 90 (2): 887-906.

Fafchamps, Marcel, Marco J. van der Leij, and Sanjeev Goyal. 2010. "Matching and Network Effects." *Journal of the European Economic Association* 8 (1): 203-231.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102: 460-465.

Gerber, Alan S., and Neil Malhotra. 2008a. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37: 3-30.

Gerber, Alan S., and Neil Malhotra. 2008b. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3: 313-326.

Hoynes H, Miller D, Simon D. Income, the earned income tax credit, and infant health. *American Economic Journal: Economic Policy* 2015;7(1):172-211.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2: e124.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23: 524-532.

Leamer, Edward E. 1983. “Let’s Take the Con Out of Econometrics.” *American Economic Review* 73: 31-43.

MacPherson, Claire M., and Olivier Sterck. 2021. “Empowering Refugees Through Cash and Agriculture: A Regression Discontinuity Design.” *Journal of Development Economics* 149: 102614.

Maniadis, Zacharias, Fabio Tufano, and John A. List. 2014. “One Swallow Doesn’t Make a Summer: New Evidence of Anchoring Effects.” *American Economic Review* 104 (1): 277-290.

Martins, Pedro S. 2009. “Dismissals for Cause: The Difference That Just Eight Paragraphs Can Make.” *Journal of Labor Economics* 27 (2): 257-279.

Menkveld, Albert, et al. (forthcoming). “Non-Standard Errors.” *Journal of Finance*.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. “Left Behind by Design: Proficiency Counts and Test-Based Accountability.” *Review of Economics and Statistics* 92 (2): 263-283.

Niehaus, Paul, Antonia Atanassova, Marianne Bertrand, and Sendhil Mullainathan. 2013. “Targeting with Agents.” *American Economic Journal: Economic Policy* 5 (1): 206-238.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. “The Preregistration Revolution.” *PNAS* 115: 2600-2606.

Ofosu, George K., and Daniel N. Posner. 2021. “Pre-Analysis Plans: An Early Stocktaking.” *Perspectives in Politics* 21 (1): 174-190.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349: aac4716.

Protzko, John. 2018. “Null-Hacking, a Lurking Problem.” PsyArXiv preprints.

Saez, Emmanuel, Manos Matsaganis, and Panos Tsakloglou. 2012. “Earnings Determination and Taxes: Evidence From a Cohort-Based Payroll Tax Reform in Greece.” *Quarterly Journal of Economics* 127: 493-533.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22: 1359-1366.

Stevens, Ann H., Douglas L. Miller, Marianne E. Page, and Mateusz Filipski. 2015. “The Best of Times, the Worst of Times: Understanding Procyclical Mortality.” *American Economic Journal: Economic Policy* 7 (4): 279-311.

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

Online Appendix; Selective reporting of placebo tests in top economics journals.

Anna Dreber, Magnus Johannesson, Yifan Yang*

* 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); Yang: Department of Economics, Stockholm School of Economics (e-mail: yifan.yang@phdstudent.hhs.se).

Inclusion/exclusion criteria

The following procedure, inclusion and exclusion criteria, was used to collect the data (with the below inclusion/exclusion criteria copied from the pre-analysis plan):

1. An algorithm that has been developed will search the articles for the terms “placebo test” or “placebo tests” and all searched articles picked up by the algorithm will be included in the “potential sample”.
2. Papers in the “potential sample” will be manually searched for placebo tests. Placebo tests to be included in the data collection have to be labeled as placebo tests in the paper and be consistent with the following definition: a test where the authors argue that they expect the null hypothesis to be true and an eventual failure to reject the null hypothesis would be interpreted by the authors as support of the validity of their research design (note that the actual result of the placebo test does not affect this criteria, but the criteria is only affected by how the authors would interpret a non-significant test result). For papers to be included in the data collection they have to report the results of at least one placebo test in line with the above definition.
3. Only placebo tests where it is clear from the text, tables, or figures in the paper whether the placebo test is significant at the 5% level in a two-sided test will be included (placebo tests reported in tables as a coefficient and standard error or t-value, will be included even if significance levels or p-values are not reported and a ratio of the coefficient and standard error ≥ 1.96 will be interpreted as significant at the 5% level if information about significance or p-values is lacking in the tables). If a paper reports in the text that none of the placebo tests were significant without reporting the test results in tables or figures, the paper will be included in the data collection even if the number of placebo tests is not mentioned in the text and even if it is not explicitly mentioned which significance level that was used (and the fraction of placebo tests that are significant at the 5% level will be defined as 0 for this paper). If the information in the text about the number of significant placebo tests differs from the information in tables/figures reporting the placebo test results, we will base the data collection on the information in the tables/figures (e.g. if it is stated in the text that 4 placebo tests out of 20 were significant in a table and there are 5 significant

placebo tests out of 20 in the table; we will record the fraction of significant placebo tests as 5/20=0.25).

4. Placebo tests that are reported in a separate Appendix file or a separate supplementary materials/information file will only be included if they are mentioned in the main text document file (i.e. the algorithm will not search such separate files; and for papers in the "potential sample" such separate files will only be manually searched for placebo tests if the main text document refer to placebo tests reported in an Appendix or supplementary materials/information file).
5. Tests labeled as placebo tests that vary the instrument used in studies using instrumental variable methods will be excluded (as such tests are difficult to reconcile with the definition of placebo tests above).
6. The 1st stage results for placebo tests using instrumental variable methods will be excluded.
7. Placebo tests for studies using randomized experiments will be excluded (note that this does not necessarily mean that papers including randomized experiments will be excluded as some papers report results from more than one method/study and these will be considered as separate "studies").
8. Placebo tests for studies using the synthetic control method will be excluded (note that this does not necessarily mean that papers using the synthetic control method will be excluded as some papers report results from more than one method/study and these will be considered as separate "studies").
9. Papers where it is not possible to infer from the paper if each significant placebo test has an effect in the same direction as the main hypothesis test will be excluded (as the primary hypothesis test is based on the fraction of significant placebo tests with the same direction as the main hypothesis test).
10. Papers with more than 100 placebo tests will be excluded (to simplify the data collection).

Papers and placebo tests fulfilling these inclusion and exclusion criteria were included in the study. The term placebo test has been used in different ways in the literature. As is clear from the second inclusion/exclusion criteria above, data was only collected for placebo tests constructed so that an eventual failure to reject the null hypothesis of the test (i.e. a test result that is not statistically significant) would be interpreted by the authors as supporting the validity of their research design.

This is to avoid ambiguity about the use of the term placebo tests in published papers. Some papers also use the term placebo test for a test of if placebo effects and treatment effects are equal (e.g. Card et al 2012), and in such a test a significant difference would be used by the authors to support their hypothesis and validity of the research design. Such tests were therefore not included in our data collection (they are ruled out by the definition above as a significant difference rather than a null result would be used to support the validity of the research design). A related category of placebo tests is tests where the authors argue that a smaller effect size in absolute terms (rather than a null result) can be expected in the placebo test than in the main hypothesis test, but the tests are still carried out as comparisons to a null effect rather than as a comparison to the main hypothesis effect size (in the pilot data collection for instance the paper by Hoehn-Velasco (2021) expected smaller effect sizes in the placebo tests for children exposed at an older age and Lippmann et al. (2020) expected a lower fraction of significant placebo tests with a lower number of Eastern Länder in group 2 in their placebo tests; these two papers were therefore excluded). This type of placebo tests were also excluded based on the definition of placebo tests in the inclusion/exclusion criteria as for these tests the authors argue that the effect sizes are likely to be smaller in the placebo tests rather than true null effects.

In some papers the term placebo test is also used when a distribution of placebo effect sizes is generated to create a null distribution that is compared to the estimated treatment effect size in the study; e.g. in the synthetic control literature as in Abadie et al. (2010). Such tests are similar in spirit to tests of a difference in placebo effects and treatment effects, and are also ruled out for the same reason (as a significant difference between the placebo null distribution and the treatment effect size would be interpreted by the authors as support of the validity of the research design; note also that these studies may not report formal hypotheses tests but more compare the treatment effect size to the placebo null distribution in a graph commenting on the rank of the estimated treatment effect to the placebo null distribution). For simplicity we exclude studies using the synthetic control method as noted in the inclusion/exclusion criteria above (and several additional papers using similar methods to estimate a distribution of placebo effect sizes were excluded due to the additional exclusion criteria of excluding papers with more than 100 placebo tests).

We furthermore excluded placebo tests that vary the instrument used (in the 1st stage) in studies using instrumental variables methods (e.g. Akcomak et al. (2016) in the pilot data collection), as these tests are difficult to interpret in terms of null results supporting the validity of the research design (these placebo tests should be excluded in any case as they are not consistent with our definition of placebo tests in our inclusion/exclusion criteria; but we added this as a separate explicit exclusion criteria to simplify the data collection). Testing an alternative supposedly valid instrument would be a robustness test where a null result would cast doubt on the research design rather than validate it, and it's not clear what one could infer from a null result testing an alternative supposedly invalid instrument.

We also excluded the 1st stage results of placebo tests using instrumental variable methods (as the 2nd stage constitutes the relevant "placebo hypothesis test" and the 1st stage only tests if the instrument has some explanatory power in the 1st stage). For simplicity, we also excluded studies using randomized experiments (i.e. studies where the researcher randomizes subjects to treatments; but so called "quasi experiments" and "natural experiments" were included). In randomized experiments it is not common with placebo tests as the validation of the identification of causal effects is redundant with randomization and tests labeled as placebo tests in such studies may be more balance or attrition tests. To simplify the data collection we also excluded papers reporting more than 100 placebo tests.

Out of the 540 papers that the algorithm identified as mentioning the term "placebo test(s)", 163 were excluded (and out of the 65 pilot observations identified by the algorithm, 15 were excluded). The 163 excluded papers were excluded based on the following exclusion criteria:

Exclusion criteria 2: 56 papers, exclusion criteria 3: 4 papers, exclusion criteria 4: 0 papers, exclusion criteria 5: 3 papers, exclusion criteria 6: 2 papers, exclusion criteria 7: 17 papers, exclusion criteria 8: 4 papers, exclusion criteria 9: 6 papers, and exclusion criteria 10: 71 papers. The largest number of exclusions is papers reporting more than 100 placebo tests (71 papers), but 61 of those papers reported a distribution of placebo effect sizes and most of these papers would have been excluded in any case as they typically compare the distribution of placebo effect sizes to the estimated effect size rather than test null hypotheses (and papers with placebo tests of this

type that are interested also in the null hypotheses of the placebo tests often do not report exact test results but only shows a distribution of effect sizes around the null). The second largest number of exclusions (56 papers) is due to exclusion criteria 2, but 27 of those papers were excluded due to that no placebo tests could be found in the papers (among those 27 papers some papers made vague statements of placebo tests that could not be interpreted in terms of the fraction of significant placebo tests).

There was ambiguity about the coding of several papers, both in terms of if papers should be included or excluded and which tests should be included for a specific paper. The main ambiguity was in interpreting inclusion/exclusion criteria 2. Whether a null result supports the validity of the research design was typically straightforward to interpret but to interpret whether the authors argue that they expect the null hypothesis to be true was difficult for some papers as this was often implicit rather than explicit in the papers. If authors argued explicitly that they expected smaller rather than zero effects in the placebo tests those tests/papers were excluded as were tests focusing on comparing the magnitude of the placebo effects rather than testing for null hypotheses. But there were some borderline cases and such tests/papers were included in the benchmark analyses, but we report a not pre-registered robustness test excluding some potentially important observations in a robustness test. In our data file we also include a "Comments" column noting the ambiguity about observations involving some uncertainty; often this ambiguity is about the number of placebo tests for papers reporting a zero fraction of significant placebo tests and that ambiguity is not important for our hypotheses tests as the number of placebo tests does not affect our outcome measures for papers with a zero fraction of significant placebo tests (several papers say that they conducted additional non-significant placebo tests but without reporting the number of tests or exact results, and the number of placebo tests for papers reporting a zero fraction of statistically significant placebo tests in our data file is more of a lower bound measure).

Algorithm, pilot study and pre-registration

The algorithm to collect the potential sample was developed by a research assistant and had already been applied to one of the 12 journals (Economic Journal) in a pilot study when we posted the pre-analysis plan. The project group (AD, MJ and YY) did not receive any information about the

papers selected by the algorithm for the remaining 11 journals in this study before the posting of the pre-analysis plan. After June 20, 2022 (after the posting of the pre-analysis plan) the research assistant that developed the algorithm provided the project group with a list of the articles published in these 11 other journals between 2009-2021 where the algorithm found that "placebo test" or "placebo tests" were mentioned in the article (as part of the pre-analysis plan the research assistant provided a signed statement that this list of papers would not be provided to the project group until after the posting of the pre-analysis plan). As part of the pre-analysis plan we posted that YY would manually collect the data about placebo tests from the "potential sample" identified by the algorithm, and would if needed discuss with MJ and AD about which papers and placebo tests to include if there was ambiguity about whether a paper should be included or not and if there was ambiguity about the value of the outcome variable for included papers. We deviated somewhat from this procedure. YY initially manually collected all the data, but MJ also double-checked all the observations and papers with any ambiguity were discussed in the entire group. We did this change as the data collection proved more challenging than we had expected (related to this see the discussion below about some papers that were difficult to code and the added robustness tests on this).

Prior to posting the pre-analysis plan we conducted a pilot study. The pilot study was conducted on Economic Journal (one of the 12 journals included in our sample). The algorithm searched the files of published papers and identified papers that mention "placebo test" or "placebo tests" in the paper. The algorithm also collected information about how many times these search words were mentioned in a paper and on which pages. The algorithm identified 55 papers in Economic Journal published between 2009 and 2021 for the pilot study. The pilot study was based on this sample of 55 papers. However, after conducting the pilot study and using the algorithm to search the remaining 11 journals the research assistant made some slight changes to the algorithm (for instance allowing for some characters between "placebo" and "test" allowing for instance for quotation marks around the term placebo). After these changes the algorithm identified 10 additional potential papers in Economic Journal that we subsequently also included in our data collection (as part of the pilot observations). The new version of the algorithm also missed one Economic Journal paper identified by the first version, which we kept as part of the pilot sample.

However, none of the Economic Journal observations are included in any hypotheses tests below, but we only report descriptive results for the Economic Journal papers.

For the initial 55 papers identified by the algorithm and included in the pilot data collection AD, MJ and YY independently collected data manually about placebo tests for 13 of these papers and met to discuss these 13 papers to discuss any ambiguity about if they should be included or not and the value of the outcome measure of included papers. After this meeting YY collected data for the remaining 42 papers and AD, MJ and YY met again to discuss any ambiguity experienced by YY for these 42 papers.

Among the 65 papers identified by the algorithm (for the pre and post pilot version of the algorithm) in Economic Journal, data about placebo tests were collected in the 50 papers that satisfied our inclusion/exclusion criteria above. As the pilot study was used to inform the data collection, no data from Economic Journal are included in the hypotheses tests (but we report descriptive results also for Economic Journal as noted above).

For these 50 papers the mean fraction of placebo tests significant at the 5% level and an effect in the same direction as the main test was 0.74% (standard error=0.40 percentage units), the mean fraction of placebo tests significant at the 5% level and an effect in the opposite direction as the main test was 2.92% (standard error=1.66 percentage units), the mean fraction of placebo tests significant at the 5% level irrespective of direction was 3.79% (standard error=1.72 percentage units), and the fraction of these papers that concluded that the placebo test results supported the validity of the research design was 100%.

As part of the pilot study we carried out an ex ante estimation of the expected minimum detectable effect (MDE) size. This was based on the 55 Economic Journal papers identified by the pre-pilot version of the algorithm and a “ballpark” estimation of the expected sample size (based on the same number of included papers in the 11 other journals as for Economic Journal). For 80% statistical power, the minimum detectable effect size (MDE) is 2.8*se for tests at the 5% level and 3.65*se for tests at the 0.5% level (where se is the standard error of the mean fraction of placebo tests with a two-sided p-value <0.05 and an effect in the same direction as the main hypothesis

test). A standard deviation of our primary outcome measure in the pilot study of 0.030 and assuming a sample size of 495 (based on the "ballpark" estimate of the expected sample size), resulted in an MDE of 0.0036 for tests at the 5% level and 0.0047 for tests at the 0.5% level (where the MDEs of 0.0036 and 0.0047 implies a 0.36 and 0.47 percentage unit deviation from 2.5%). As stated in the pre-registration this estimation should be interpreted cautiously as the standard deviation was based on a small sample and there was considerable uncertainty about the sample size. As part of our results, we report the MDE (for 80% power and tests at the 5% and 0.5% levels) based on the observed standard error of our outcome measures used in primary hypothesis test 1 and secondary hypothesis tests 1-4. It turned out that the standard deviation in our overall sample for the primary outcome variable was somewhat larger than the STD used above (0.043 instead of the 0.030 used in the estimation above), and the sample size estimation above was overly optimistic as the final sample of included papers was 377 instead of the "ballpark" estimate of 495 above. The eventual statistical power is thus lower than in the ex ante estimation (with an MDE of 0.62 percentage units for tests at the 5% level and 0.81 percentage units for tests at the 0.5% level; compared to 0.36 and 0.47 percentage units in the ex ante MDE estimations).

Not pre-registered robustness test: Ambiguous observations

Here we report more details about the ambiguity of the five papers excluded in a not pre-registered robustness test (Fafchamps et al. 2010; Neal and Schanzenbach 2010; Saez et al. 2012; Niehaus et al. 2013; Borcan et al; 2017). Four of the papers were excluded in the robustness test and for the fifth paper some of the placebo tests were excluded in the robustness test. The results of this robustness test are reported in Online Appendix Table 3.

Fafchamps et al. (2010) provide arguments for both a null effect and a positive effect in the placebo tests and use the placebo test to distinguish between these arguments. If a null effect is observed this supports their main results of the effect of network proximity on first collaboration and if a positive effect is observed this suggests that their main result is due to network proximity being correlated with time varying unobserved match quality. On page 225 the authors provide arguments for why a significant effect may be observed in the placebo tests (on subsequent collaboration) when they write "In contrast, if network proximity is correlated with time-varying

unobserved match-quality, then it should remain significant for subsequent collaborations as well." On page 211 they similarly provide arguments for why a positive and a null effect could be observed. They find significant negative effects in both placebo tests and use these results to support the validity of their research design and their main findings (as the effect in the placebo test is not positive as in their main results). After observing the negative effects in the placebo tests they find that the negative effect of network proximity on subsequent collaboration (the placebo test) is driven by a negative effect at a distance of two in the network and they provide arguments for why there could be a negative effect of the placebo test at a distance of two. They argue that at a network distance of two there is one positive and one negative effect at work for first collaboration (the main test) and that the positive effect dominates, but for subsequent collaboration (the placebo test) only the negative effect exists. This paper was included as 2/2 significant placebo tests (both tests with effects in the opposite direction of the main results) in the benchmark analysis, but due to the ambiguity about if this paper fulfills our inclusion criteria it is excluded in this robustness test.

Neal and Schanzenbach (2010) report placebo tests in Figure 2A and B and 11/20 of these tests are significantly different from 0 (10 with effects in the opposite direction of the main results) in the Figure (in the Figures it is difficult to see if 10 or 11 of the placebo tests are significant, but the observation was included as 11 significant placebo tests). The placebo tests show the change in reading and math scores in different deciles between 2004 and 2005 in the placebo group used to contrast with the corresponding results in Figure 1A and B for the "treatment group". However, in interpreting the placebo test results the authors are primarily concerned with comparing the patterns in Figure 2A and B with Figure 1A and B to verify that the patterns differ; and the authors do not argue that the null hypothesis is expected to hold. The authors interpret the 11/20 statistically significant placebo tests to support the validity of the research design as the patterns differ between Figure 2A and B and Figure 1A and B. The only comment the authors make about the null hypothesis is that they on page 271 write that "We do not know why there are some statistically significant deviations from 0 in these figures. In any pair of years, especially during the early years of a new policy regime, there may be differences in test administration or curricular priorities that create such differences. Our main point is that these figures describe differences between two cohorts that experienced broadly similar accountability environments, and these

differences in no way fit the pattern observed in figures 1A and 1B." This paper was included as 11/20 significant placebo tests in the benchmark analysis, but due to the ambiguity about if this paper fulfills our inclusion criteria it is excluded in this robustness test.

Saez et al. (2012) involves a different form of ambiguity compared to the other studies in this robustness test. For this paper the ambiguity is not about the inclusion criteria of the paper and placebo test, but that the text of the paper and the Figures differs substantially from the placebo test results reported in Table A2 Panel B (15/30 significant placebo tests; 7 in the same direction as the main results and 8 in the opposite direction of the main results). In the main text (page 526) the authors refer to the placebo tests as "We provide in the Online Appendix placebo tests showing that there is no discontinuity at the cut-off entry date in the distribution of (gross, posted, and net) earnings below the old cap." In the Online Appendix these results are also reported in Figure A4 panel A and B and the Figure legend concludes "Both graphs confirm that there are no discontinuities in any of those series at the cut-off date." On page 38 of the Online Appendix describing these placebo tests the author writes for Figure A4 panel A that "the series indeed do not display any discontinuity" and for Figure A4 panel B that "we do not find any systematic discontinuity at the cut-off date. Some of the coefficients are significant in some specifications but the magnitudes are much smaller than in Table V and a number of coefficients are not significant." But several coefficients in Online Appendix Table A2 panel B are larger than those in Table V. The text and Figure A4 (which does not report significance of any tests) are thus strongly at odds with the many significant placebo tests (and the magnitude of those effect sizes) in Table A2. The authors also use the placebo test results as support of the validity of the research design. This paper was included as 15/30 significant placebo tests in the benchmark analysis, but due to the large discrepancies between the text and Figures on the one hand and the Table A2 results on the other hand we excluded this study in this robustness test. Some other papers also have discrepancies between the text, Figures and Table results, but those differences are not as large as for this paper.

Niehaus et al. (2013) use a variable referred to as "Number of placebo violations" in their placebo tests reported in Table 4 in the paper. A null effect of these variables supports the validity of the used research design, but the authors do not argue that they expect the null hypothesis in these placebo tests to be true but give various ex ante arguments for effects in different directions of the

placebo tests such as on page 230 writing "We also estimate whether placebo violations are correlated with BPL card ownership, though here it is less clear what to expect: if placebo violations are correlated with both demand shifters and prices, then their effect on card ownership is ambiguous." The 2/6 significant placebo tests (with effects in the same direction as the main results) were included in the benchmark analyses above, but given the ambiguity about if these placebo tests are consistent with our inclusion criteria we excluded this paper in this robustness test.

Borcan et al. (2017) carry out 6 non-significant placebo tests of a “placebo camera” whose inclusion was not ambiguous. However, in their Table 3 labeled “Placebo Test” (which included three of the “placebo camera” tests) they also include a trend variable referred to as "Year12" and year11" (Table 3 columns 5-8). The results of these time trend dummies are also discussed as placebo tests in the paper but the authors do not argue that the null hypothesis is likely told (but on page 193 they describe these time trends being zero as “under the very strong assumption”). On page 197 the authors interpret the positive time trends (in opposite direction to the main results) in the placebo tests as evidence that the main results were not negatively affected by a general year trend. 7 out of 8 time trend effects in the placebo tests were significant (in the opposite direction to the main findings) and were included in our benchmark results leading to 7/14 significant placebo tests in total for this paper. The authors concluded that the 7 out of 8 statistically significant time trend placebo tests supported the validity of their research design. Given that the authors describe the null hypothesis in these placebo tests as a very strong assumption and do not argue that the null hypothesis in the placebo tests is likely to hold the inclusion of these tests is questionable and in this robustness tests we excluded these tests and included this observation as 0/6 significant placebo tests.

Not pre-registered robustness test: Ambiguous coding of outcome variable 4

For four papers there was ambiguity about coding the yes/no variable for if the authors concluded that the placebo test results supported the validity of the research design or not (Martins 2009; Hoynes et al. 2015; Bollinger at al. 2020; MacPherson and Sterck 2021). In the baseline analysis we conservatively coded this variable as “no” for these four observations, but we also carried out

a robustness test coding these four observations as “yes” (see Online Appendix Table 1 for results for each journal when these observations are coded as “no” in the baseline analysis; when they are coded as “yes” the outcome variable is 100% “yes” for all journals). We also excluded one paper from this test in the baseline analysis that explicitly concluded that the test results could be used to both conclude that the research design was valid as the placebo tests were not significant and that it could be used to conclude that it was not valid due to the wide confidence intervals (Stevens et al. 2015). Below we give more details about the ambiguity about the authors’ conclusion for these five papers.

Martin (2009) writes about the placebo tests (page 272-273) that "Overall, I interpret the results of this analysis based on "artificial" thresholds as evidence that the main results concerning job and worker flows and job performance are not picking up effects that emanate from differences in firm size and that happen to coincide with the thresholds defined by the new law. However, these wage results suggest that the firm-size wage premium may have increased over the 1990s or that there may be systematic differences in wage growth between firms of different size, which will explain the significant wage results in my benchmark analysis." And on page 275 the author writes "Several results also indicate slower wage growth at treated firms, which would be consistent with bargaining models, but these findings are not robust to all placebo tests."

Hoynes et al. (2015), write about the placebo tests results page 205) that “The gap between fourth and third births does raise a cautionary note about potential parity-specific trends in birth weight, and our analysis should be interpreted in light of this caution. We believe that despite this, the preponderance of evidence indicates that the EITC does improve child health. First the timing of these spurious trends does not correspond cleanly with the policy change. And second, in our “maxcredit models”, results are robust to inclusion of parity-specific trends.”

Bollinger et al. (2020) write about the significant placebo test results (page 121) that “This immediately raises concerns about the identification of peer effects in the OLS specifications, even with the rich set of fixed effects. We view this result as indicating that there is an endogeneity issue, likely due to trends that affect both peer group consumption and the household’s water consumption. On the other hand, the IV results are noisy, but are quite close to 0 and show no

statistically significant relationship (see Table A.6 for the first-stage results). While this result alone cannot rule out all possible identification concerns, it shows that there is no evidence of unobservable trends confounding identification, further supporting the validity of our primary results. It also highlights the importance of an instrumental variables strategy in identifying peer effects in our setting." In the Conclusion section they further write (page 130) that "We further perform a series of placebo tests and robustness checks that uniformly support the contention that our IV strategy allows us to identify causal effects." The author here uses the significant placebo tests in the OLS regression as an additional argument for their IV strategy.

MacPherson and Sterck (2021) write about the placebo tests (page 11) that "The coefficients in the regression on the independence from aid variable are significant at the 1% threshold, which is probably due to the high variability of this subjective measure and might explain why the findings for this variable are not fully robust." But in the Introduction section (page 2) they write that "These results are robust to various tests and specification checks."

Stevens et al. (2015) explicitly concluded that the test results could be used to both conclude that the research design was valid as the placebo tests were not significant and that it could be used to conclude that it was not valid due to the wide confidence intervals. In footnote 25 they write "The fact that neither interaction is close to statistical significance can support an argument that the placebo test "passes". On the other hand the point estimates for the interaction are larger than those for the 65+ group, which argues that the test "fails". Our interpretation is that the very large standard error estimates make this test uninformative. We note that these large standard error estimates do contrast with those for the 65+ group."

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105: 493-505.

Akcomak, I. Semih, Dinand Webbink, and Bas ter Weel. 2016. “Why Did the Netherlands Develop So Early? The Legacy of the Brethren of the Common Life.” *Economic Journal* 126: 826-860.

Almond, Douglas, Bhashkar Mazumder, and Reyn van Ewijk. 2015. “In Utero Ramadan Exposure and Children’s Academic Performance.” *Economic Journal* 125: 1501-1533.

Bollinger, Bryan, Jesse Burkhardt, Kenneth T. Gillingham. 2020. “Peer Effects in Residential Water Conservation: Evidence from Migration.” *American Economic Journal: Economic Policy* 12 (3): 107-133.

Borcan, Oana, Mikael Lindahl, and Andreea Mitrut. 2017. “Fighting Corruption in Education: What Works and Who Benefits?” *American Economic Journal: Economic Policy* 9 (1): 180-209.

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez. 2012. “Inequality at Work: The Effect of Peer Salaries on Job Satisfaction.” *American Economic Review* 102: 2981-3003.

Fafchamps, Marcel, Marco J. van der Leij, and Sanjeev Goyal. 2010. “Matching and Network Effects.” *Journal of the European Economic Association* 8 (1): 203-231.

Hoehn-Velasco, Lauren. 2021. “The Long-Term Impact of Preventative Public Health Programmes.” *Economic Journal* 131: 797-826.

Hoynes, Hilary, Doug Miller, and David Simon. 2015. “Income, the Earned Income Tax Credit, and Infant Health.” *American Economic Journal: Economic Policy* 7 (1): 172-211.

Lippmann, Quentin, Alexandre Georgieff, and Claudia Senik. 2020. “Undoing Gender with Institutions: Lessons from the German Division and Reunification.” *Economic Journal* 130: 1445-1470.

Martins, Pedro S. 2009. “Dismissals for Cause: The Difference That Just Eight Paragraphs Can Make.” *Journal of Labor Economics* 27 (2): 257-279.

MacPherson, Claire, and Olivier Sterck. 2021. "Empowering Refugees Through Cash and Agriculture: A Regression Discontinuity Design." *Journal of Development Economics* 149: 102614.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92 (2): 263-283.

Niehaus, Paul, Antonia Atanassova, Marianne Bertrand, and Sendhil Mullainathan. 2013. "Targeting with Agents." *American Economic Journal: Economic Policy* 5 (1): 206-238.

Saez, Emmanuel, Manos Matsaganis, and Panos Tsakloglou. 2012. "Earnings Determination and Taxes: Evidence From a Cohort-Based Payroll Tax Reform in Greece." *Quarterly Journal of Economics* 127: 493-533.

Stevens, Ann H., Douglas L. Miller, Marianne E. Page, and Mateusz Filipski. 2015. "The Best of Times, the Worst of Times: Understanding Procyclical Mortality." *American Economic Journal: Economic Policy* 7 (4): 279-311.

APPENDIX TABLES

Online Appendix Table 1: The Table shows the number and fraction of papers for each journal and overall coded as that the authors conclude that the placebo tests support the validity of their research design (based on that the four papers whose coding on this variable was ambiguous are coded as "no"). 95% confidence intervals that overlap 1 are bounded at 1 in the Table (as the fraction cannot exceed 1).

Journal	Number of papers	Authors conclude that placebo tests support validity	Fraction	Standard Error	95% Confidence Interval
Econ J (Pilot)	50	50	1		
Am Econ J Applied Econ	11	11	1		
Am Econ J Econ Policy	65	63	0.9692	0.0216	[0.9269, 1]
Am Econ Rev	16	16	1		
Econometrica	7	7	1		
J Dev Econ	96	95	0.9896	0.0104	[0.9692, 1]
J Eur Econ Assoc	28	28	1		
J Labor Econ	27	26	0.9630	0.0370	[0.8904, 1]
J Polit Econ	21	21	1		
Q J Econ	26	26	1		
Rev Econ Stat	58	58	1		
Rev Econ Stud	21	21	1		
Total (except pilot)	376	372	0.9894	0.0053	[0.9790, 0.9997]

Online Appendix Table 2. Results for the tests of primary hypothesis 1 and secondary hypothesis 1-4. Pre-registered robustness test excluding the 5 papers where we lack information about the number of placebo tests in the paper (this is the case for papers only reporting in the text that none of the placebo tests were significant without reporting the number of tests).

	(1)	(2)	(3)	(4)	(5)
	Primary	Secondary			
	Hypothesis 1	Hypothesis 1	Hypothesis 2	Hypothesis 3	Hypothesis 4
Mean Fraction	0.0131	0.0184	0.0315	0.9892	
Standard Deviation	0.0435	0.0777	0.0937	0.1026	
Standard Error	0.0023	0.0040	0.0049	0.0053	
95% CI	[0.0087 0.0175]	[0.0105, 0.0263]	[0.0219, 0.0410]	[0.9788, 0.9997]	
Benchmark of Test	0.025	0.025	0.05	0.975	
Difference	-0.0119	-0.0066	-0.0185	0.0142	0.0053
Standard Error of Difference	0.0023	0.0040	0.0049	0.0053	0.0044
t/z-value	-5.2665	-1.6324	-3.8137	2.6695	1.2195
p-value	<0.00001	0.10345	0.000160	0.00760	0.22344
95% CI of Difference					[-0.0032, 0.0138]
DF	371	371	371		371
Observations	372	372	372	371	372

Online Appendix Table 3. Results for the tests of primary hypothesis 1 and secondary hypothesis 1-4. Not pre-registered robustness test excluding four ambiguous papers and excluding some ambiguous test results from a fifth paper.

	(1)	(2)	(3)	(4)	(5)
	Primary	Secondary			
	Hypothesis 1	Hypothesis 1	Hypothesis 2	Hypothesis 3	Hypothesis 4
Mean Fraction	0.0114	0.0123	0.0237	0.9892	
Standard Deviation	0.0385	0.0445	0.0636	0.1026	
Standard Error	0.0020	0.0023	0.0033	0.0053	
95% CI	[0.0075, 0.0153]	[0.0078, 0.0168]	[0.0172, 0.0301]	[0.9788, 0.9997]	
Benchmark of Test	0.025	0.025	0.05	0.975	
Difference	-0.0136	-0.0127	-0.0263	0.0142	0.0009
Standard Error of Difference	0.0020	0.0023	0.0033	0.0053	0.0028
t/z-value	-6.8173	-5.5140	-7.9989	2.6785	0.3157
p-value	< 0.00001	< 0.00001	< 0.00001	0.00740	0.75244
95% CI of Difference					[-0.0046 0.0063]
DF	372	372	372		372
Observations	373	373	373	372	373