DEPARTMENT OF ECONOMICS AND FINANCE

SCHOOL OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Do Negative Replications Affect Citations?

NOTE: This paper is a revision of University of Canterbury WP No. 2021/14

Tom Coupé
W. Robert Reed

WORKING PAPER

No. 16/2022

Department of Economics and Finance
UC Business School
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

WORKING PAPER No. 16/2022

Do Negative Replications Affect Citations?

Tom Coupé[1]
W. Robert Reed[1†]

September 2022

Abstract: This study examines the effect of negative replications on the citation rates of replicated studies. We study a set of 204 replicated studies in economics and compare their citation performance with an initial sample of 112,000 potential controls taken from Scopus. From this initial pool, we match each replicated study with multiple controls based on having comparable citation histories. Our main finding is that there is no evidence that studies that receive negative replications suffer a penalty in the form of fewer citations. We also find that replicated studies receive somewhat more citations than their matched control studies, though here the causal interpretation is more suspect.

Keywords: Replications, Citations, Matching, Meta-science, Self-correcting science

JEL Classifications: A11, A14, B41, C18

Acknowledgements: We acknowledge helpful feedback from Patrick Bayer, Andrew Chang, James Druckman, Uri Gneezy, Nicholas Huntington-Klein, Benjamin Jones, Miklós Koren, Joshua Krieger, Felix Schafmeister, Kyle Jensen, Marta Serra-Garcia, and participants at Metascience 2021, aimos 2021, and the BITSS 10th Annual Meeting. Uati Saofai provided able research assistance. Data from the website Replication Wiki is gratefully acknowledged and is used in accordance with the terms and conditions of the Creative Commons Attribution 4.0 International Public License.

[1] Department of Economics and Finance, University of Canterbury, NEW ZEALAND

† Corresponding author: W. Robert Reed. Email: bob.reed@canterbury.ac.nz

I. Introduction

Is science self-correcting? In other words, are there mechanisms in the market for scientific ideas that discourage the proliferation of facts and theories that have been refuted? Answering this question requires providing clarification around key questions. What is the "market for scientific ideas"? How can one measure the "proliferation of facts and theories"? When is a fact or theory considered to be "refuted"? Recent research has focused on citations of journal articles as a measure of effect and found evidence both in favour and against the notion that science is self-correcting.

Previous research examining the effect of negative/unsuccessful replications on citations has exploited the database generated by the Reproducibility Project: Psychology (Open Science Collaboration, 2015). This large-scale, multi-lab project investigated the reproducibility of 100 highly-cited experiments in psychology. They reported that only 36% of the replicated experiments produced statistically significant estimates in the same direction as the originals.[1] Three subsequent studies used the outcomes from these replications to study differences in citation patterns between studies with positive and negative replications. Yang et al. (2020) and Schafmeister (2021) found no difference. Serra-Garcia & Gneezy (2021), using an expanded database that added replications from economics and general science journals, found that studies with negative replications were actually cited more. Yang et al. (2020) focused on aggregate differences in annual citation rates. Schafmeister (2021) and Serra-Garcia & Gneezy (2021) used regression adjustment to adjust for heterogeneity across studies.

A somewhat more positive view of self-correcting science comes from the literature studying the effect of retractions on citations. Furman et al. (2012) estimated that retracted

[1] The terminology around replications can be confusing. "Replication" can sometimes be meant to imply successful confirmation of a previous result, as in "their finding was replicated by Smith and Jones (XXXX)". We use replication to mean any study whose main purpose is to confirm a key finding(s) from a previously published study.

studies in biomedicine received 65% fewer yearly citations over the post-retraction period compared to a matched control sample. Azouley et al. (2015) performed a similar analysis for retracted studies in PubMed and estimated a 69% reduction in annual citations. They also investigated the possibility of "spillover effects"; that is, that studies whose content was "related" to the retracted study might also face a citation penalty. They report a 5-10% reduction in annual citations for these "related" articles compared to matched control studies.

Lu et al. (2013) explored another aspect of spillovers; that retractions impact the citations of the retracted authors' other research. They focused on citations to research retracted authors had published prior to the date of the retraction. They found an annual citation penalty of 6.9% in the years following retraction, though there was no effect if the retraction arose from a self-reported error. Jin et al. (2019) further explored spillover effects and found greater citation penalties for "less eminent" co-authors of a paper, something they called the "Reverse Matthew Effect".

Nevertheless, there is ample evidence that retracted and discredited studies continue to be cited, even favourably cited, after the original claims have been repudiated (Budd et al., 1998; Bornemann-Cimenti et al., 2016; Tatsioni et al., 2007; Candal-Pedreira, 2020; Schneider et al., 2020; Fernández et al., 2021; Hardwicke et al., 2021; Hsiao & Schneider, 2021; Piller, 2021). Retraction Watch's "Top 10 Most Highly Cited Retracted Papers" tracks number of citations before and after retraction, several of which have more citations after retraction (Retraction Watch, n.d.).

In summary, the evidence for self-correcting science from the retraction literature is mixed, while the evidence from the replication literature is negative, though thinner and concentrated in psychology. In weighing these different findings, it can be argued that replications provide a more meaningful perspective on whether science self-corrects. Retracted studies are extreme events. It takes a lot for a journal to retract a paper. For example, the academic publisher Wiley states the following criterion for retraction: "There is major scientific error which would invalidate the conclusions of the article, for example where there is clear evidence that findings are unreliable, either as a result of misconduct (e.g. data fabrication) or honest error (e.g. miscalculation or experimental error)."[2]

Given such a high bar, many inferior studies will fail to be culled from the literature through retraction. Replication provides the only way to address these studies. Observing how the literature responds to replications arguably provides a better gauge of how well the academic market of ideas is functioning.

Accordingly, this study examines the effect of negative replications on the citation rates of replicated studies in economics. We study a set of 204 replicated studies and compare their citation performance with an initial sample of 112,000 potential controls taken from Scopus. Approximately half of the replicated studies had their results refuted by their replications, with the remaining half receiving either a confirmation or a mixed conclusion.

Using matching criteria that accommodate (i) differences in the lengths of time between publication of the original study and its replication, as well as (ii) differences in the number of citations, we match each replicated study with multiple, non-replicated controls based on having comparable, year-by-year citation histories. Our main samples consist of 74, 103, and 142 replicated studies (the "Treated") and 7,044, 7,552, and 11,202 matched control studies, respectively.[3] We have two main findings. First, studies that are replicated receive more citations than their matched control studies. Second, there is no evidence that studies that receive negative replications suffer a penalty in the form of fewer citations.

[2] From Wiley's website: https://authorservices.wiley.com/ethics-guidelines/retractions-and-expressions-of-concern.html, retrieved November 11, 2021.

[3] Note that some controls are matched to more than one treated. There are 6,571, 7,056, and 10,330 unique controls in the three samples, respectively.

II. Matching Strategy

The "Treated". The first issue we need to address is how to define a "replication". The term is used in many different ways. It can mean simply verifying the original authors' code reproduces the results, performing the same analysis on a different dataset, re-estimating the original specification using a different estimator, substituting alternative variables that arguably measure the same thing, running an experiment on a new set of participants, etc. We relied on two sources that collect information on replications in economics: The Replication Network and ReplicationWiki (Höffler, 2017). Both sites aim to provide an exhaustive list of all replications in economics. We followed *The Replication Network* in defining a replication as any study for which the main purpose was to determine the correctness of a previously published study. As a quality constraint, we only included replications that were published in peer-reviewed journals.

We then matched each replication with the study it replicated. We excluded (i) replication studies that replicated more than one original paper, and (ii) original papers that were replicated by more than one replication study. This gave us pairs of a replication and an original study that were not linked to any other replications or original studies. We further excluded pairs for which we had less than 3 years of post-replication citation data (i.e., papers published after 2016); and less than two full years of citation histories on which to match treated and controls. This resulted in a sample of 204 original studies with corresponding replications.

The next issue we had to address was how to define a "negative" replication. Here again there are numerous ways to define replication "failure/success". A researcher may decide that a replication has produced a negative outcome if a key coefficient is statistically insignificant but was significant in the original. Or if the sign of the coefficient reversed. Or if the size of the estimated effect is substantially smaller, or larger, than the original. Different criteria may be applicable for different circumstances. When testing a theory, statistical significance may the pertinent criterion. When estimating the effect of a policy intervention, effect size may be what's important.

Our approach is to rely on the assessment of the replicating author. If the author of the replication concludes that his/her research refutes the original article, we classify the replication outcome as "negative". If the author concludes that the replication confirms the original study, or the results are mixed or unclear, we categorize the replication as "positive" or "mixed/unclear", respectively. This approach assumes that the replicating author is in the best position to determine which results of the original study are most important, and which criterion (statistical significance, effect size) is most appropriate.

There is another reason for relying on the replicating author's assessment. In terms of "self-correcting science", what matters is how a replication is perceived by other researchers. Our view is that readers' perceptions of a replication are likely to be strongly influenced by the replicating author's assessment, especially when the replication has gone through a peer review process. If the replication concludes that it has confirmed/refuted another study, and that conclusion has passed the review of journal referees, then most readers are likely to take that as a reliable interpretation of the results.

In almost all cases, the replication authors' assessments were clearly stated in the abstract and/or conclusion of their papers. TABLE 1 gives two examples from each category. A complete list of how each of the replications in our study were classified, along with the corresponding authors' statements, is provided in a supplementary file. Of the 204 treated studies in our initial sample, 111 (54%) had negative replications, 41 (20%) had positive replications, and 52 (25%) were mixed.

TABLE 1 here

Selection of Controls: Stage 1. We collected the Scopus identification numbers for all of the replicated studies in our sample (the "Treated").[4] With this number, we were able to extract their corresponding year-by-year citation histories from Elsevier's API. From the same source we also extracted information about their year of publication, the journal in which they were published, and their volume and issue number[5]. Our selection procedure for finding control studies consisted of two stages. In the first stage, we collected a large pool of studies from which to select controls.

Our collection procedure used information about the "publication type" of the replicated study (i.e., "article" or "review article") and its "Field". The latter categories are quite broad. Examples include Economics and Econometrics; Finance; General Business, Management and Accounting; Public Health, Environmental and Occupational Health; Energy (miscellaneous); and Statistics, Probability and Uncertainty. Studies can be assigned to more than one field. For each "treated" study, we found all non-replicated studies that (i) were published in the same year, (ii) shared the same document type and field, and (iii) were published in a journal in which at least one of the 204 replicated studies also appeared.

We then extracted the citation histories for each of these. At the end of Stage 1, our sample consisted of 204 treated and 112,000 potential controls, though many of these controls were matched to more than one treated. Stage 2 of our selection process consisted of filtering through these potential controls to find control studies that "closely matched" the treated studies. Because this step is essential for assessing the reliability of our results, we describe this second stage in much detail.

[4] Originals without their own page in Scopus also were excluded

[5] For the replication papers, we extracted the year of publication from Scopus. For those replication papers not included in Scopus, we searched for the date of publication from other sources include the Replication Wiki pages and the journals themselves.

Selection of Controls: Stage 2. The goal of Stage 2 was to find control studies that closely matched the year-by-year citation histories of the replicated studies. This task was complicated by two factors. First, studies had different lengths of citation histories because they differed in how many years had passed between when the original was published and the replication was published. Second, studies differed in how many citations they had, with some studies having only a few citations, and others having hundreds. In general, it is harder to find close matches for studies with many replications.

FIGURE 1 here

Every treated study in our dataset was selected so that there were at least two years of citation history to match treated and controls. Correspondingly, there needed to be at least three years difference in the publication years of the replication and the original. For example, if an original study was published in 2014 and the replication was published in 2017, we searched for controls that were also published in 2014 and had identical or very similar citations in 2015 and 2016. FIGURE 1 plots a histogram of number of studies for each length of time between publication of the treated study and its replication. 176 treated, or 78% of the sample, had replications published 3 to 8 years after the originals. The remaining 49 studies (22%) had intervening periods of between 9 and 21 years. The differing time gaps between publication of the treated and its replication generate citation histories of different lengths on which to match treated and controls.

FIGURE 2 here

FIGURE 2 shows how we used the respective citation histories to match up controls with the treated. For each treated, we track the citations in the years between when it and its replication were published. We then take all the potential controls for the treated study from Stage 1 and compare citations over the same period. Matching is based on the year-by-year differences over the citation history.

We don't match on total citations. Rather, we choose controls that match the year-by-year record of citations for the original. Specifically, we take the absolute value of the difference in citations for each year and sum over the citation history. For example, suppose a treated study has 15 citations over a 3-year period equal to (2,5,8), and potential control studies A and B have citations (1,4,10) and (2,6,7). Both potential control studies have 15 citations over the period, but B more closely matches the year-by-year evolution of citations of the original.

For studies with a three-year difference between the publication years of the replication and the original (*K* = 3), we have two years of citation history to match on. For studies with a four-year difference (*K* = 4), we have three years of citation history. We follow this procedure for studies up to and including an eight-year difference. For studies with more than 8 years between publication of the original and its replication (*K* > 8), we only compare citation histories in the seven years preceding publication of the replication.

Thus, for each treated and potential control from Stage 1, we calculate the following sum of absolute differences for *K* = 3,4,5,6,7,8,

(1) $$TotAbsDiff_K = \sum_{k=1}^{K-1} \left| Citations_{T-k}^{Control} - Citations_{T-k}^{Treated} \right|$$

where *T* is the year the replication was published. For *K* > 8, we calculate

(2) $$TotAbsDiff_{GT8} = \sum_{k=1}^{7} \left| Citations_{T-k}^{Control} - Citations_{T-k}^{Treated} \right|$$

When $TotAbsDiff_K = 0$, then the year-by-year citation record of the control exactly matches the year-by-year citation record of the treated.

TABLE 2 here

Even with our large pool of potential control studies, it is difficult to get perfect matches. As shown in TABLE 2, there are only 2,201 perfect matches out of 112,000 possible controls. As a result, if we want a larger pool of control studies, we have to loosen the criterion for matching controls to treateds.

Our approach is to use a “sliding scale” matching criterion. For a treated with just a few citations at the time the replication study was published, we want the match to be exact or almost exact. For a treated with a lot of citations, we allow the year-by-year differences to be larger. Let $TotOrigCites_K$ measure the total number of citations for the treated up to (but not including) the year the replication was published,

(3) $TotOrigCites_K = \sum_{k=1}^{K-1} Citations_{T-k}^{Treated}$.

FIGURE 3 here

FIGURE 3 shows that the treated studies in our sample differ widely in the number of citations they had at the time the replication was published. 63 (31%) of the treated had less than 10 total citations at the time the replication was published. 87 (43%) had between 10 and 50 citations. On the other end, 27 (13%) had between 100 and a 1000 citations, and 3 (1%) had more than a 1000.

A control is matched with a treated whenever the sum of the absolute value of the year-by-year differences in citations are less than a given threshold, where the threshold is increased for original studies with more citations ($TotOrigCites_K$). Specifically, the decision rule matches a control with the treated if

(4) $TotAbsDiff_K \leq ceil(PCT \times TotOrigCites_K + 0.001)$

where *PCT = 0%, 10%,* or *20%*, and the *ceil(x)* function rounds up *x* to the nearest integer. TABLE 3 shows the threshold values corresponding to different values of $TotOrigCites_K$ and PCT.

TABLE 3 here

For example, when $PCT = 0\%$, the matching criterion states that the sum of year-by-year, absolute differences in citations between the treated and the control cannot be larger than 1 citation over the citation history period. The threshold value is the same no matter how many citations the treated has. This threshold rule disproportionately selects treated/control pairs with relatively few citations because there are many more studies with just a few citations compared to those with many citations (cf. FIGURE 3).

When $PCT = 10\%$, the matching threshold increases with $TotOrigCites_K$. For example, consider a treated and matched control that share a three-year citation history, where the treated study has a total of 20 citations. According to TABLE 3, in order to be a successful match, the potential control study can differ by no more than 3 citations over the citation history, or no more than an average of 1 citation per year. If, instead, the original study had 200 citations, the potential control could differ by no more than 21 citations, or an average of 7 citations per year. For $PCT = 20\%$, the threshold values are slightly less than twice as large compared to $PCT = 10\%$.

The foregoing matching rule produces a customized set of matches for each treated study. Each set of a single treated and its matched controls shares the same publication year and belongs to the same Scopus Field category. We call an individual set of a treated study and its matched controls an "issue", as in journal issue. The subsequent analysis will cluster on "issues." Our estimation strategy is to observe the difference in citations between each treated study and its matched controls in the years following publication of the replication.

This raises yet another issue. How many years should we track citations after the replication was published? There is a trade-off between length of post-replication period and number of treated. The longer the post-replication period, the fewer treated we have to study.

TABLE 4 here

This trade-off is evident in TABLE 4. 88 (43%) of the treated studies in our sample have 10 or more post-replication years of available citation data. 161 (79%) have 5 or more years, and 204 (100%) have 3 or more years. The subsequent analysis focuses on studies that have at least 3 years of post-replication citation data. However, we perform an identical analysis for studies having at least 5 years of post-replication data. None of our conclusions are altered when using this alternative sample of observations.

TABLE 5 here

TABLE 5 reports the number of treated and matched controls for each value of *K* and matching criterion *PCT.* The first thing to note is that we lose a lot of treated studies when we require good matches. For example, when we require that each treated and matched control pair differ by no more than 1 citation over their respective citation histories (*PCT = 0%*), the number of corresponding treated falls from 204 to 75. If we loosen the matching criteria to *PCT = 10%* and *20%,* the number of treated is somewhat larger at 110 and 167 studies, respectively, but still falls short of 204. Further, if we require 5 years of post-replication data instead of 3, the numbers fall to 55, 82, and 130 treated, respectively.

Given the paucity of studies having more than 8 years between publication of original and replication, and to facilitate comparison across the different matching criteria (*PCT=0%, 10%,* and *20%*), our subsequent analysis will focus on the samples with *K* = 3 to 8.

TABLE 6 here

TABLE 6 reports on the closeness of the matches for the three different matching criteria. As expected, matches are very close when *PCT=0%.* The maximum absolute deviation over the entire citation history for the 7,044 controls in the sample *K*=3 through 8 is 1 citation. The mean absolute deviation is 0.69 citations. When we loosen the matching criterion to *PCT=10%*, adding an additional 508 controls, the mean rises slightly to 0.82 citations. 90% of the controls in the *PCT = 10%* sample have a total absolute deviation of 1 citation or less over the citation history period. Loosening the criterion further to *PCT=20%* adds another 3,650 controls. However, the additional controls comes at the cost of poorer matches. The mean absolute deviation rises to 1.76 citations. While the median deviation is still 1 citation and 75%

of the controls differ by 2 citations or less, the worst 5% of matches deviate by 6 or more citations, and the worst 1% deviate by 13 citations or more.

TABLE 7 here

A consequence of selecting treatments and controls based on closeness of match is that we disproportionately select studies with fewer citations. This occurs because it is harder to match studies that have many citations. This is evident in TABLE 7. The first column reports quantile values of total citations for the full set of 204 treated studies at the time the replication was published. The 25th, 50th, and 75th quantile values for total citations of the treated are 8, 23, and 54.5 citations, respectively.

The subsequent six columns report quantile values of total citations for the matched set of treated and controls that correspond to the three matching criteria (*PCT* = 0%, 10%, and 20%). For example, when imposing the requirement that treated and controls differ by no more than 1 citation over their respective citation histories (*PCT* = 0%), the matched treated and controls have 25th, 50th, and 75th quantile values of 3, 6, and 12; and 1, 2, and 4 citations, respectively. Note that the quantile values for the controls are less than the treated. This illustrates that the controls have a disproportionate number of studies with relatively few citations; an outcome of the fact that it is easier to find controls for treated that do not have many citations.

Sample selection bias. Sample selection bias occurs when causation runs from the error term to the treatment variable. This study investigates two "treatments": (i) being replicated, and (ii) having a negative replication, where we are primarily interested in the effect of the latter treatment. The error term represents unexplained differences in citations between replicated studies and controls.

With respect to (i), sample selection bias would occur if the likelihood a study was selected for replication was related to the number of citations it would receive. In other words, it was the type of studies being selected for replication that caused differences in citation rates, rather than the act of replication itself. We address this concern by matching citations during the pre-treatment period. The better the match during the pre-treatment period, the smaller the potential problem with sample selection. Even after matching on citation histories, it is still possible for sample selection to occur if replicators chose studies that they correctly anticipated would receive more citations in the future compared to non-replicated studies *with identical citation histories*. This would positively bias estimates of the causal effect of replication on citations. We discuss this further below.

With respect to (ii), sample selection bias would occur if the likelihood a study had a negative replication was related to the number of citations it would receive. This is less likely to be a problem than (i) because the outcome of the replication is unknown at the time the original study was chosen for replication. For there to be sample selection bias, it would have to be the case that there was some feature of original studies chosen for replication that caused them to be cited systematically differently than other studies with identical citation histories, and this feature was correlated with the outcome of the replication, an outcome that was not observed at the time the original study was selected. It is difficult to imagine scenarios where this is likely to be a case, especially since both negative and positive replications are matched to controls with identical or near identical numbers of citations in the pre-treatment period.

III. Results: The Effect of Replications on Citations

Before we estimate the effect of a negative replication, which is the main focus of our study, we first investigate the overall difference in citations between replicated and matched controls. We define the difference in citations such that positive differences indicate that the treated study has more citations than its matched control in a given year *t*.

$$(5) \qquad DIFF_{it,i\in K} = Citations_{it,i\in K}^{Treated} - Citations_{it,i\in K}^{Control}$$

We estimate the following regression for each year of the seven year period: t = *-3,-2,…,2,3;* encompassing the three years before the replication was published, the year the replication was published, and the three years after the replication was published.

(6) $DIFF_{it,i\in K} = \beta_0 + \varepsilon_{it,i\in K}$

We expect $\beta_0 = 0$ for t = -3,-2,-1 if our matching criteria are effective in selecting good controls. We note that β_0 will equal at least 1 for t = 0, ceteris paribus, because the treated study is always cited by the replication study.

In selecting an estimator, we note that the construction of the dependent variable in Equation (6) induces a correlation in all observations from the same "issue". This occurs because each observation from the same issue shares the same $Citations_{K,t}^{Treated}$ value. Appropriate estimators need to accommodate this clustering. Accordingly, we begin by reporting results using a hierarchical linear model (HLM) estimator with robust standard errors that cluster on issue. This allowed us to incorporate within-cluster heterogeneity while also addressing their associated lack of independence. Later on we consider a variety of alternative estimators.

FIGURE 4 here

HLM uses maximum likelihood and assumes normality, particularly in the dependent variable. FIGURE 4 plots histograms for $DIFF_{it,i\in K}$ for the combined samples of *K = 3,4,…,7,8* and *PCT = 0%, 10%,* and *20%.* The distributions are symmetric and approximately normally distributed. Intra-class correlations for each of the three samples are 0.454, 0.630, and 0.489, respectively, so HLM estimation seems appropriate.

TABLE 8 here

TABLE 8 reports the associated estimates. Looking first at the pre-replication period, *t* = -3,-2, and -1, we see that differences exist even after matching. For example, under the *PCT = 20%* matching regime, treated received 0.319 more citations, on average, than their matched controls three years before the replication was published (t = -3*).* At t = -2 and t = -1, they received 0.454 and 0.550 additional citations. The latter two values are statistically significant at the 1-percent level. In contrast, the differences are substantially smaller for the *PCT = 0%* and *PCT = 10%* matching regimes. For example, under *PCT = 0%,* treated received 0.125, 0.088, and 0.087 more citations than controls in periods t = -3, t = -2, and t = -1, respectively.

Statistically significant differences in the pre-replication period raise concerns about balance in citations between treated and controls. They suggest that observed differences in the post-replication period may be carryovers from the pre-replication period. Accordingly, while we continue to report results for *PCT = 20%*, our subsequent discussion will focus on the cases *PCT = 0%* and *10%* as the pre-replication differences are generally smaller. However, even for these two cases, there are some significant differences between treated and controls in the pre-treatment period. As a result, one should be careful about attaching causal interpretations of the subsequent results.

Turning now to the post-replication period (t = 1, 2, and 3) we estimate that replicated studies receive 1.8 to 2.5 (*PCT = 0%*) and 2.9 to 5.3 (*PCT = 10%*) additional citations a year compared to their matched controls. Each of the six estimated coefficients are significant at the 5% level, with five significant at the 1% level. The estimated effects are relatively large in size. Rows (8) and (9) report the mean and median values of total citations for the 74 and 103 treated studies, respectively, at the time their replication was published. These are 2.9 and 2 citations, and 4.2 and 2 citations, respectively. Thus, yearly increases in citations of the order of 2 to 5 are quite large, almost implausibly large. This is of some concern and we explore this further below.

Why are replicated studies more likely to be cited than their matched, unreplicated controls? One possibility is that replications raise awareness of the replicated studies. Raised awareness could come in the form of readers of the replication learning of the existence of the replicated study where they otherwise would have been unaware. Another possibility is that readers update the importance they attach to a study when they see it replicated, and hence are more likely to cite it. A third possibility is that the estimated effect is not causal. It may be that the reason the studies were selected for replication in the first place is because the results were of larger interest to the discipline than the matched controls, even after controlling for citations during the pre-treatment period.

Lastly, we consider the estimated effect of being replicated when *t* = 0; that is, in the year the replication study is published. Our estimates indicate that treated studies receive between 2.7 and 3.6 more citations at time *t* = 0 than their matched controls. However, it must be remembered that these numbers include the citation from the published replication. Thus a better estimate would be 1.7 to 2.6 citations. Is it reasonable that replications could affect citations of the original study in the same year the replication was published? While some of this may be attributed to carryover from the pre-replication period, we suspect that most of this increase is due to the replications having been circulated as working papers prior to publication. This would give time for readers of the replication to attract readers, gain increased awareness of the original study, and cite it in their own research.

IV. Results: The Effect of Negative Replications on Citations

The primary focus of this study is to estimate the impact of a negative replication. Our measure of effect uses the same dependent variable as above: the difference in citations between the treated and the matched controls in a given year *t*, with positive values indicating that the treated study receives more citations. We estimate the following regression,

(7) $$DIFF_{it,i\in K} = \beta_0 + \beta_1 NEGATIVE_{it,i\in K} + \varepsilon_{it,i\in K} \,,$$

for t = -3,-2,-1, 0, 1, 2, 3, where *NEGATIVE* is a dummy variable that takes the value 1 if the treated study in $DIFF$ was refuted by the associated replication study, and 0 if it was confirmed or the results were mixed. The treatment effect is measured by β_1. It can be thought of as a difference-in-difference estimator. It measures the difference in citations between treated and controls for replicated studies with negative replications minus the difference in citations between treated and controls for replicated studies with positive/mixed replications. If negative replications adversely affect a study's citations, β_1 will be negative for $t > 0$. To estimate Equation (7) we again use a hierarchical linear model, clustering at the level of issues, with robust (clustered) standard errors. We further allow β_1 to be random, allowing negative replications to have different effects for different issues.

As before, we estimate separate regressions for each time period, starting from three years before the replication was published (t = -3) to three years after (t = 3). We expect $\beta_1 = 0$ for t = -3,-2,-1 because the replication had not yet been published during this time period. This provides a further "balancing" check that our matching process has not biased the selection of controls to produce post-replication citation results that continue pre-replication citation behaviour.

In addition to estimating separate regressions for each year, we also pool the yearly observations to allow us to conduct multi-year tests of treatment effects. Specifically, we estimate

(8) $DIFF_{it,i\in K} = \sum_{t=-3}^{3} \beta_{0t} \times T(t)_{it,i\in K} + \sum_{t=-3}^{3} \beta_{1t} NEGATIVE_{it,i\in K} \times T(t)_{it,i\in K} + \varepsilon_{it,i\in K}$,

where $T(-3)$ through $T(3)$ are dummy variables that take the value 1 when t = *-3,-2,…,2,3,* respectively.

We test for an overall, post-replication treatment effect by testing the null hypothesis:

(9) $H_0 = \sum_{t=1}^{3} \beta_{1t} = 0$.

We also test for an overall pre-replication "treatment effect" by testing the null hypothesis:

(10) $H_0 = \sum_{t=-3}^{-1} \beta_{1t} = 0$

We expect $\sum_{t=-3}^{-1} \beta_{1t} = 0$ if we can assume that the outcome of the to-be-published-later replication study were unknown during the pre-replication period. TABLE 9 reports the associated results. Since this section focuses on the effect of negative replications on citations, the table only report estimates for β_1 in Equation (7).

TABLE 9 here

Our expectation that the estimated effects of a negative replication would be zero during the pre-replication period (*t* = -3,-2,-1) is confirmed. All of the estimated coefficients are small in size. For example, when *PCT = 0%* and *t* = -3, we estimate a mean difference of 0.076 citations between studies with negative replications and studies with positive/mixed replications. Of the six estimated coefficients associated with the pre-replication periods for *PCT = 0%* and *PCT = 10%*, four are positive and two are negative; all are statistically insignificant. Row (8) in the table presents the results of a test of an overall pre-replication effect. We fail to reject the hypothesis that the sum of the estimated effects during the pre-replication periods is equal to zero with p-values well above 0.05 (p = 0.605 and 0.320). These results are consistent with the assumption of random assignment of treatment.

Turning to the post-replication period, we find no evidence that negative replications impact the amount of citations received by replicated studies. While the estimated effects are generally larger in absolute value compared to the pre-replication period, they are again all statistically insignificant. Of the 6 associated estimates for *PCT = 0%* and *PCT = 10%*, four are positive and two are negative. When we perform a test of overall significance of the estimated treatment effects in the post-replication period (cf. Row 9), we cannot reject the null hypothesis that the cumulative effects over this period are zero. The associated p-values are 0.170 and 0.775. Further, in contrast with our results from TABLE 8, we more confident that the estimates of TABLE 9 represent causal effects.

The only statistically significant, estimated treatment effect occurs when *t = 0*, but only for our strictest matching criterion, *PCT = 0%*. For that case, we estimate a positive citation effect of a negative replication of 2.3 citations. As discussed above, we attribute estimated treatment effects associated with replications at time *t = 0* to the fact that these studies likely circulated prior to publication as working papers.

The regression results for all three samples in TABLE 9 hint that the effects of negative replications may turn negative over time. With only three years of post-replication observations, however, any observed patterns are potentially misleading. TABLE 10 repeats our analysis, this time with all treated and matched controls that have at least five years of post-replication data.

TABLE 10 here

The main results concerning pre- and post-replication effects remain the same so we skip over them and instead inspect the estimates in Rows (5) through (9). None of the estimated coefficients for times *t = 1* to *5* are statistically significant. None of the three samples (*PCT = 0%, 10%,* and *20%*) show evidence of declining estimates over time. In fact, the *PCT = 0%* sample produces positive estimates for each time period, with the largest estimated effect occurring in the final period (*t* = 5). Given the large standard errors, we cannot rule out the possibility of a declining trend, but there is no evidence that adverse effects from negative replications get stronger over time.

One of the concerning observations from TABLES 8 and 9 is the large size of the estimates. Specifically, it seems improbable that being replicated can add 2 to 5 additional citations *a year* to an article when that is approximately equal to the *total* number of citations the article had at the time it was replicated (cf. Rows 8 and 9 in TABLE 8). A possible explanation is that the estimates are being driven upwards by studies with relatively many citations. To address this, we re-estimate the specifications in TABLES 8 and 9 with quantile regression (Chamberlain, 1994; Koenker, 2005). The associated estimates reflect how variables relate to the median, rather than the mean, of the dependent variable, which makes them less influenced by extreme values.

There is another way our sample may produce a misleading picture of the effects of a replication/negative replication. As noted above, not all the treated studies have the same number of matches. Studies with few citations are easier to match, and thus have more controls. Using individual observations implicitly gives greater weight to these studies. To address this problem, we collapse the multiple observations associated with each treated study into a single observation, so that the observation now represents mean values of the respective variables (similar to how a "between estimator" works).

A final change we make recognizes that some of the control studies are used for more than one treated study. The degree of overlap isn't large. Of the 7,044, 7,552, and 11,202 control studies in our three subsamples, 6,571, 7,056, and 10,330 are unique. This implies that approximately 5-8% of the control studies are matched to more than one treated, violating the assumption of observation independence. To address this problem, we bootstrap the standard errors.

TABLE 11 here

TABLES 11 and 12 report the results of re-estimating the specifications of TABLES 8 and 9 using quantile regression. Looking first at the effect of replication in TABLE 11, whereas we previously found significant differences between treated and control studies in the pre-treatment period, we now find no significant differences for the *PCT* = 0% and *PCT* = 10% samples. In fact, the estimated median difference in citations during this period is zero for both samples and all three time periods. This differs from the *PCT* = 20% sample, where two of the differences are positive, one of which is significant. As a result, we continue to focus on the *PCT* = 0% and 10% samples.

Rows (5) through (7) report estimates of the effect of replication on the original studies' citations. Compared to TABLE 8, all of the estimates are smaller, ranging from 0.5 additional citations per year to 2.0 additional citations. Not only has quantile regression produced smaller estimates, but the measures of total citations prior to the replication being published are larger. Mean total cites range from 7.8 to 19.5 citations, and median total cites range from 4.7 to 8.4 citations.

The reason for the larger numbers is the HLM estimates from TABLE 8 were based on individual observations, and there were more studies with fewer citations because these were easier to match. In the quantile regressions, these were collapsed into a single value for each treated observation, which produced a total citation profile closer to that of the treated studies. In summary, the estimates from TABLE 11 find evidence of a positive citation effect from being replicated, but the effects are small, ranging from 0.5 to 2.0 citations per year. These compare to mean and median total citations of 7.8-19.5 and 4.7-8.4, respectively.

TABLE 12 here

We next turn to quantile regression estimates of the effect of a negative replication on citations (cf. Equation 7). These are reported in TABLE 12, where once again we only report estimates for β_1, the coefficient on the *NEGATIVE* dummy variable. As before, and as expected, the estimates of a negative replication in the pre-replication period is close to zero and statistically insignificant.

The estimates in the post-replication period range from -0.406 to 1.667 citations per year. All are insignificant except for the estimate of 1.667 at $t = 2$ for sample *PCT = 0%.* Also as before, the tests of overall effect during the pre- and post-replication periods are insignificant. There continues to be no evidence that a negative replication has an adverse effect on the citations received by the original article. While the estimates from TABLES 8 and 9 are generally consistent with those from TABLES 11 and 12, we prefer the latter because of the econometric problems they address and the fact that the associated estimates seem more reasonably sized.

It is worth emphasizing that absence of evidence is not evidence of absence: statistical insignificance is a function of the power to detect a significant effect. Can we say anything about the statistical power of the estimated treatment effects in TABLE 12? While the inadequacies of ex post power calculations are well-known (Hoenig & Heisey, 2001; Levine & Ensom, 2001; Yuan & Maxwell, 2005), these stem primarily from variability in the estimated effect. As McKenzie & Ozier (2019) note, the estimated standard error is far less noisy. Accordingly, they suggest using estimated standard errors to calculate (ex post) statistical power. Tian (2022) investigated the performance of ex post power analyses based on estimated standard errors. In his experiments, he found that 95% sample ranges for estimated ex-post power estimates generally ranged between 70-90% over a large span of experimental conditions when the true power was 80%.

With this variability in mind, we calculate ex post estimates of statistical power using the standard errors associated with the estimated treatment effects in TABLE 12. We consider two effect sizes: (i) a treatment effect of 1 citation per year, or a cumulative effect of 3 citations over three years; and (ii) a treatment of 2 citations per year, or a cumulative effect of 6 citations over three years. We do this both for the *PCT = 0%* and *PCT = 10%* regressions. These are reported in TABLE 13.

For an effect size of (1 citation/year, 3 citations/3 years), none of our estimates of statistical power achieve 80%. The closest is the "Test of overall post-replication effect" for matching criterion *PCT = 0%.* We estimate the probability of estimating a significant effect for this effect size to be 64% (but remember the wide "confidence intervals" around this number reported by Tian (2022)). For an effect size of (2 citations/year, 6 citations/3 years), the estimates of statistical power are substantially larger. The associated "Test of overall post-replication effect" is estimated to have power of 100% and 91% for the *PCT = 0%* and *10%* regressions. The power estimates for the individual year effects are somewhat less.

How should we interpret these ex post estimates of statistical power? They provide evidence that our statistical design is generally sufficient to identify a cumulative effect size of 6 citations/3 years for in both the *PCT = 0%* and *PCT = 10%* regression; and (2 citations/year) for the individual year effects in the *PCT = 0%* regression. For effect sizes much smaller than this, one is likely to obtain an insignificant estimate even if an effect is present.

V. Further Robustness Checks

The preceding analysis considered a number of robustness checks. Three different matching criteria were used (*PCT = 0%, 10%,* and *20%*). Two different estimators were employed (HLM and quantile regression). Further, while most of the analysis focused on a three-year post-replication period, we also re-estimated our negative replication specification with a five-year post-replication period. The longer post-replication period entailed using a smaller set of treated and control studies, providing another robustness check over sample composition.

In this last section we summarize the results of additional robustness checks. Specifically, we use three alternative estimators: (i) OLS with cluster robust standard errors; (ii) OLS-averaged with heteroscedasticity robust standard errors, where individual differences in citations between treated and controls are averaged for each treated study; and (iii) random effects with cluster robust standard errors, where the multiple observations per treated are organized as unbalanced panel data. Each of these were estimated for both 3- and 5-year post-estimation periods, and for matching criteria of *PCT = 0%* and *10%.* In addition, any combinations involving quantile or HLM regressions that were not previously reported were also estimated.

We did this separately both for estimates of the overall effect of replication on citations (cf. TABLES 8 and 11), and the effect of negative replications on citations (cf. TABLES 9, 10, and 12). If reported in table form, these estimates would produce 16 and 14 additional tables of results, respectively. To facilitate interpretation of such a large number of estimates, we produce time series graphs of the respective treatment effects over the pre- and post-treatment periods, with confidence intervals. These are reported in APPENDIX A (effect of negative replication) and APPENDIX B (overall effect of replication).

The additional robustness checks confirm the previous results. The estimated effects of a negative replication are almost always statistically insignificant, and almost always positive. Of 58 estimated post-treatment effects: 54 are statistically insignificant, and 47 are positive, with the four significant estimates being positive. None of the cumulative, post-treatment effects are statistically significant. With respect to the overall effect of a replication on citations, almost all the individually estimated post-treatment effects are statistically significant, and all are positive: 68 out of 68 estimated post-treatment effects are positive, and 60 are statistically significant.

V. Conclusion

This study examined the effect of negative replications on the citation rates of replicated studies. We study a set of 204 replicated studies and compare their citation performance with an initial sample of 112,000 potential controls taken from Scopus. Using matching criteria that accommodate differences in the lengths of time between publication of the original study and its replication, as well as differences in the number of citations across studies, we match each replicated study with multiple controls based on having comparable citation histories prior to publication of the replication.

Our main finding is that there is no evidence that studies that receive negative replications suffer a penalty in the form of fewer citations. This result is robust across many samples and estimation procedures. It is robust if we use a three-year post-replication period or a five-year post-replication period; whether we restrict our sample to the closest matches (*PCT = 0%*), or allow looser matching criteria (*PCT = 10%* or *20%*); whether we use hierarchical linear model estimation, panel data random effects, OLS-cluster estimation, or quantile regression. It is robust if we estimate separate effects for each year relative to when the replication was published, or whether we pool the data in a window around following the replication publication date. In any and every circumstance, we find no evidence of a citation penalty for studies whose findings are later refuted by replications. Relatedly, there is no evidence that any adverse effects of negative replications gather strength over time.

We also find that studies that are replicated receive significantly more citations than their matched control studies. Our best estimates place the size of the effect between 0.5 and 2.0 additional citations a year. This compares to mean and median total citations at the time the replication was published of 7.8 to 19.5 citations, and 4.7 to 8.4 citations, respectively. However, causal interpretations are on less secure footing when evaluating these estimates.

Can our results be interpreted as evidence that science is not “self-correcting”? There are alternative interpretations. One possibility is that science is self-correcting when researchers are aware of a replication study, but researchers were not familiar with the results from the replication studies we studied. If a replication produces a negative result, but researchers are unaware of its existence, one would not expect to see any effect. The problem with this explanation is that we observe statistically significant, higher citation rates for studies that have been replicated. While the effect is not large, it does suggest that replications are being read.

Another candidate explanation is that negative replications are not persuasive. Just because a replicating author declares that his/her paper has refuted a previous study does not mean that the research community agrees. Still, one would think that relative to a positive replication, a negative replication would convey less confidence in the findings of a study; and less confidence would translate into fewer citations.

Some researchers argue that citations are not well-suited to play a “self-correcting” role. In their study of citations, Aksnes et al. (2019) write the following:

> One might think that in cases where the solidity or plausibility is assessed as poor, the work will not be considered as worth citing (i.e., will be neglected), and in cases where more than one study shows similar results, an author may choose to cite the study she perceives as the most solid. As a consequence, solidity/plausibility—as perceived at the time of citing—may to a certain extent be reflected in citation patterns. There is, however, little knowledge about the extent to which this actually is the case, and (as explained in "Understanding Citations" section) studies of citation behavior have identified a multitude of factors that are not per se associated with the solidity of the studies. *Therefore, it seems unlikely that citations can be seen as valid indicators of the solidity of the publications* [italics added].

The findings of this study are consistent with the view that researchers cite papers for many reasons, some of which are unrelated to the "solidity or plausibility" of a study. If that is the case, then whatever services replications may play in science, self-correction of unreliable results is not one of them. The issue is an important one. If replications do not play a self-correcting role in science, then what does? Where is the avenue that leads from discredited findings to reduced influence? That remains a topic for future research.

References

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*, 9(1), 2158244019829575.

Anderson, L. B., & Delgado, M. S. (2010). Another round of fraternity membership and binge drinking. *Journal of Economic and Social Measurement*, 35(1-2), 129-147.

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1), 57-67.

Azoulay, P., Furman, J.L., Krieger, J.L., & Murray, F.E. (2015). Retractions. *Review of Economics and Statistics,* 97(5), 1118-1136.

Azoulay, P., Stuart, T., & Wang, Y. (2014). Matthew: Effect or fable? *Management Science*, 60(1), 92-109.

Baltagi, B. (2010). Narrow Replication of Serlenga and Shin (2007) gravity models of intra-EU trade: application of the CCEP-HT estimation in heterogeneous panels with unobserved common time-specific factors. *Journal of Applied Econometrics*, 25(3), 505-506.

Bar-Ilan, J. & Halevi, G. (2018) Temporal characteristics of retracted articles. *Scientometrics,* 116(3): 1771–1783.

Bornemann-Cimenti H, Szilagyi IS, & Sandner-Kiesling A. (2016). Perpetuation of Retracted Publications Using the Example of the Scott S. Reuben Case: Incidences, Reasons and Possible Improvements. Sci Eng Ethics. Aug;22(4):1063-1072. doi: 10.1007/s11948-015-9680-y. Epub 2015 Jul 7. PMID: 26150092.

Budd, J. M., Sievert, M., & Schultz, T. R. (1998). Phenomena of retraction: reasons for retraction and citations to the publications. *JAMA*, 280(3), 296-297.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer,, Altmejd, T. A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa,, Heikensten, A. E., Hummer, L., Imai, T., Isaksson, S, Manfredi, D., Rose, J., Wagenmakers, E.-J., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.

_______________ . (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644.

Candal-Pedreira, C., Ruano-Ravina, A., Fernández, E., Ramos, J., Campos-Varela, I., & Pérez-Ríos, M. (2020). Does retraction after misconduct have an impact on citations? A pre–post study. *BMJ Global Health*, 5(11), e003719.

Cawley, J., Markowitz, S., & Tauras, J. (2004). Lighting up and slimming down: the effects of body weight and cigarette prices on adolescent smoking initiation. *Journal of Health Economics*, 23(2), 293-311.

Chamberlain, G. (1994). *Quantile regression, censoring, and the structure of wages*. In Advances in Economics Sixth World Congress, ed. Christopher A. Sims, 171-209. Cambridge University Press: Cambridge.

DeSimone, J. (2007). Fraternity membership and binge drinking. *Journal of Health Economics*, 26(5), 950-967

Devereux, P. J., & Hart, R. A. (2010). Forced to be rich? Returns to compulsory schooling in Britain. *The Economic Journal*, 120(549), 1345-1364.

Fernández, L. M., Hardwicke, T.E., & Vadillo, M. A. (2021). Retracted papers die hard: Diederik Stapel and the enduring influence of flawed science. PsyArXiv. https://doi.org/10.31234/osf.io/cszpy

Furman, J. L., Jensen, K., & Murray, F. (2012). Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy*, 41(2), 276-290.

Hamoudi, A. (2010). Exploring the causal machinery behind sex ratios at birth: does hepatitis B play a role?. *Economic Development and Cultural Change*, 59(1), 1-21.

Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., van den Akker, O. R., Nuijten, M. B., & Ioannidis, J. P. (2021). Citation patterns following a strongly contradictory replication result: Four case studies from psychology. *Advances in Methods and Practices in Psychological Science*, 4(3), 25152459211040837.

Höffler, J. H. (2017). Replicationwiki: Improving transparency in social sciences research. *D-Lib Magazine*, 23(3), 1.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 19-24.

Hsiao, T. K., & Schneider, J. (2021) Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies*, 1-53. https://doi.org/10.1162/qss_a_00155

Jin, G. Z., Jones, B., Lu, S. F., & Uzzi, B. (2019). The reverse Matthew Effect: Catastrophe and consequence in scientific teams. *The Review of Economics and Statistics*, 101(3), 492–506.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press: New York.

Levine, M., & Ensom, M. H. (2001). Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 21(4), 405-409.

Lu, S. F., Jin, G. Z., Uzzi, B., & Jones, B. (2013). The retraction penalty: Evidence from the Web of Science. *Scientific Reports*, 3(1), 1-5.

McKenzie, D. & Ozier, O. (2019, May 16). Why ex-post power using estimated effect sizes is bad, but an ex-post MDE is not. *World Bank Blogs: Development Impact*. https://blogs.worldbank.org/impactevaluations/why-ex-post-power-using-estimated-effect-sizes-bad-ex-post-mde-not

Nakov, A. (2010). Jackknife instrumental variables estimation: replication and extension of Angrist, Imbens and Krueger (1999). *Journal of Applied Econometrics*, 25(6), 1063-1066.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review*, 96(1), 152-175.

Oster, E. (2005). Hepatitis B and the case of the missing women. *Journal of Political Economy*, 113(6), 1163-1216.

Piller, Charles (2021). Many scientists citing two scandalous COVID-19 papers ignore their retractions. *Science*. https://doi.org/10.1126/science.abg5806

Rees, D. I., & Sabia, J. J. (2010). Body weight and smoking initiation: Evidence from Add Health. *Journal of Health Economics*, 29(5), 774-777.

Retraction Watch (n.d.). Top 10 most highly cited retracted papers. Retrieved from https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/

Schafmeister, F. (2021). The Effect of Replications on Citation Patterns: Evidence From a Large-Scale Reproducibility Project. *Psychological Science*, 32(10), 1537-1548.

Schneider, J., Ye, D., Hill, A. M., & Whitehorn, A. S. (2020). Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics*, 125(3), 2877-2913.

Serlenga, L., & Shin, Y. (2007). Gravity models of intra-EU trade: application of the CCEP-HT estimation in heterogeneous panels with unobserved common time-specific factors. *Journal of Applied Econometrics*, 22(2), 361-381.

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, 7(21), eabd1705.

Tatsioni, A., Bonitsis, N. G., & Ioannidis, J. P. (2007). Persistence of contradicted claims in the literature. *JAMA*, 298(21), 2517-2526.

Tian, J. (2022). *An Analysis of Statistical Power in Empirical Economics*. PhD thesis, University of Canterbury (unpublished).

Yang, Y., Youyou, W., & Uzzi, B. (2020). Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(20), 10762-10768.

Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, 30(2), 141-167.

TABLE 1:
Examples of Replication Assessments

Original	Replication	Assessment	Statement from Paper
Oster (2005)	Hamoudi (2010)	Negative	“I find that repeating Oster’s original analysis in a different data set—one that is better suited to addressing the question—produces strikingly different results” (page 2)
Oreopoulos (2006)	Devereux & Hart (2010)	Negative	“Re-analysing this dataset, we find much smaller returns of about 3% on average with no evidence of any positive return for women” (page 1345)
Cawley et al. (2004)	Rees & Sabia (2010)	Positive	“...we reexamine the relationship between body weight and smoking initiation. Our results are generally consistent with those of Cawley, Markowitz and Tauras” (page 774)
DeSimone (2007)	Anderson & Delgado (2010)	Positive	“This paper describes a successful attempt to replicate DeSimone” (page 129)
Angrist et al. (1999)	Nakov (2010)	Mixed	“I replicate Angrist et al.' s Monte Carlo simulations in Table I for Models 1, 2, 4, and 5, as well as their estimates of returns to schooling in Table II. I am unable to replicate the authors' Carlo results for Model 3” (page 1063)
Serlenga &Shin (2007)	Baltagi (2010)	Mixed	“While most of the estimates remain about the same…Their conclusion that the HT estimate…is fragile” (page 505)

TABLE 2:
Perfect Matches by Number of Years Difference between Publication of Original and Replication

Years Difference	Number of Control Studies	Percent of Total Perfect Matches
3	1,204	54.7%
4	99	4.5%
5	425	19.3%
6	3	0.1%
7	466	21.2%
8 or more	4	0.2%
Total	2,201	100.0%

NOTE: A “perfect match” is defined by $TotAbsDiff = 0$ (see Equations 1 and 2 in the text).

TABLE 3:
Threshold Values for $TotAbsDiff_K$ for Various Combinations of $TotOrigCites_K$ and PCT

		TotAbsDiff	
TotOrigCites	***PCT* = 0%**	***PCT* = 10%**	***PCT* = 20%**
0	1	1	1
10	1	2	3
20	1	3	5
50	1	6	11
100	1	11	21
200	1	21	41
1000	1	101	201
2000	1	201	401

NOTE: Threshold values are calculated using Equation (4) in the text.

TABLE 4:
Number of Treated by Years of Post-Replication Data

Years of Post-Replication Data	Number (Frequency)	Number (Cumulative)
3	22	204
4	21	182
5	16	161
6	12	145
7	17	133
8	15	116
9	13	101
10	18	88
11	4	70
12	8	66
13	5	58
14	6	53
15	4	47
16	8	43
17	3	35
18	4	32
19	8	28
20	3	20
21	4	17
22	1	13
24	1	12
27	2	11
28	1	9
29	1	8
30	1	7
31	2	6
34	1	4
37	1	3
38	1	2
42	1	1

NOTE: The values in the table report the number of treated studies for which we have the given years of post-replication data. We highlight 3 and 5 because our two main samples are constructed to have at least 3- and 5-years, respectively, of citation data following publication of the replication study.

TABLE 5:
Number of Originals and Matched Controls for Different Values of *K* and *PCT*

K	*Matching Criteria* *PCT = 0%* *(Treated/Controls)*	*PCT = 10%* *(Treated/Controls)*	*PCT = 20%* *(Treated/Controls)*
3	34/4,553	38/4,873	39/6,873
4	16/662	21/791	26/1,791
5	8/940	17/976	21/1,284
6	8/72	14/87	21/260
7	4/772	7/778	19/857
8	4/45	6/47	16/137
3-8	74/7,044	103/7,552	142/11,202
>8	1/1	7/9	25/146
ALL	75/7,045	110/7,561	167/11,348

NOTE: *K* is defined as the difference in years between the publication of the replication and the original. *PCT* adjusts the matching criteria based on the total number of citations a study has at the time the replication was published (see Equation 4 in the text and corresponding discussion). The table reports the numbers of treated and controls for each pair of (*K/PCT*) values. We highlight the row *K* = *3-8* because we focus on this sample in our reporting and discussion of results.

TABLE 6:
Distribution of $TotAbsDiff_{38}$ for Different Matching Criteria

	Matching Criteria		
	PCT = 0%	***PCT = 10%***	***PCT = 20%***
Min	0	0	0
1%	0	0	0
5%	0	0	0
10%	0	0	0
25%	0	0	1
50%	1	1	1
75%	1	1	2
90%	1	1	3
95%	1	2	6
99%	1	3	13
Max	1	17	34
Mean	0.689	0.820	1.760
N	7,044	7,552	11,202

NOTE: This table reports distribution statistics for the total, absolute value of the annual differences in citations between treated and controls for the three samples defined by *K* = 3-8 and *PCT* = 0%, 10%, and 30%; where *K* is defined as the difference in years between the publication of the replication and the original, and *PCT* adjusts the matching criteria based on the total number of citations a study had immediately prior to when the replication was published (see Equation 4 in the text and corresponding discussion).

TABLE 7:
Distribution of Total Citations at Time Replication Published for Treated and Matched Control Studies for Different Matching Criteria

	FULL SAMPLE: *Treated*	**SUBSAMPLES**					
		PCT = 0%		***PCT = 10%***		***PCT = 20%***	
		Treated	***Controls***	***Treated***	***Controls***	***Treated***	***Controls***
Min	0	0	0	0	0	0	0
1%	0	0	0	0	0	0	0
5%	2	0	0	0	0	1	0
10%	3	1	0	2	0	2	0
25%	8	3	1	3	1	5	1
50%	23	6	2	9	2	15.5	4
75%	54.5	12	4	26	5	38	8
90%	138	19	7	48	10	77	18
95%	355	22	9	77	13	108	33
99%	1131	48	14	138	40	171	77
Max	2239	48	50	180	192	180	234
Mean	80.6	7.9	2.9	20.2	4.2	29.3	8.2
N	204	74	7,044	103	7,552	142	11,202

NOTE: The table reports distribution statistics for the four samples: (i) the full sample of 204 treated studies, and the three analysis samples defined by (*K/PCT*) = (3-8/0%), (3-8,10%) and (3-8,20%), where *K* is defined as the difference in years between the publication of the replication and the original, and *PCT* adjusts the matching criteria based on the total number of citations a study has immediately prior to when the replication was published (see Equation 4 in the text and corresponding discussion). Note that the difference between the Max values for the treated and controls can be greater than the $TotAbsDiff_{38}$ values in TABLE 6 if there is a difference in citations in the year the papers were published, since the TABLE 6 values do not include these citations.

TABLE 8:
Mean Difference in Citations between Treated and Controls by Years Relative to Publication of the Replication

			Matching Criteria	
		PCT = 0%	***PCT = 10%***	***PCT = 20%***
(1)	***t = -3***	***0.125*** [1.38] (0.168)	***0.345**** [1.68] (0.094)	***0.319**** [1.84] (0.065)
(2)	***t = -2***	***0.088****** [2.92] (0.004)	***0.150*** [1.58] (0.114)	***0.454****** [3.28] (0.001)
(3)	***t = -1***	***0.087****** [3.30] (0.001)	***0.170***** [2.10] (0.036)	***0.550****** [3.61] (0.000)
(4)	***t = 0***	***2.744****** [4.86] (0.000)	***3.564****** [5.72] (0.000)	***3.587****** [6.82] (0.000)
(5)	***t = 1***	***1.826***** [2.32] (0.020)	***2.890****** [3.68] (0.000)	***2.517****** [3.62] (0.000)
(6)	***t = 2***	***2.024****** [4.22] (0.000)	***3.296****** [4.07] (0.000)	***2.536****** [3.28] (0.001)
(7)	***t = 3***	***2.452****** [4.85] (0.000)	***5.316****** [4.39] (0.000)	***4.145****** [3.96] (0.000)
(8)	**Mean Total Cites (t = -1)**	2.9	4.2	8.2
(9)	**Median Total Cites (t = -1)**	2	2	4
(10)	**N/Controls**	7,044	7,552	11,202
(11)	**N/Treated**	74	103	142

NOTE: The table reports the results of estimating β_0 in Equation (6) for three different samples defined by (*K/PCT*) = (3-8/0%), (3-8,10%) and (3-8,20%), where *K* is defined as the difference in years between the publication of the replication and the original, and *PCT* adjusts the matching criteria based on the total number of citations a study has immediately prior to when the replication was published (see Equation 4 in the text and corresponding discussion). Separate regressions are estimated for each of seven years (*t=-3,-2,-1,0,1,2,3*), where years are measured relative to the year the respective replication study was published. The dependent variable measures the difference in citations for the given year between replicated studies and their matched, unreplicated control studies. Estimates in brackets are *t*-values. Estimates in parentheses are *p*-values. *t*-values are based on cluster robust standard errors, where clusters are defined by "issue". An "issue" consists of all the control studies that are matched to a given treated study.

Estimates should be interpreted as the mean difference in citations at time *t* between studies that were replicated and matched control studies that were not replicated. To facilitate an assessment of the size of the estimated effects, Rows (8) and (9) report the mean and median total cites of the studies at time *t = 0.*

*, **, and *** indicate statistical significance at the 10-, 5-, and 1-percent levels.

TABLE 9:
Estimated Effect of Negative Replication on Citations of the Treated: 3-Year Post-Replication Period

		Matching Criteria		
		PCT = 0%	*PCT = 10%*	*PCT = 20%*
(1)	***t = -3***	***0.076*** [0.42] (0.674)	***0.438*** [1.15] (0.249)	***0.353*** [1.07] (0.285)
(2)	***t = -2***	***-0.013*** [-0.21] (0.833)	***0.145*** [0.75] (0.454)	***0.087*** [0.31] (0.755)
(3)	***t = -1***	***0.024*** [0.46] (0.649)	***-0.068*** [-0.39] (0.695)	***0.216*** [0.70] (0.483)
(4)	***t = 0***	***2.301***** [2.21] (0.027)	***1.813*** [1.50] (0.133)	***1.456*** [1.40] (0.162)
(5)	***t = 1***	***2.324*** [1.61] (0.107)	***0.394*** [0.26] (0.795)	***0.526*** [0.38] (0.706)
(6)	***t = 2***	***1.079*** [1.13] (0.261)	***-0.224*** [-0.14] (0.888)	***-0.333*** [-0.21] (0.830)
(7)	***t = 3***	***0.595*** [0.60] (0.549)	***-1.628*** [-0.66] (0.506)	***-2.324*** [-1.10] (0.273)

(8)	**Test of overall pre-replication effect:**	$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.11}$
		t = 0.52
		p = 0.605
			$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.51}$
			t = 0.99
			p = 0.320
				$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.70}$
				t = 1.41
				p = 0.159

(9)	**Test of overall post-replication effect:**	$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{4.01}$
		t = 1.37
		p = 0.170
			$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{-1.48}$
			t = -0.29
			p = 0.775
				$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{-2.11}$
				t = -0.44
				p = 0.658

(10)	**Mean Total Cites (t=0)**	2.9	4.2	8.2
(11)	**Median Total Cites (t=0)**	2	2	4
(12)	**N/Controls**	7,044	7,552	11,202
(13)	**Treated**	74	103	142

NOTE: The table reports the results of estimating β_1 in Equation (7) for three different samples defined by (*K/PCT*) = (3-8/0%), (3-8,10%) and (3-8,20%), where *K* is defined as the difference in years between the publication of the replication and the original, and *PCT* adjusts the matching criteria based on the total number of citations a study has immediately prior to when the replication was published (see Equation 4 in the text and corresponding discussion). Separate regressions are estimated for each of seven years (*t=-3,-2,-1,0,1,2,3*), where years are measured relative to the year the respective replication study was published.

The dependent variable measures the difference in citations for the given year between replicated studies and their matched, unreplicated control studies. Estimates in brackets are *t*-values. Estimates in parentheses are *p*-values. *t*-values are based on cluster robust standard errors, where clusters are defined by "issue". An "issue" consists of all the control studies that are matched to a given treated study.

Estimates should be interpreted as the mean difference in citations at time *t* between studies that were replicated and received "negative" assessments, and studies that were replicated and received "positive" or "mixed" assessments. Rows (8) and (9) report the results of combining observations from years *t=-3,-2,-1,0,1,2,3* and then estimating the joint hypotheses that the effects $\beta_1 = 0$ in each year of the "pre-" and "post-"replication periods, respectively. To facilitate an assessment of the size of the estimated effects, Rows (10) and (11) show the mean and median total cites of the studies at time *t = 0.*

*, **, and *** indicate statistical significance at the 10-, 5-, and 1-percent levels.

TABLE 10:
Estimated Effect of Negative Replication on Citations of the Treated: 5-Year Post-Replication Period

			Matching Criteria	
		PCT = 0%	***PCT = 10%***	***PCT = 20%***
(1)	***t = -3***	***0.047*** [0.31] (0.754)	***0.438*** [0.92] (0.356)	***0.316*** [0.78] (0.436)
(2)	***t = -2***	***-0.021*** [-0.27] (0.787)	***0.227*** [0.88] (0.381)	***0.238*** [0.69] (0.489)
(3)	***t = -1***	***0.011*** [0.18] (0.854)	***-0.004*** [-0.02] (0.986)	***0.347*** [0.89] (0.373)
(4)	***t = 0***	***2.044*** [2.11] (0.035)	***2.019*** [1.43] (0.153)	***1.688*** [1.38] (0.169)
(5)	***t = 1***	***2.383*** [1.34] (0.181)	***0.627*** [0.33] (0.744)	***0.731*** [0.43] (0.669)
(6)	***t = 2***	***0.843*** [0.74] (0.458)	***0.174*** [0.09] (0.930)	***-0.300*** [-0.16] (0.875)
(7)	***t = 3***	***1.059*** [1.32] (0.185)	***-0.683*** [-0.22] (0.824)	***-1.626*** [-0.62] (0.534)
(8)	***t = 4***	***1.254*** [0.88] (0.381)	***-0.368*** [-0.11] (0.911)	***-0.456*** [-0.16] (0.872)
(9)	***t = 5***	***3.059*** [1.42] (0.155)	***-0.520*** [-0.10] (0.922)	***-0.033*** [-0.01] (0.994)

(10)	**Test of overall pre-replication effect:**	$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.15}$
		t = 0.88
		p = 0.377
			$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.69}$
			t = 1.02
			p = 0.309
				$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.93}$
				t = 1.50
				p = 0.134

(11)	**Test of overall post-replication effect:**	$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{8.65}$
		t = 1.33
		p = 0.182
			$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{-0.75}$
			t = -0.05
			p = 0.959
				$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{-1.66}$
				t = -0.13
				p = 0.895

		Matching Criteria		
		PCT = 0%	*PCT = 10%*	*PCT = 20%*
(12)	**Mean Total Cites (t=0)**	4.0	5.9	11.5
(13)	**Median Total Cites (t=0)**	2	3	5
(14)	**N/Controls**	6,171	6,587	8,689
(15)	**Treated**	55	79	112

NOTE: This table reports the same information as TABLE 9, except that it restricts the sample to studies that have 5 years of post-replication data (compared to 3 years of post-replication data in TABLE 9).

*, **, and *** indicate statistical significance at the 10-, 5-, and 1-percent levels.

TABLE 11:
Mean Difference in Citations between Treated and Controls by Years Relative to Publication of the Replication: Quantile Regression

		Matching Criteria		
		PCT = 0%	*PCT = 10%*	*PCT = 20%*
(1)	*t = -3*	0.000 [0.00] (1.000)	0.000 [0.00] (1.000)	0.000 [0.00] (1.000)
(2)	*t = -2*	0.000 [0.00] (1.000)	0.000 [0.00] (1.000)	0.139* [1.83] (0.069)
(3)	*t = -1*	0.000 [0.00] (1.000)	0.000 [0.00] (1.000)	0.204** [2.40] (0.018)
(4)	*t = 0*	1.558*** [3.22] (0.002)	1.930*** [3.28] (0.001)	1.933*** [4.40] (0.000)
(5)	*t = 1*	0.500 [1.46] (0.149)	0.833* [1.89] (0.062)	0.889** [2.24] (0.026)
(6)	*t = 2*	0.750 [1.57] (0.120)	1.295** [2.64] (0.010)	1.061** [2.46] (0.015)
(7)	*t = 3*	1.717*** [5.24] (0.000)	2.000*** [5.77] (0.000)	1.898*** [4.61] (0.000)
(8)	**Mean Total Cites (t = -1)**	7.8	19.5	27.8
(9)	**Median Total Cites (t = -1)**	4.7	8.4	14.8
(11)	**Observations**	74	103	142

NOTE: The estimates in the table come from the same general procedure described in TABLE 8 with two main differences. First, the individual observations associated with each treated study have been collapsed to a single observation. $DIFF_{it,i\in K}$ is now the mean value of the difference in citations for the given year between a replicated study and its matched, unreplicated control studies. Second, we use quantile regression to estimate Equation (6) with bootstrapped standard errors (1000 replications). Accordingly, the estimates should be interpreted as the median, mean value of $DIFF_{it,i\in K}$.. Rows (8) and (9) report the mean and median values of the treated-specific, average total cites of the studies at time *t = 0.*

*, **, and *** indicate statistical significance at the 10-, 5-, and 1-percent levels.

TABLE 12:
Estimated Effect of Negative Replication on Citations of the Treated: Quantile Regression

		Matching Criteria		
		PCT = 0%	*PCT = 10%*	*PCT = 20%*
(1)	*t = -3*	***0.000*** [0.00] (1.000)	***0.000*** [0.00] (1.000)	***-0.037*** [-0.26] (0.798)
(2)	*t = -2*	***0.000*** [0.00] (1.000)	***0.000*** [0.00] (1.000)	***-0.222*** [-1.66] (0.100)
(3)	*t = -1*	***0.037*** [0.52] (0.602)	***0.105*** [1.61] (0.111)	***0.204*** [1.25] (0.213)
(4)	*t = 0*	***0.980*** [1.10] (0.275)	***1.053*** [0.90] (0.371)	***1.538**** [1.87] (0.063)
(5)	*t = 1*	***0.045*** [0.05] (0.958)	***-0.406*** [-0.39] (0.700)	***0.257*** [0.34] (0.734)
(6)	*t = 2*	***1.667***** [2.44] (0.017)	***1.217*** [1.00] (0.320)	***0.200*** [0.18] (0.860)
(7)	*t = 3*	***0.777*** [1.18] (0.243)	***0.457*** [0.48] (0.630)	***0.265*** [0.25] (0.804)

(8)	**Test of overall pre-replication effect:**	$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.04}$
		t = 0.32
		p = 0.746
			$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.11}$
			t = 0.61
			p = 0.540
				$\sum_{t=-3}^{-1} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{-0.06}$
				t = -0.22
				p = 0.825

(9)	**Test of overall post-replication effect:**	$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{2.49}^{*}$
		t = 1.92
		p = 0.055
			$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{1.27}$
			t = 0.69
			p = 0.488
				$\sum_{t=1}^{5} \hat{\boldsymbol{\beta}}_{1t} = \mathbf{0.72}$
				t = 0.43
				p = 0.666

(10)	**Mean Total Cites (t=0)**	7.8	19.5	27.8
(11)	**Median Total Cites (t=0)**	4.7	8.4	14.8
(12)	**Observations**	74	103	142

NOTE: The estimates in the table come from the same general procedure described in TABLE 9 with two main differences. First, the individual observations associated with each treated study have been collapsed to a single observation. $DIFF_{it,i\in K}$ is now the mean value of the difference in citations for the given year between a replicated study and its matched, unreplicated control studies. Second, we use quantile regression to estimate Equation (7) with bootstrapped standard errors (1000 replications). Accordingly, the estimates should be interpreted as the estimated effect that a negative replication has on the median, mean value of $DIFF_{it,i\in K}$ relative to a positive or mixed replication. Rows (10) and (11) report the mean and median values of the treated-specific, average total cites of the studies at time *t = 0.*

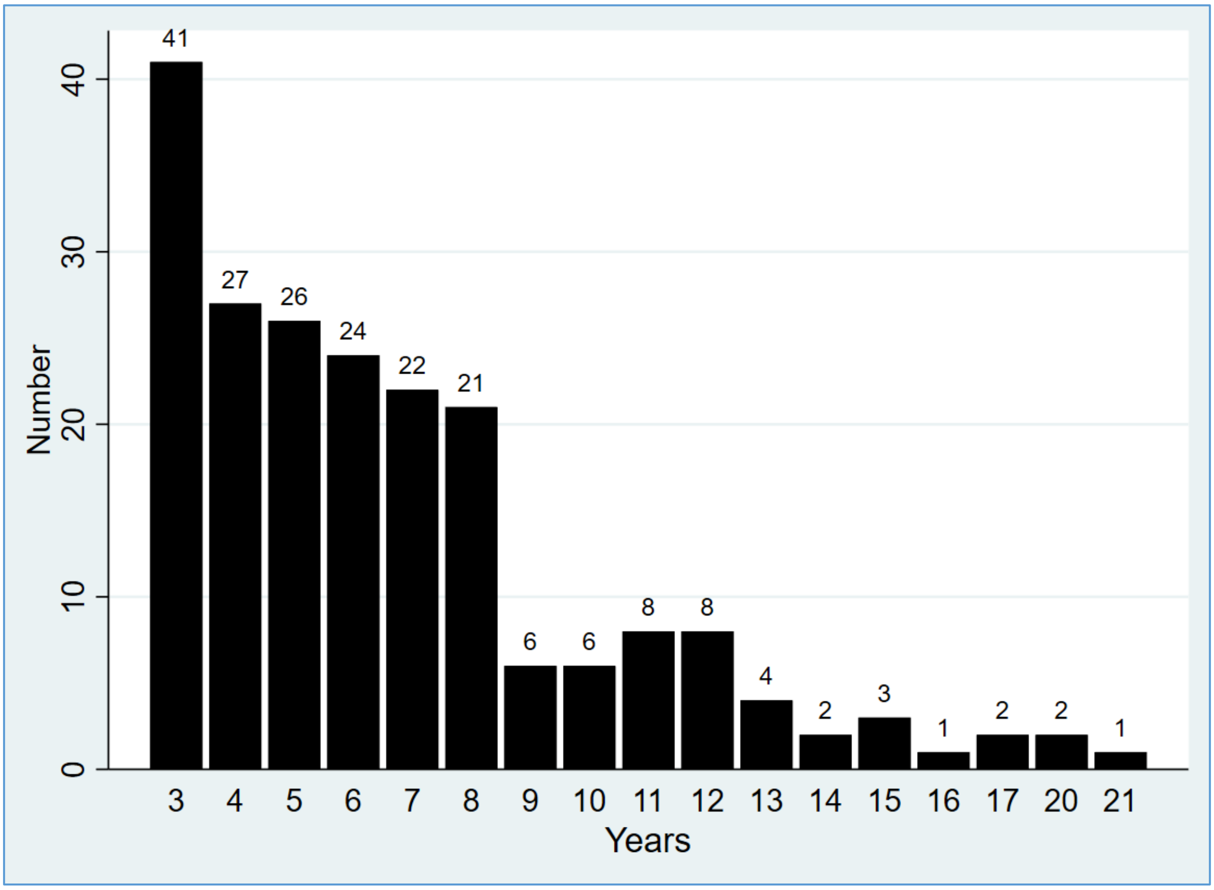
*, **, and *** indicate statistical significance at the 10-, 5-, and 1-percent levels.

TABLE 13
Ex-post Estimates of Statistical Power

Time	*Effect Size = 1 citation/year; 3 citations/3 years*		*Effect Size = 2 citation/year; 6 citations/3 years*	
	PCT = 0%	***PCT = 10%***	***PCT = 0%***	***PCT = 10%***
t = 1	21%	16%	64%	47%
t = 2	31%	13%	83%	37%
t = 3	33%	18%	86%	56%
t = (1,2,3)	64%	37%	100%	91%

NOTE: Estimates of statistical power are calculated by $Prob\left(\left(z < \frac{Effect\ Size}{Estimated\ SE}\right) - 1.9645\right)$, where z is distributed standard normal, Effect Size is either (1 citation/year; 3 citations/3 years) or (2 citations/year; 6 citations/3 years), and Estimated SE comes from the standard errors associated with the estimated treatment effects in TABLE 12.

FIGURE 1:
Years between Publication of Treated and Its Replication



NOTE: The table displays number of studies by the difference in years between when an original study was published and when its replication was published ("Years") for the full sample of 204 treated studies. Note that a study with 3 years difference -- say the original was published in 2014 and the replication was published in 2017 -- has two full years of intervening data (2015, 2016) by which to match citation histories. In our sample, most of the treated studies have 8 or fewer years' difference between when they were published and when their replications were published.

FIGURE 2:
Matching Controls with Treated Based on Citation History

A. Three-year gap (K=3) between publication of original and replication

	Year Original Published (T = -3)	Citations (T = -2)	Citations (T = -1)	Year Replication Published (T = 0)
Control		$Citations_{T-2}^{Control}$	$Citations_{T-1}^{Control}$	
Treated		$Citations_{T-2}^{Treated}$	$Citations_{T-1}^{Treated}$	

$$TotAbsDiff_3 \;=\; \left|Citations_{T-2}^{Control} - Citations_{T-2}^{Treated}\right| + \left|Citations_{T-1}^{Control} - Citations_{T-1}^{Treated}\right|$$

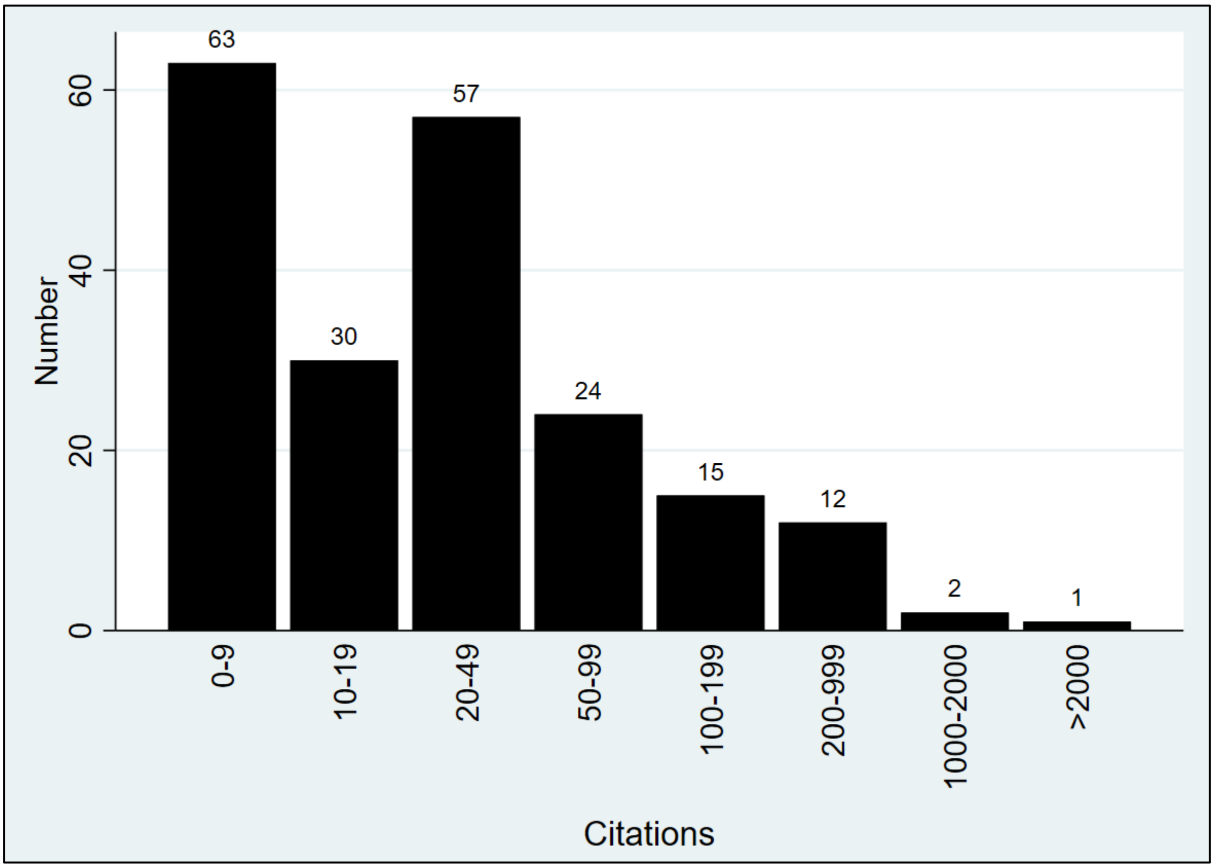
B. Four-year gap (K=4) between publication of original and replication

	Year Original Published (T = -4)	Citations (T = -3)	Citations (T = -2)	Citations (T = -1)	Year Replication Published (T = 0)
Control		$Citations_{T-3}^{Control}$	$Citations_{T-2}^{Control}$	$Citations_{T-1}^{Control}$	
Treated		$Citations_{T-3}^{Treated}$	$Citations_{T-2}^{Treated}$	$Citations_{T-1}^{Treated}$	

$$TotAbsDiff_4 \;=\; \left|Citations_{T-3}^{Control} - Citations_{T-3}^{Treated}\right| + \left|Citations_{T-2}^{Control} - Citations_{T-2}^{Treated}\right| + \left|Citations_{T-1}^{Control} - Citations_{T-1}^{Treated}\right|$$

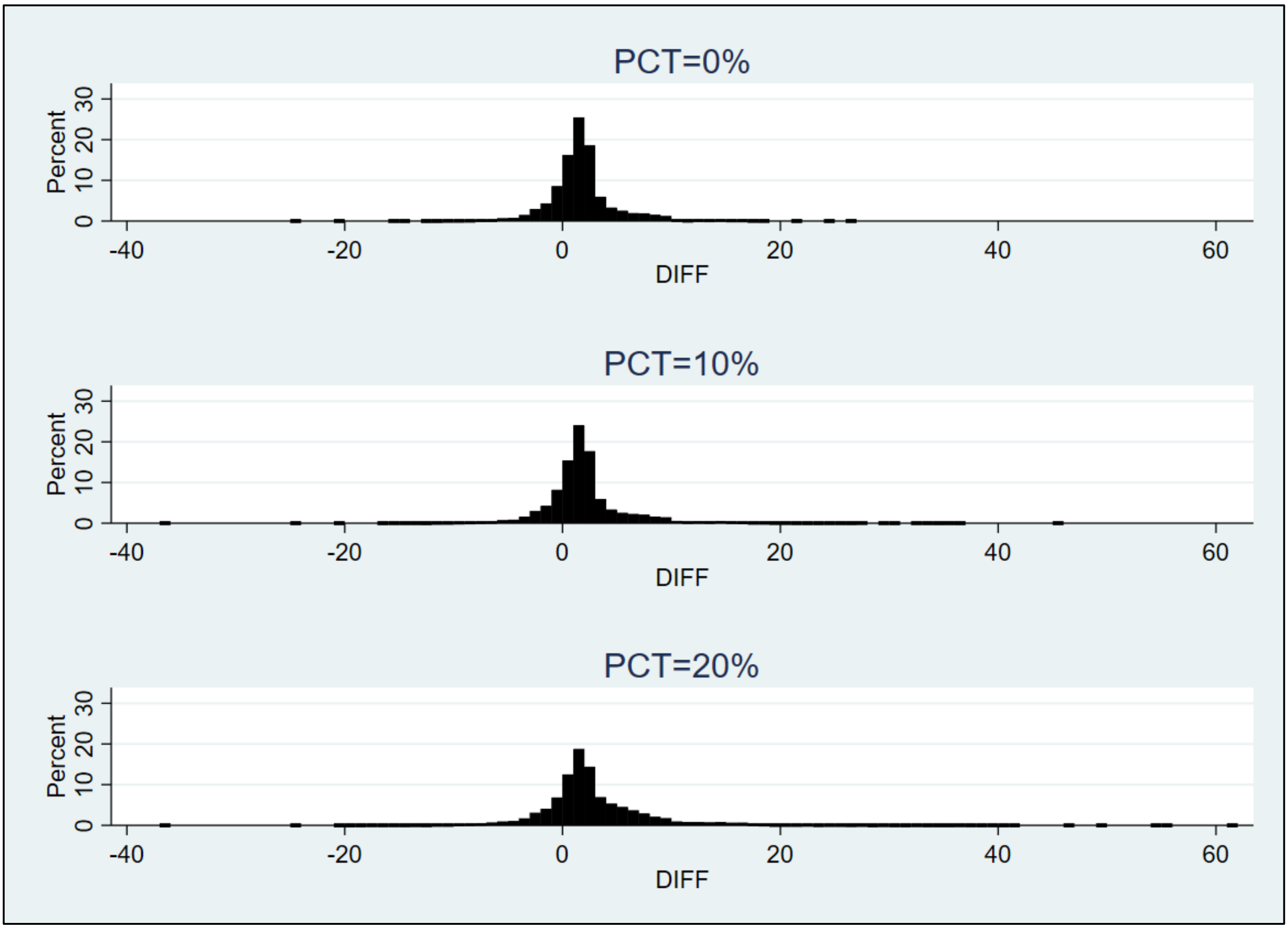
NOTE: This figure illustrates the relationship between years difference between when an original and its replication were published, and number of intervening years available to compare citation histories.

FIGURE 3:
Total number of citations of treated studies at time replication was published



NOTE: The figure shows the distribution of total citations for the full sample of 204 treated studies up to the time immediately before their replications were published. Note that the subsamples used in our analyses are disproportionately drawn from the lower end of the citation distribution.

FIGURE 4:
Representative Histograms for the Variable DIFF



NOTE: Each of the panels above show the distribution of the dependent variable in Equations (6) and (7) at time *t=0* for the three analysis samples defined by (*K/PCT*) = (3-8/0%), (3-8,10%) and (3-8,20%), where *K* is defined as the difference in years between the publication of the replication and the original, and *PCT* adjusts the matching criteria based on the total number of citations a study has immediately prior to when the replication was published (see Equation 4 in the text and corresponding discussion).

APPENDIX A:
Plots of Effects of Negative Replications on Citations

Panel A: OLS-Pooled (3-Year Post-Estimation Period)

Panel B: OLS-Between (3-Year Post-Estimation Period)

Panel C: Random Effects (3-Year Post-Estimation Period)

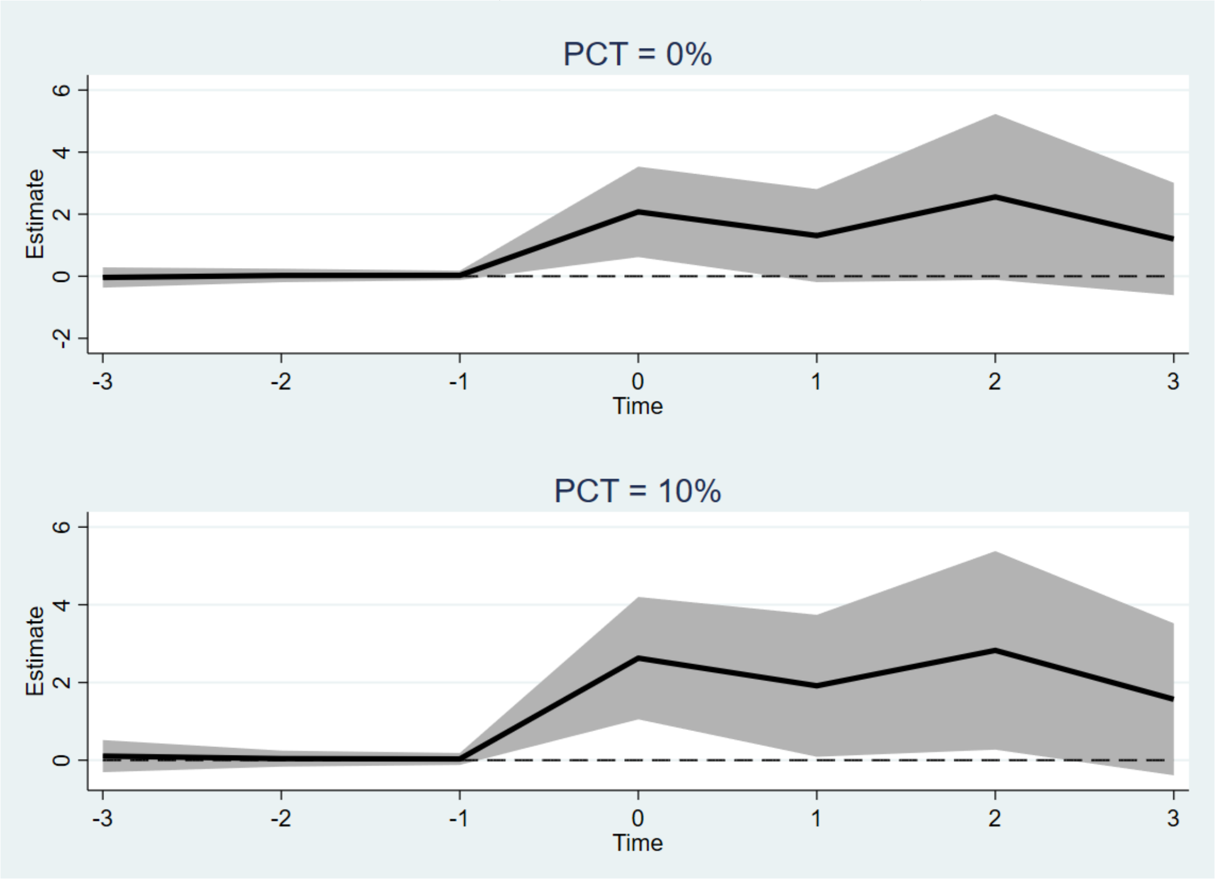
Panel D: OLS-Pooled (5-Year Post-Estimation Period)

Panel E: OLS-Between (5-Year Post-Estimation Period)

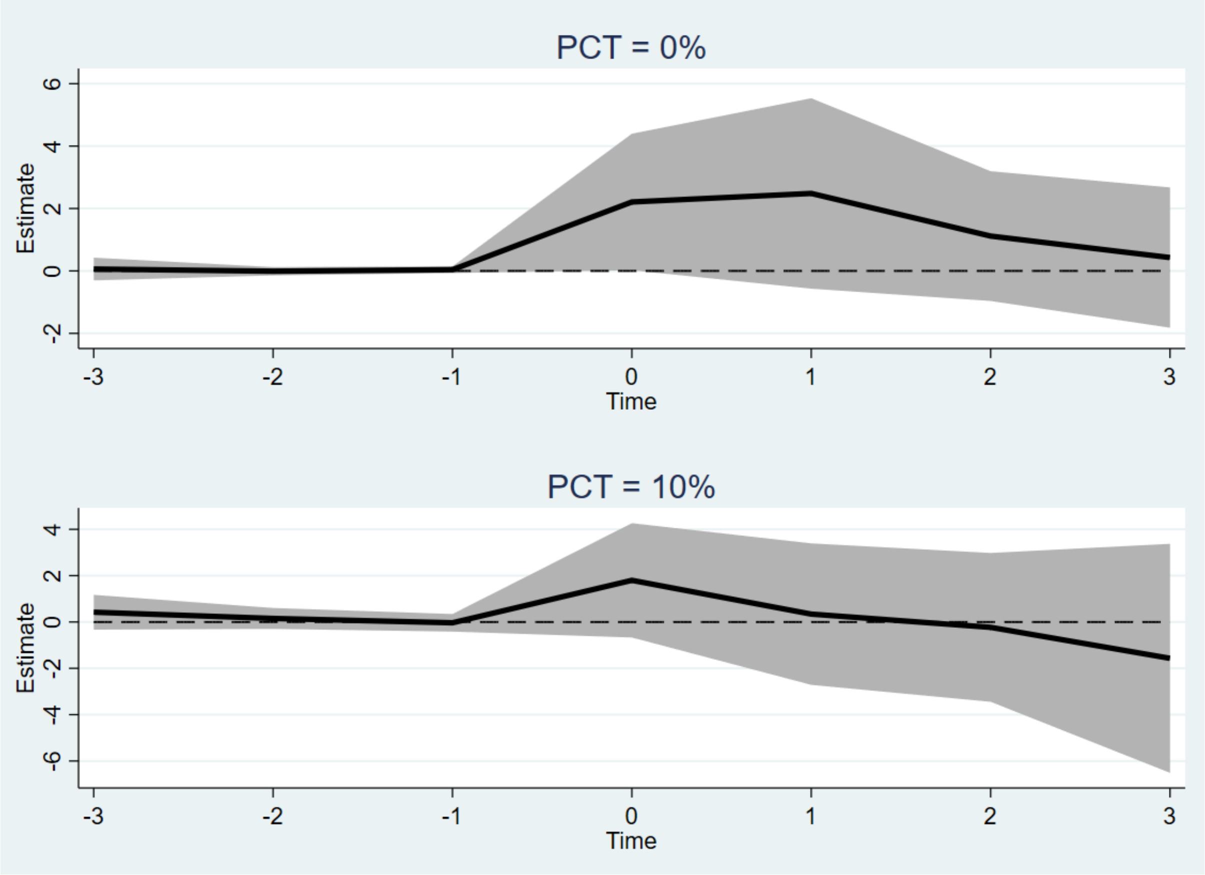
Panel F: Random Effects (5-Year Post-Estimation Period)

Panel G: Quantile Regression (5-Year Post-Estimation Period)

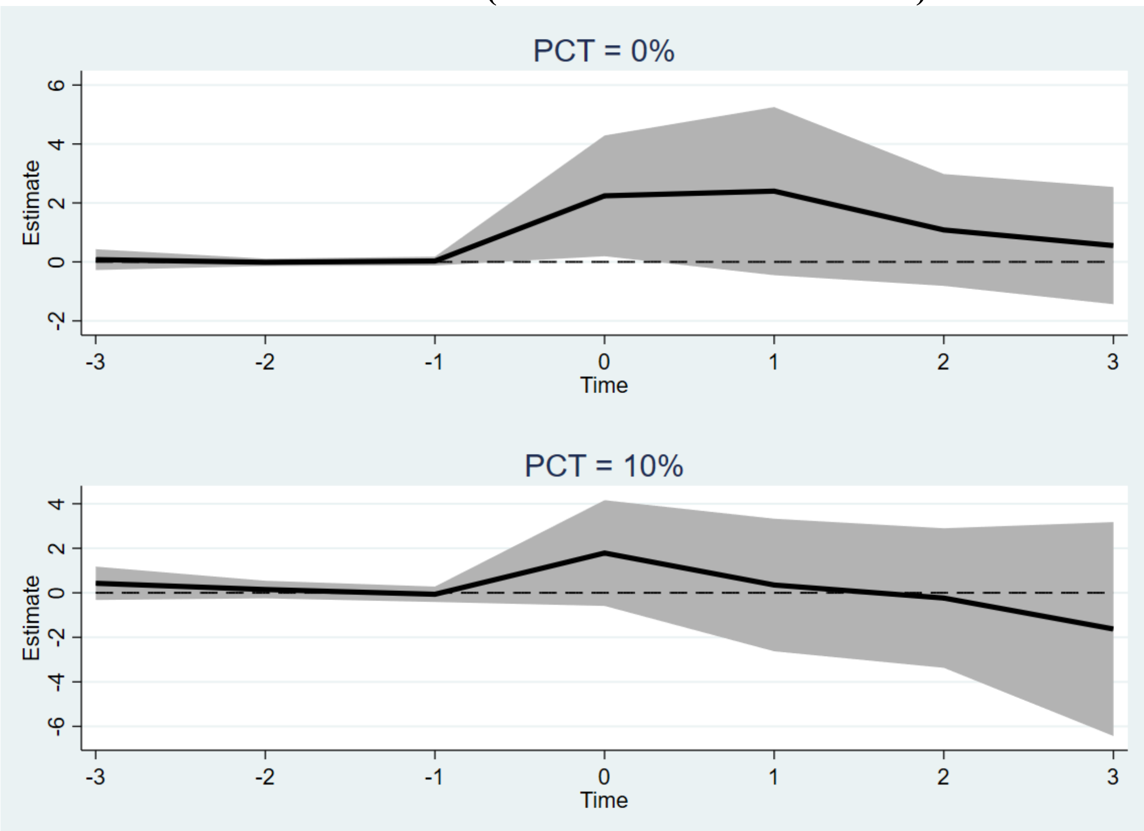
A. OLS-Pooled (3-Year Post-Estimation Period)



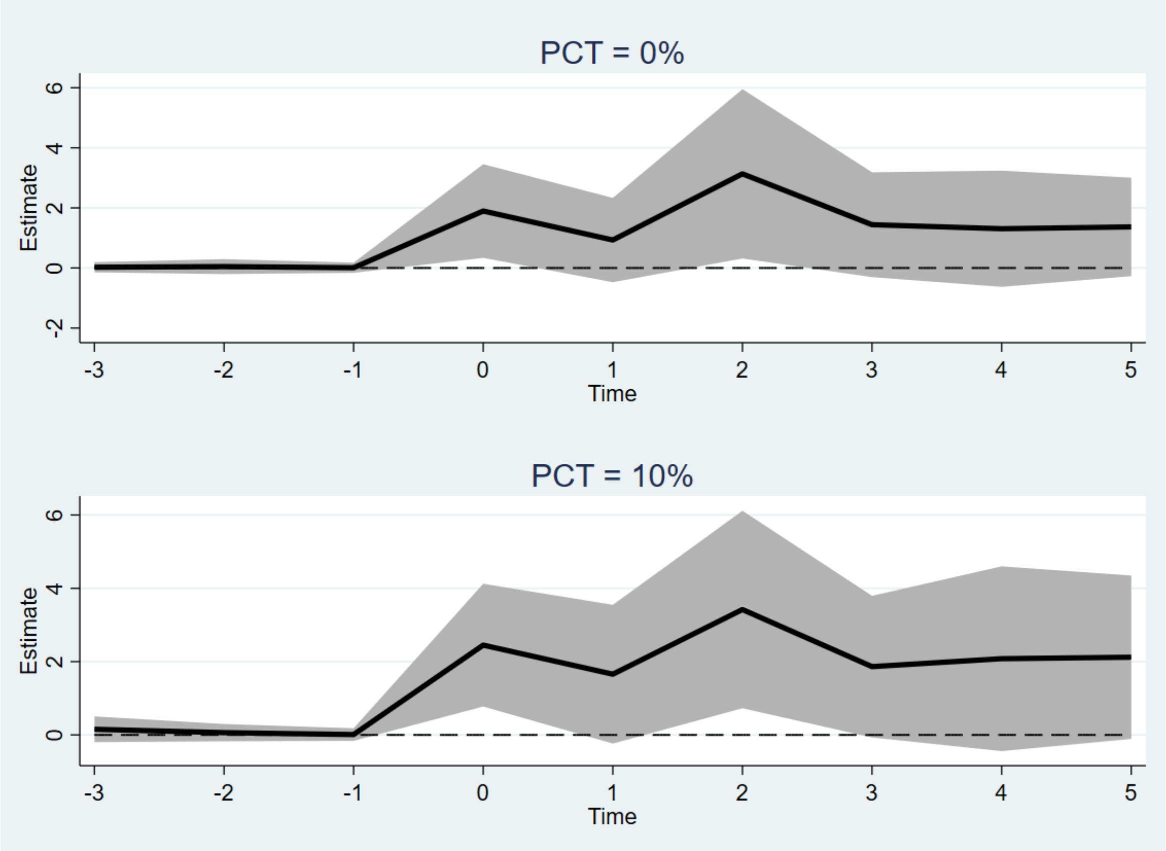
B. OLS-Between (3-Year Post-Estimation Period)



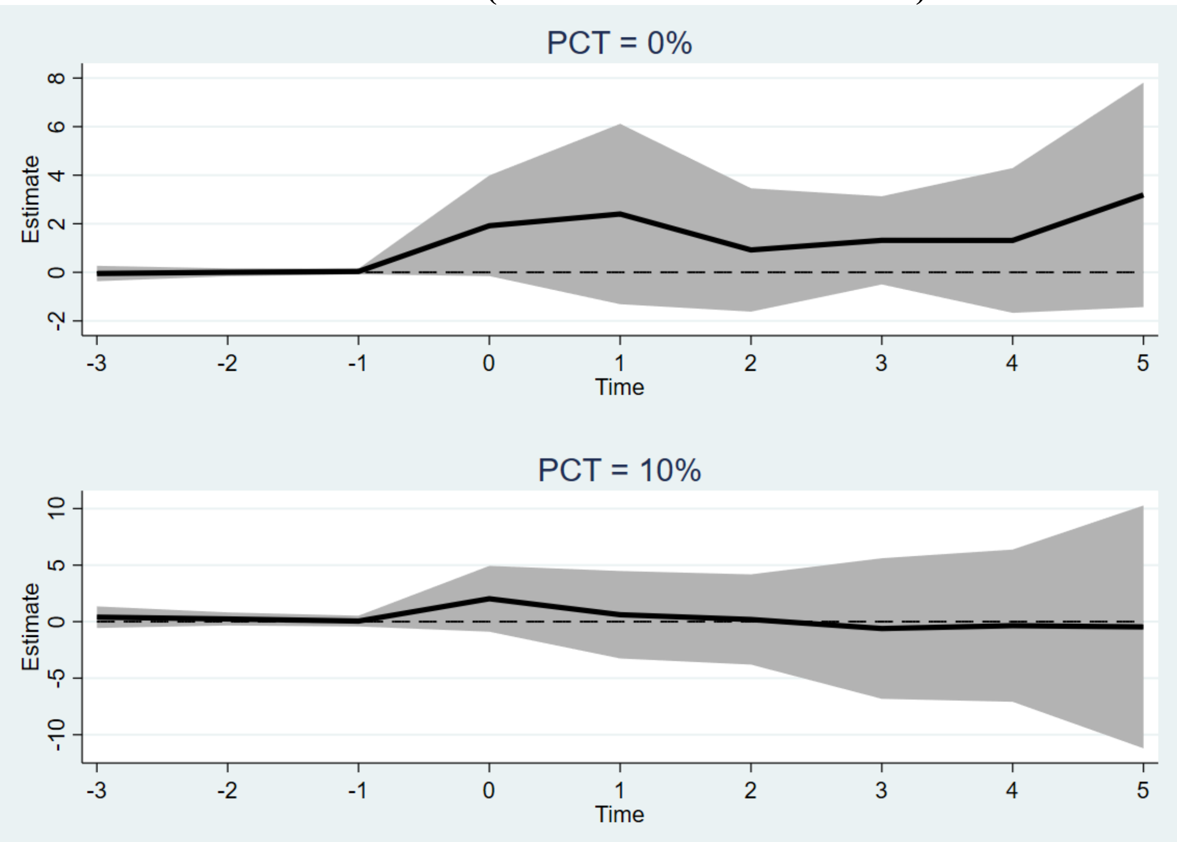
C. Random Effects (3-Year Post-Estimation Period)



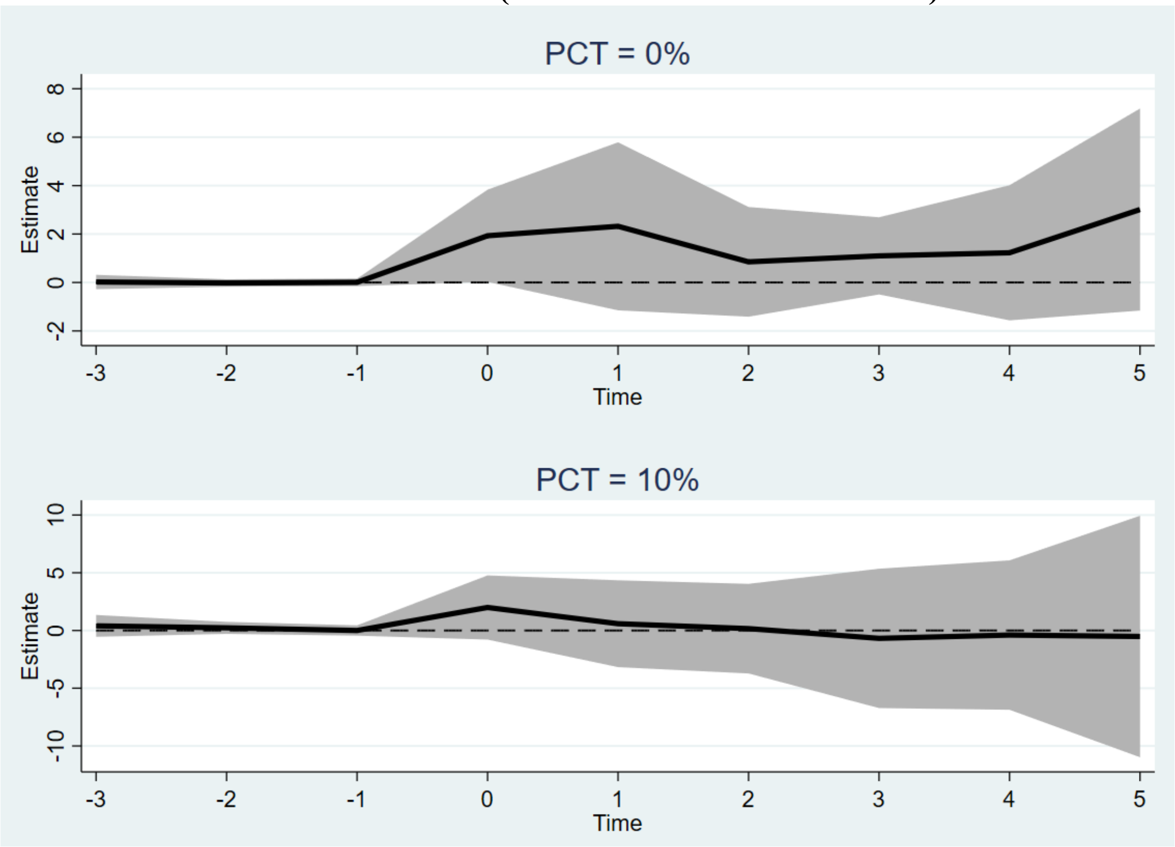
D. OLS-Pooled (5-Year Post-Estimation Period)



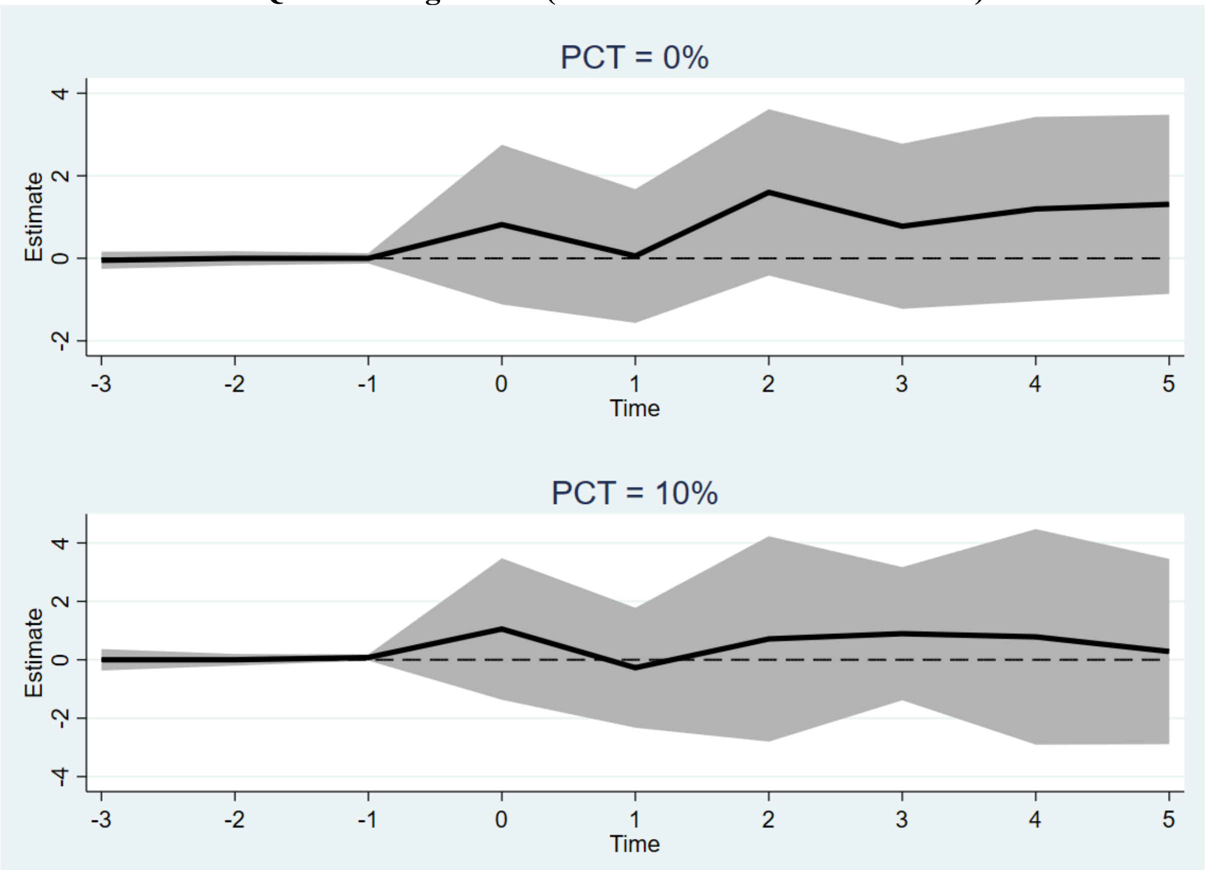
E. OLS-Between (5-Year Post-Estimation Period)



F. Random Effects (5-Year Post-Estimation Period)



G. Quantile Regression (5-Year Post-Estimation Period)



APPENDIX B:
Plots of Overall Effect of Replications on Citations

Panel A: OLS-Pooled (3-Year Post-Estimation Period)

Panel B: OLS-Between (3-Year Post-Estimation Period)

Panel C: Random Effects (3-Year Post-Estimation Period)

Panel D: OLS-Pooled (5-Year Post-Estimation Period)

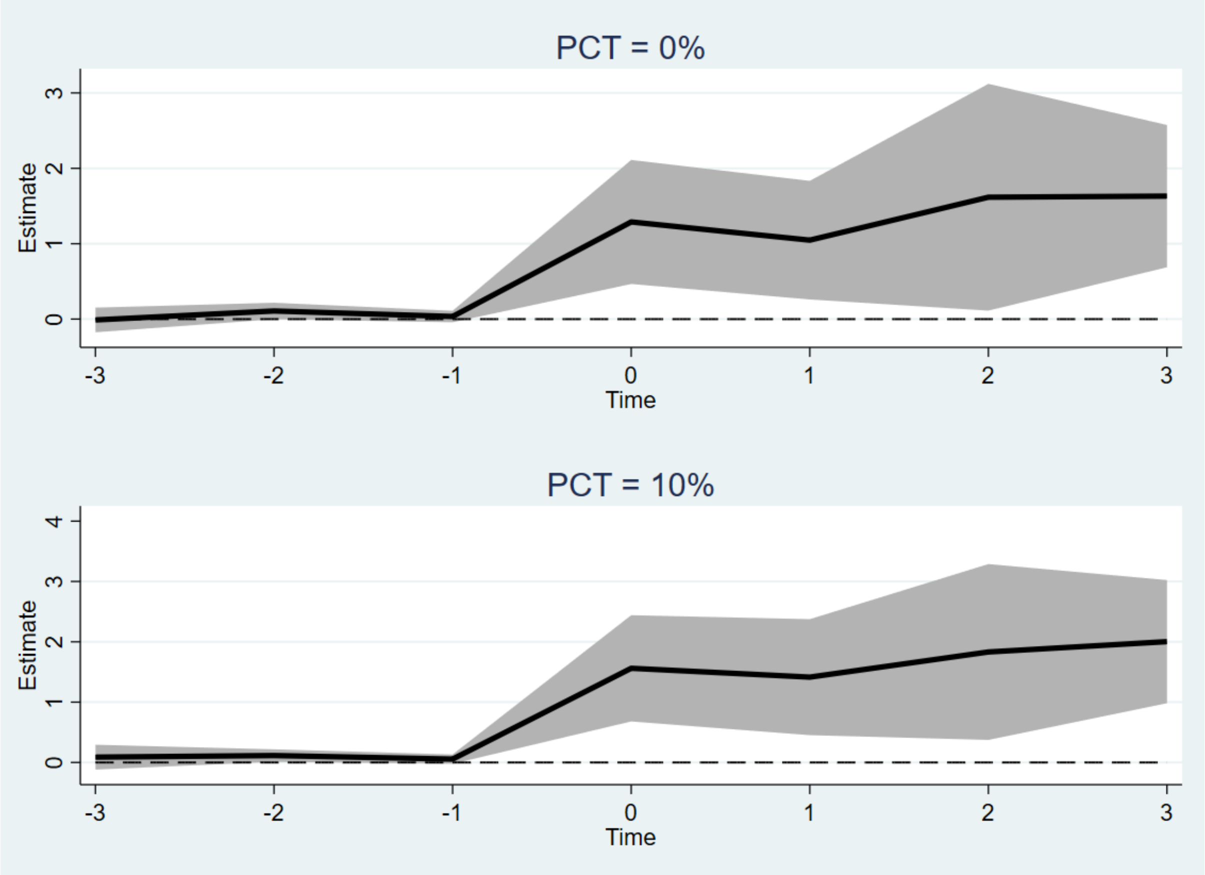
Panel E: OLS-Between (5-Year Post-Estimation Period)

Panel F: Random Effects (5-Year Post-Estimation Period)

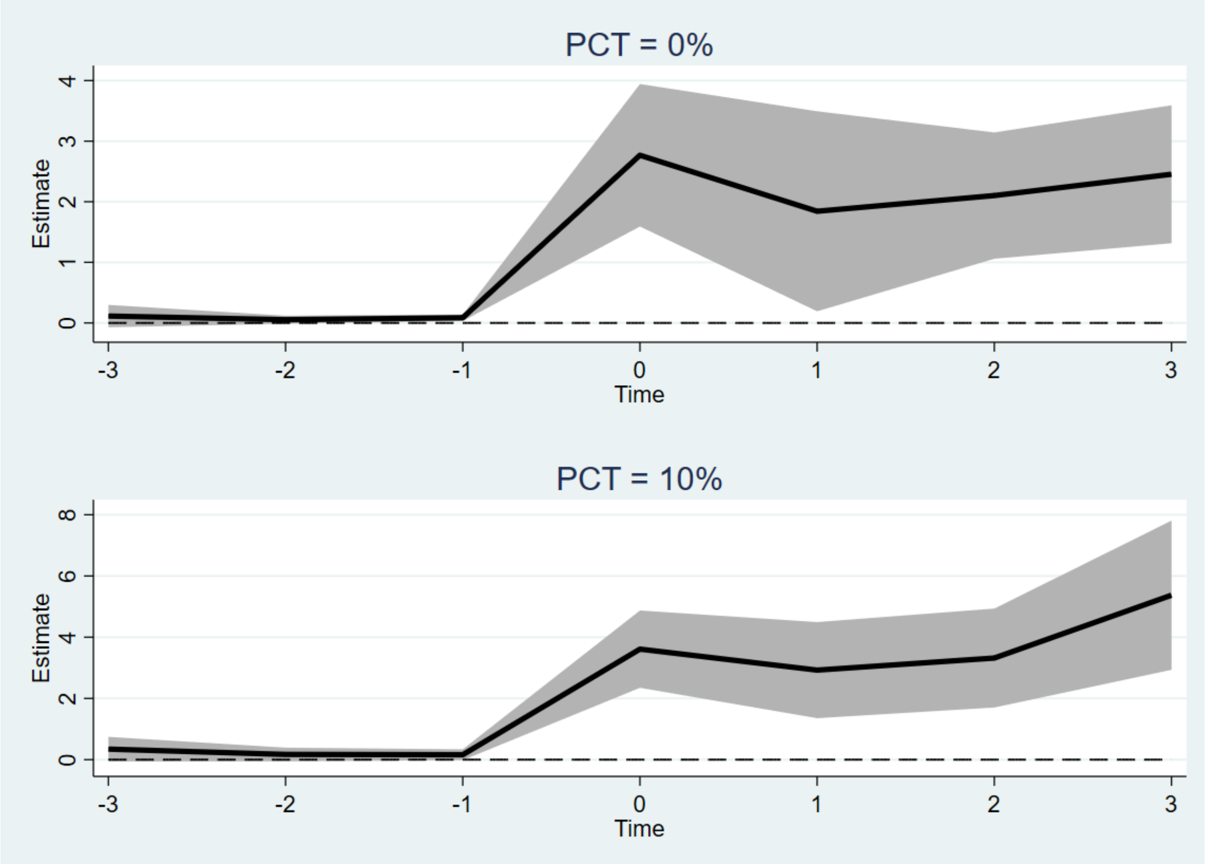
Panel G: Quantile Regression (5-Year Post-Estimation Period)

Panel H: HLM (5-Year Post-Estimation Period)

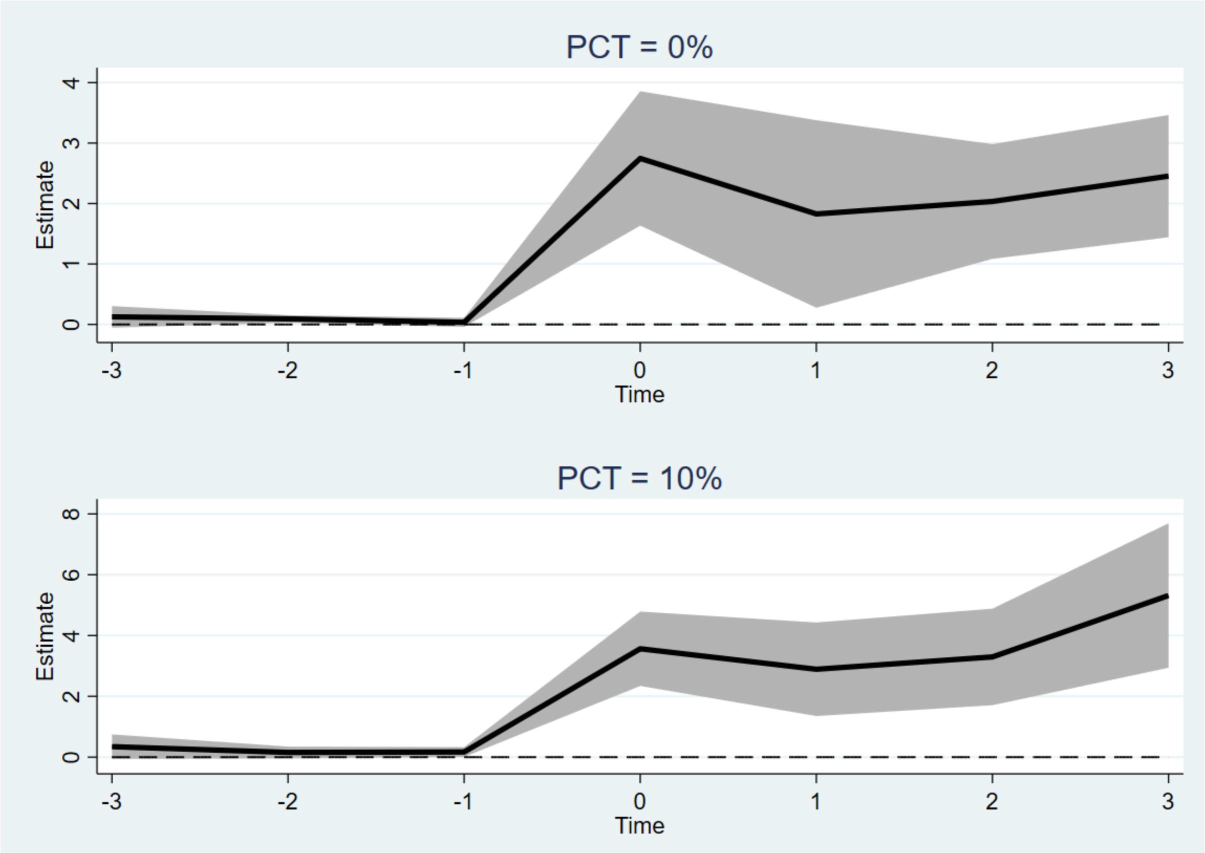
A. OLS-Pooled (3-Year Post-Estimation Period)



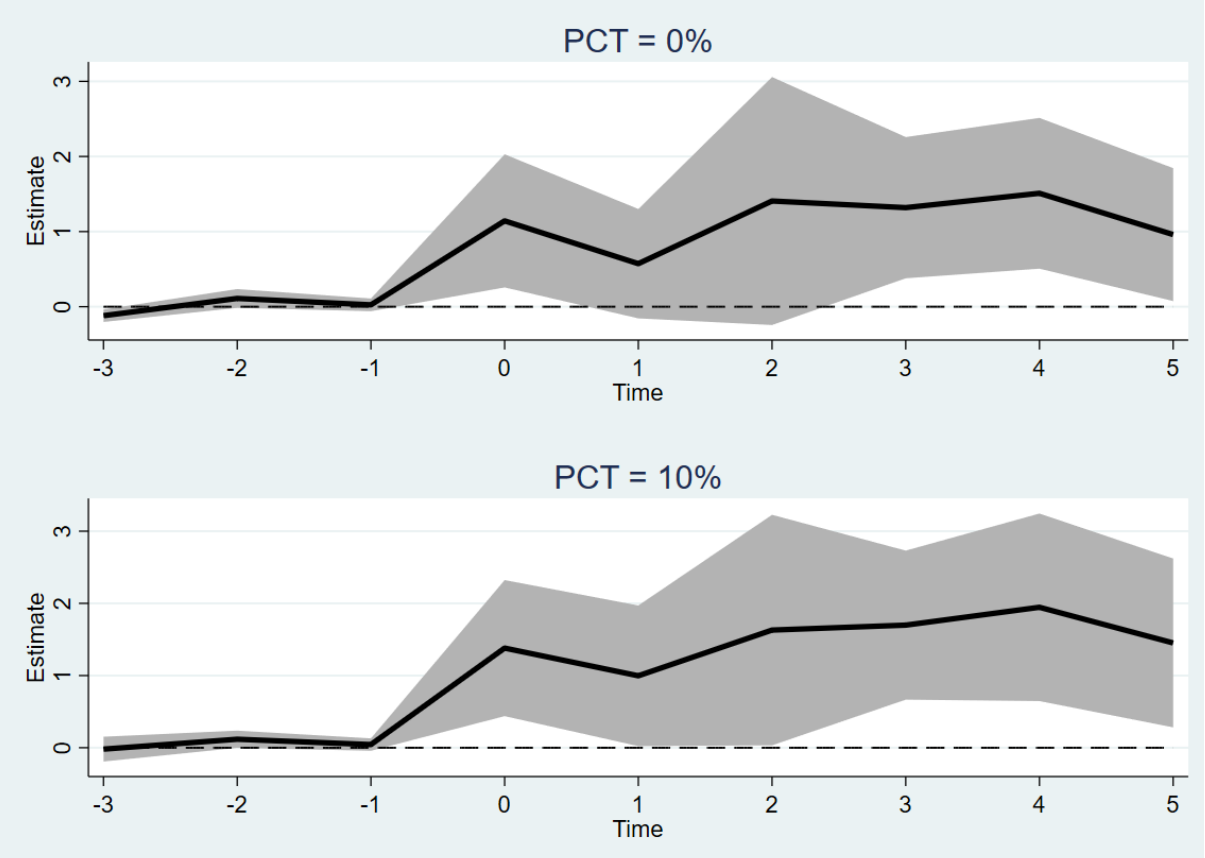
B. OLS-Between (3-Year Post-Estimation Period)



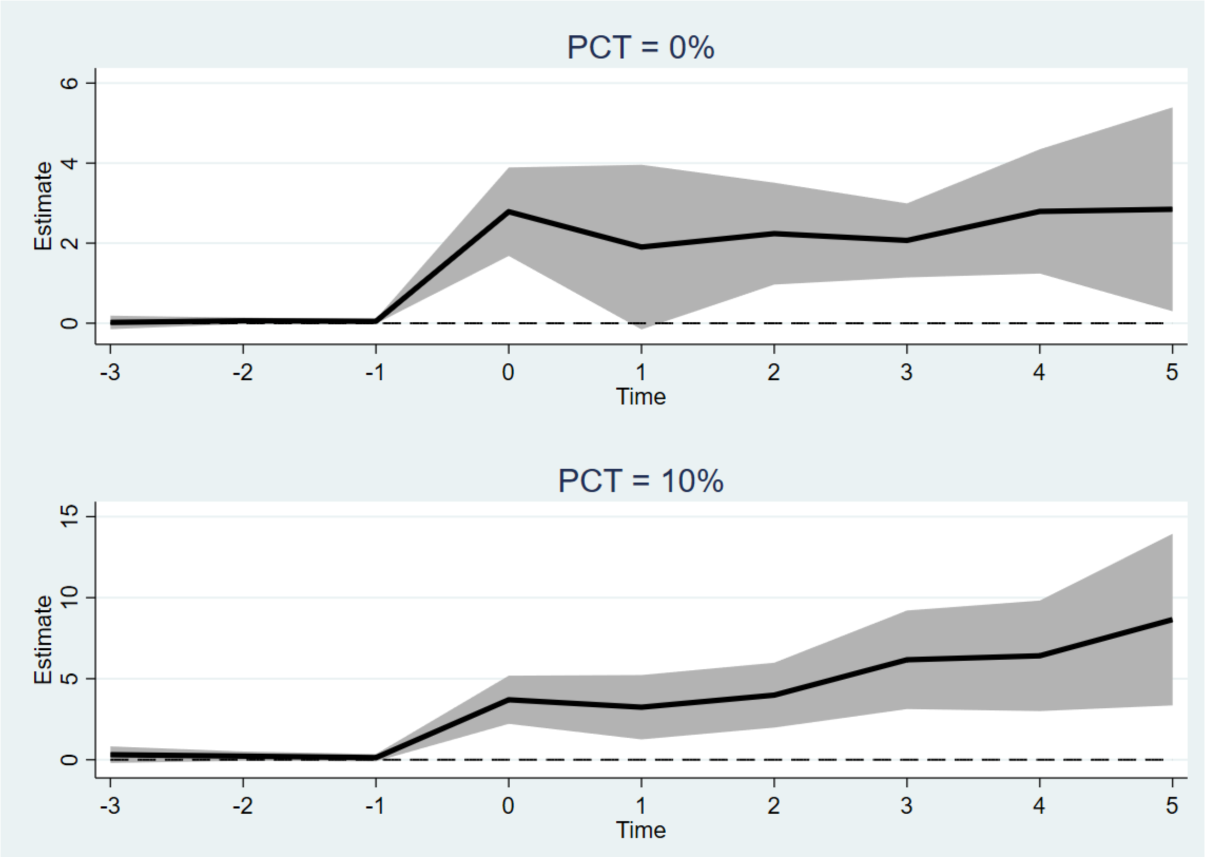
C. Random Effects (3-Year Post-Estimation Period)



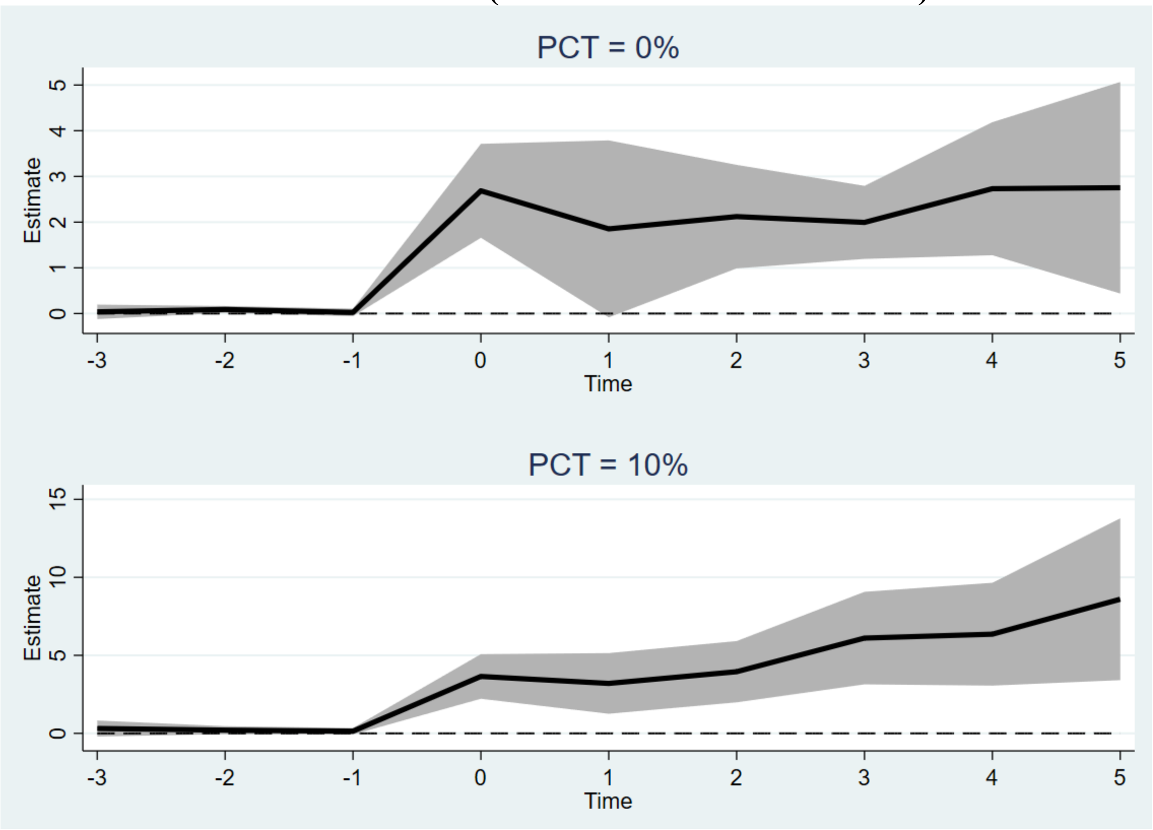
D. OLS-Pooled (5-Year Post-Estimation Period)



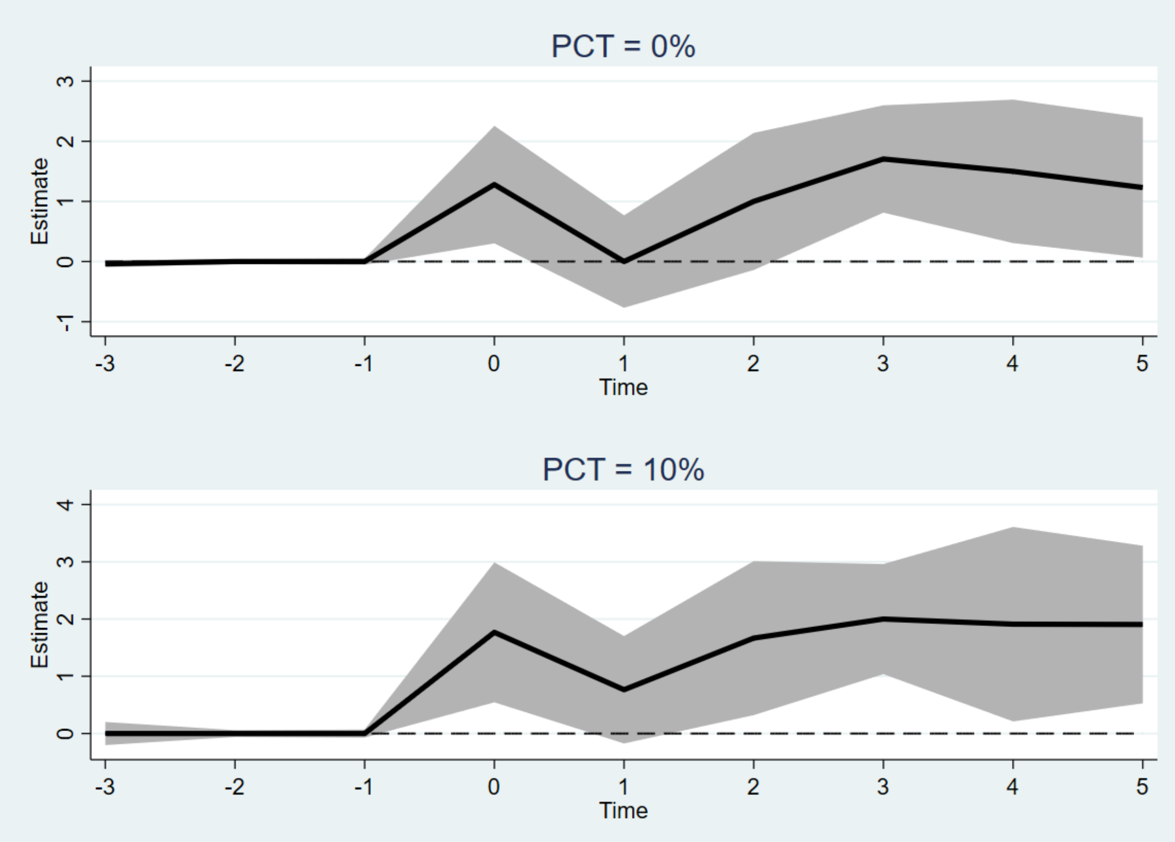
E. OLS-Between (5-Year Post-Estimation Period)



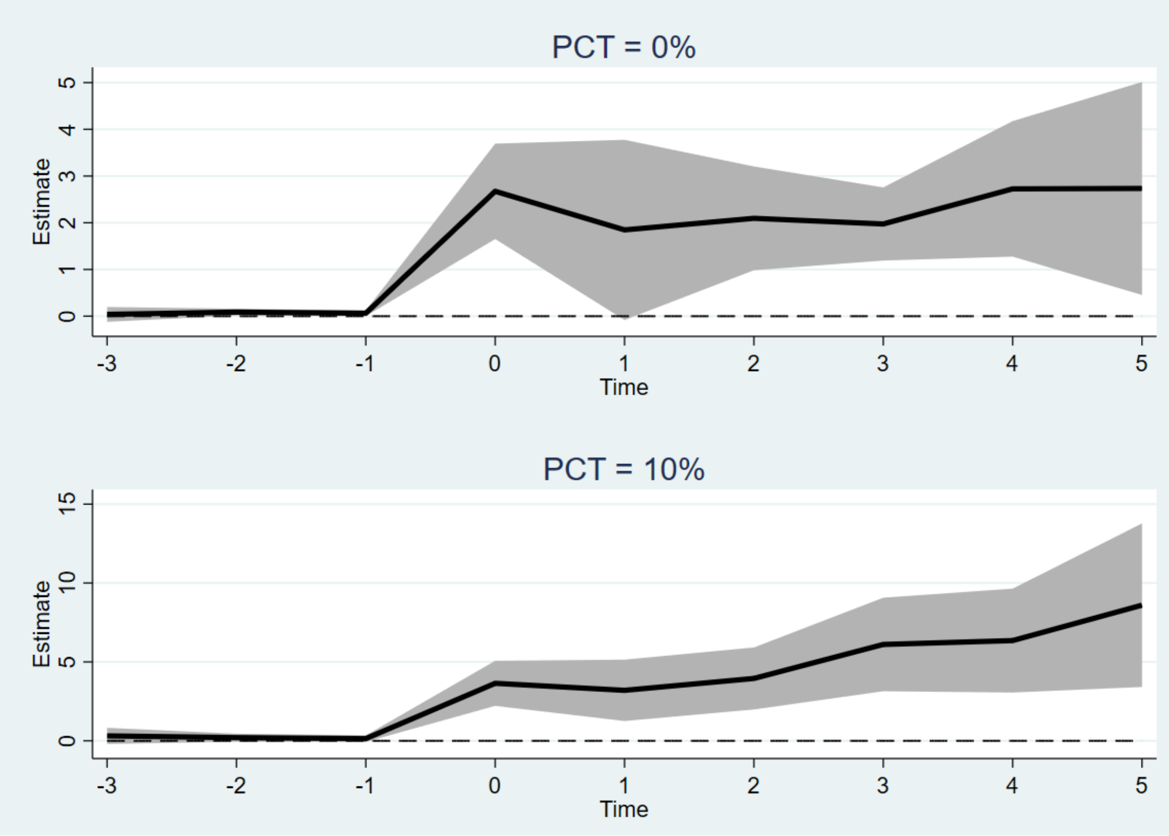
F. Random Effects (5-Year Post-Estimation Period)



G. Quantile Regression (5-Year Post-Estimation Period)



H. HLM (5-Year Post-Estimation Period)



SUPPLEMENT:
Categorization of Replications as Positive, Negative, or Mixed/Unclear

Replication	Positive	Negative	Mixed / Unclear	Key sentence (page number)
Anderson & Delgado (2010)	1	0	0	This paper describes a successful attempt to replicate D…D's published results are replicable…identically replicate the paper. (1/20)
Abascal & Baldassarri (2015)	0	1	0	Our evidence suggests there is no meaningful relationship between ethnic diversity and measures of trust and cooperation. (Page 748)
Abrevaya & Puzzello (2012)	0	1	0	In this comment, we have re-examined this claim by AC and find little systematic evidence to support it. (11/14)
Agell, Ohlsson & Thoursie (2006)	0	1	0	But in failing to control for…the regressions reported by FH are flawed. Using…we find that…is statistically insignificant and highly unstable across specifications. (1/8)
Aharoni, Grundy & Zeng (2013)	1	0	0	Our empirical work supports the Fama French valuation insight. (10/11)
Aiken et al. (2015)	0	0	1	Although most results were reproduced as originally reported, we identified discrepancies of several types between the original findings and re-analysis. (Abstract)
Albouy (2012)	0	1	0	This comment argues that there are several reasons to doubt the reliability and comparability of their European settler mortality rates and the conclusions that depend on them. (3/19). When the first-stage estimate of ß is not significantly different from zero (15/19) ...Cross-country growth regressions cannot disentangle the effect of settler mortality from that of other variables that may explain institutions and growth... (16/19)
Allen & Price (2015)	0	1	0	Their claim “that there can be no rational argument” is therefore wrong. (3/20) As other studies come forward and the inevitable ‘counting up’ of studies takes place, the Langbein and Yost paper should not be counted. It found nothing. (13/20)
Amati (2009)	1	0	0	All the Tables and results were replicated using Stata 10. (Page 1054)
Amin (2011)	0	0	1	This comment focuses on their primary conclusion, illustrating that whilst the theoretical predictions are robust, the point estimate of a return to education of 7.7 percent is driven by an outlier twin pair in the dataset. If one removes that twin pair, the estimated return to education in their paper's favored specification is 5.1 percent and only significant at the 10 percent level... (7/8)
Antonovics & Goldberger (2005)	0	1	0	We find issues in the construction and coding of their data that lead us to question their primary conclusion. (1/8) Using our coding for…BR's answer to the question…is not robust. (3/8)

Arai, Karlsson & Lundholm (2011)	0	1	0	After removing missing values we are left with 818 observations. We cannot replicate any of their results and our estimations yield no support for their claims.
Ash & Robinson (2009)	0	1	0	…a coding error…Correcting the error changes the basic results of the paper…We also propose an alternative interpretation of the other main result…We offer two critiques of DL…in DL is unstable…neither inequality nor race…is adequately theorized in DL. (1/5)
Atkinson & Brandolini (2009)	0	0	1	the role of two of the three globalisation variables appears to be slightly strengthened in impact and significance, but there is no longer a significant effect of the outflow of direct investment (Page 399)
Aughinbaugh (2000)	1	0	0	The results confirm previous findings about … (1/12)
Bakke & Whited (2012)	0	0	1	To alleviate this issue, we use observations near funding thresholds and find causal effects of … but not on investment. (2/30) Although we can replicate the result of a strong negative correlation between mandatory contributions and investment… (28/30)
Balistreri & Hillberry (2007)	0	1	0	We find this claim tenuous.(1/13) AvW argue that the inclusion of theoretic structure…solves the border puzzle posed by MC…we show that AvW's adoption of…substantially reduces…This effect is magnified by AvW's pre-estimation treatment of US…we note that the AvW procedure does not account for...we show that neglecting this...reduces AvW's estimated border coefficient...theory-consistent border effects are only marginally smaller than...(11/13) ...once the symmetric cost assumption is relaxed, AvW's argument...can be overturned...cannot be empirically supported...(12/13)
Baltagi (2006)	0	0	1	While the…estimates are replicated, the fixed effects…are not. (2/6) This paper confirms…This result should be tempered by the fact that…(5/6)
Baltagi (2010)	0	0	1	While most of the estimates remain about the same…Their conclusion that the HT…is fragile… (2/3)
Beare (2008)	0	1	0	Unfortunately, the results of EF's study are faulty because… (4/11) The apparent support provided to…is clearly nothing more than… (8/11)
Bell & Miller (2015)	0	1	0	Using more appropriate methodologies than have previously been used, we find that dyads in which both states possess nuclear weapons are not significantly less likely to fight wars, nor are they significantly more or less belligerent at low levels of conflict. This stands in contrast to previous work. (Abstract)
Berger & Everaert (2009)	0	1	0	NNO … conclude that there is a cointegrating relationship between unemployment and labour market institutions. We challenge this result for two reasons. (p. 480)
Bhargava & Pathania (2013)	0	1	0	Our estimates imply an upper bound in the crash risk odds ratio of 3.0, which rejects the 4.3 asserted by Redelmeier and Tibshirani (1997). (Abstract)

Bloom, Canning & Fink (2014)	0	1	0	In a more general empirical framework… the AJ results are reversed. (10/13)
Born & Pfeifer (2014)	0	0	1	When we recalibrate the corrected model for the benchmark case of Argentina, we actually find that risk shocks matter more for output…and the contribution of interest rate risk shocks to business cycle volatility more than doubles. (3/10) The failure of FGRU's model to capture the cyclicality and volatility of net exports suggests that further research on the contribution of risk shocks to emerging market business cycles is needed. (9/10)
Bottasso, Castagnetti & Conti (2015)	1	0	0	We replicate the baseline specification of their study and we show that main results are robust to the use of a different estimation strategy…we also find a larger role for knowledge spillovers. (1/3)
Brandt et al. (2009)	0	1	0	This evidence suggests that the increase…was not a time trend but…(1/38)
Breznau (2015)	0	1	0	In 99.5 percent of the cases, addition of the main effect removes Brooks and Manza's empirical findings completely. (Abstract)
Brosig (2002)	1	0	0	Even with the more stringent restrictions that we placed on our experimental method, the findings of Frank et al. (1993) can be confirmed. (Page 282)
Brzezinski (2015)	1	0	0	Their results can be successfully replicated using a more refined power-law fitting methodology and a more comprehensive dataset. (1/6)
Budria et al. (2013)	0	0	1	We find evidence that risk attitudes are relevant but support is mixed. (1/25) We draw the general conclusion that Shaw's results are not very robust. We found little support for … However, we do find evidence… (23/25)
Burnside (2011)	0	1	0	I argue that consumption risk explains none of the cross-sectional variation in the expected returns of their portfolios. (2/22) I conclude that…the evidence for LV's consumption-based model is extremely weak. (4/22)
Campbell (2013)	0	1	0	This finding, while distinct from other estimated in the literature based on the same time period as GR… (2/17)
Carlsson & Johansson-Stenman (2010)	1	0	0	This implies that some central conclusions of CEHM survive regardless of whether or not their observed voting differences are due to scale differences. (6/7)
Cebi (2007)	0	1	0	The findings fail to support the predictions of the model… (1/15)
Chanda, Cook & Putterman (2014)	0	0	1	We are able to reproduce the AJR reversal in terms of the territorial entities that constitute present-day countries. But we show that with respect to the people who live in countries and their descendants, there was no reversal. (3/29)
Chakravarty (2015)	1	0	0	I find the theory is strongly supported by empirical results from Egyptian. (1/8)

Chang (2012)	0	0	1	Although we find very similar results on medical care utilization, I find that some yearly mortality rates that they report…are unreasonably low…I also find the NHI effect on mortality is larger for…(2/14) …finds similar estimates…on medical care utilization...the mortality rates reported in C...inconsistent with my findings...(13/14)
Chen (2012)	0	0	1	This paper replicates their original findings successfully…the evidence suggested that relative PPP does not hold in the long run. (1/10)
Cheung (2015)	0	0	1	I find that when the possibility of diversification is removed, while the element of risk is maintained, the effect observed by Andreoni and Sprenger (2012b) is reduced in magnitude by just over one-half. (p. 2259)
Ciccone (2011)	0	1	0	I show this finding is driven by a positive correlation between conflict in t and rainfall levels in t-2. In the latest data, conflict is unrelated to rainfall. (2/14)
Cohen, Lai & Steindel (2015)	0	1	0	Correcting for this, we find a statistically significant increase in the out-migration of taxpayers…our estimates may understate the actual impact of the tax on New Jersey net out-migration. (2/20)
Conti & Hansman (2013)	0	1	0	We show that, alternatively to the authors' conclusions, personality contributes to the education-health gradient to an extent nearly as large as that of cognition. (1/6)
Cook (2009)	0	0	1	While little evidence of stationarity was detected…rejection of the unit root hypothesis was observed under panel data unit root testing…Using…strong evidence in favour of stationarity in 11 of 13 economies examined. In contrast to…the results obtained...provide evidence in support of their I (0) inference...(2/8) the unit root hypothesis is rejected for three economies at 5% level, while....However, the unit root hypothesis is not rejected for the remaining eight economies. (4/8)
Ćorić (2016)	0	0	1	I find evidence affirming a positive relationship between CIA interventions and imports from the U.S., as well as evidence affirming most of the other results reported by BENS. However, my estimates indicate…can be explained by changed in tariffs. (2/9)
Couch & Placzek (2010)	0	0	1	The results presented here for Connecticut differ in some ways from those found in the work of JLS…(2/19) While some differences are observed in the experiences of…both studies find that there are no long-term losses…(3/19) While this difference in results across the types of estimators is not large...an important common finding in both samples...(17/19)
Cranfield et al. (2000)	0	1	0	the formulation presented here provides an alternative means to operationalize the estimation of AIDADS. Based on limited testing, the procedure appears to have some advantages over earlier estimation approaches. (Page 1914)

Croushore & Marsten (2016)	1	0	0	Our results show that the Rudebusch-Williams findings are robust in all dimensions. (Abstract)
Crump, Goda & Mumford (2011)	0	1	0	We do not find strong evidence to justify the model specification from the original paper. We also show that even if the original specification is correct, the results of Whittington are specific only to…and not robust to broader measures of tac subsidies...this finding casts additional doubt on the results. However, the total short-run effect is not statistically different from zero. (4/14)
Dalton & Norton (2000)	0	1	0	We find: (1) the functional form…is supported…no evidence of a threshold effect…no longer evidence of…no need to incorporate re-transformation factors into the payment formula. (1/20)
Dammon, Dunn & Spatt (1989)	0	1	0	Our estimate of the incremental value of restarting…is generally much smaller than that reported by C. (2/33) However, we do not find compelling the summary interpretation offered by C…There are two reasons for our scepticism. (3/33)
Davis (2007)	0	1	0	This paper reports on a failed attempt to replicate work by S…Further analysis of her original data set reveals numerous errors…(1/11)
Devereux & Hart (2010)	0	1	0	We find much smaller returns of about 3% on average with no evidence of any positive return for women and a return for men of 4%-7%. (1/21)
Douglas & Reed (2016)	0	1	0	...we demonstrate that this empirical support vanishes when…we conclude that there is insufficient evidence to support the hypothesis of a "small-state effect". (2/10)
Dube & Vargas (2013)	1	0	0	This replicates Angrist and Kugler's (2008) finding (Page 1409)
Durham, Geweke & Ghosh (2015)	1	0	0	The FILTER-DJI model proposed here provides one possible workaround for the estimation issues left unresolved by CJO. While the estimation results and model comparisons reported by CJO are not valid in the context of the GARCH-DJI model, they are valid...This note thus provides a constructive solution that reaffirms the usefulness of CJO's empirical findings. (2/5)
Duvendack & Palmer-Jones (2012)	0	1	0	The mainly insignificant impacts of microfinance differ greatly by gender of borrower, but are all vulnerable to selection on unobservables. We are therefore not convinced that the relationship between microfinance and outcomes are causal with these data.
Edlund (2000)	0	1	0	R's findings are not robust…I fail to replicate…(2/8)
Eisenhauer (2005)	0	1	0	In contrast to the results of ordinary least squares regressions, robust estimation and weighted least squares results indicate that the HVP may not be rejected at conventional levels of significance. (p. 465)

Elgers, Pfeiffer & Porter (2003)	0	1	0	We argue that the method DP use to measure…mechanically biases the evidence…we find that DP's results cannot be distinguished from those achieved…the "backing-out" approach…is ineffective. (1/18)
Everaert & Vierke (2016)	0	1	0	the strong relationship (i) disappears when cross-sectional dependence is accounted for using the CCEP estimator (Abstract)
Fain (1998)	1	0	0	All but…have the same signs as those reported by LS. Which tends to support their model. (2/8)
Fan & Mahal (2011)	0	0	1	we are able to replicate their results quite closely in terms of the outcome prevalence and the significance level, as we show next, the effects on every form of diarrhoea are not robust. (13/32) Our results on the effect of piped water on diarrhoea are consistent with a previous study by JR...but only if the outcome measure used is acute dysentery. (19/32)
Farbmacher (2012)	1	0	0	I have replicated their simulations for the linear model using the continuous updating estimator with CUE…(2/5) These are very similar to NW's Monte Carlo results. (3/5) The effects of education is slightly stronger when I use the CUE compared with 2SLS. (4/5)
Findlay & Santos (2012)	1	0	0	Our result, using the correct data, imply that player race has no effect on card prices-a result consistent with that originally reported by HMOR. (10/20) In this study we follow HMOR's approach of using a single performance statistics. Our study finds that race and ethnicity of baseball greats do not matter to the price of rookie cards. (17/20)
Fisher et al. (2012)	0	1	0	DG argues that predicted impacts are not economically significant…earlier research by us and others has found large potential impacts…Likely explanations for the divergence in finding include…the balance of evidence weighs heavily on the side of severe adverse potential impacts to US agriculture...(12/13) The predicted impacts are insignificant if we use DG's data...but are statistically significant...if we use our replicated weather variables and year fixed effects. While DG find insignificant impacts in their original paper...We find significant damages...(8/13)
Fisman & Love (2007)	0	0	1	We propose a more direct measure of growth…our direct growth measure outperforms their financial dependence measure and is less vulnerable…This still suggests an important role for…(1/11)…once this "growth opportunities" view is taken into account, RZ's interaction...is no longer statistically significant...our findings suggest that financial development is indeed important in...(3/11)
Fraas & Lutter (2012)	0	1	0	We argue that…its conclusion that…is simply incompatible with the empirical evidence presented in EPR…the source-specific trading ratios that EPR advocates lead to unattractive outcomes not likely to be efficient. (2/7)

Freeman (1999)	1	0	0	Corrected estimation using logarithmic first differences confirms Ruhm's original finding of pro-cyclical alcohol consumption, but these results, unlike Ruhm's, are robust to sample period. (p. 661)
Fu (2009)	0	1	0	A find that monthly stock returns are negatively related to…thus, their findings should not be used to imply the relation…I find a significantly positive relation…(1/14)
Garcia (2013)	0	1	0	A procedural replication of Ross's (2006) controversial finding that democracy has no effect on child mortality shows this null finding to be an artifact of the way quinquennial averages were computed, and the static nature of the preferred model. (Abstract)
Gardner & Diaz-Saiz (2008)	0	1	0	We reexamine this study and show that the accuracy of…can be improved by…Contrary to F, we show that…simple exponential smoothing with drift is the best smoothing method…the same accuracy as the robust trend. (1/5) For all method, little difference in forecast accuracy was found...(4/5)
Gerdtham & Trivedi (2001)	0	0	1	Our results indicate that the support for the inequity hypothesis reported by G is sensitive to model…(1/8) Our alternative modelling framework suggests a more qualified support for the inequity hypothesis…although…the FM results for doctor visits are qualitatively similar to the hurdle results...there is some evidence of higher utilization in the higher income group in poor health, but much weaker evidence...in good health...the FM model provides some qualitative evidence suggesting that "non-need" factors such as income and educational status boost health care utilization...(8/8)
Gerdtham et al. (1999)	1	0	0	Our results thus support the validity of the WvD method. (1/8)
Gilbert & Pfuderer (2014)	0	1	0	The consensus conclusion that CIT trading has no discernible impact on futures returns…is exaggerated if applied to the complete range of grains and oilseed markets. (Page 318)
Gisselquist (2014)	0	1	0	This paper argues that the findings of this article have been significantly overstated. Through a simple re-analysis of the data, it shows that ethnic diversity does not straightforwardly undermine public goods provision. (Abstract)
Goel & Mazhar (2015)	0	1	0	We fail to find a statistically robust effect of corruption on electoral outcomes. (1/12) These findings make KM's study less persuasive…we cannot support the hypothesis that corrupt incumbents are necessarily punished by voters. (10/12)

Goeree & Zhang (2014)	0	0	1	Our findings highlight the fragility of cheap-talk communication and may serve as a guide to refine existing behavioral theories. (1/18) We replicate recent findings by CD… however, this positive effect of communication is absent in our treatment with agent competition...However, none of the models by themselves can explain the substitute patterns between competition and communication that we observe in th experiments. (16/18) We find that in the "no-competition" treatments, communication raises efficiency. (2/18)
Gomes & Paz (2011)	0	0	1	The narrow replication exercise of Yogo finds results identical to the original papers…The null hypothesis of the Sargan test is rejected in several cases, in particular for US and UK quarterly data. These rejections cast doubts on either instrument validity or model specification. (2/3) ...for all instrument combinations US and UK samples present rejections of the null in the Sargan test...when the Sargan test was not rejected, Yogo's result that the estimated EIS is not statistically different from zero still holds. (3/3)
Gordon & Wang (2004)	0	1	0	The results of our statistical tests challenge L's finding…Finding from the estimations of both models do not support the conclusions about…reported by L…(6/31)…results shown here-which, contrary to L, find no effect on…Not only is the beneficial effect of ...cast into doubt, but...(10/31)
Gorodnichenko & Tesar (2009)	0	1	0	We show that the border effect identified by ER is entirely driven by the difference in the distribution of prices within the US and Canada. (2/24) Our arguments suggest that a comparison of…is insufficient for identifying border frictions. (4/24) we show the sensitivity of the border effect to the presence of cross-country heterogeneity. (5/24) Our results strongly suggest that reduced-form coefficients in border effect regressions should not be generally interpreted as...cannot credibly identify the impact of the border. (23/24)
Gottschalk (1981)	0	1	0	Using a consistent estimator of the income effect significantly modifies their conclusion. (Abstract)
Grafeneder-Weisstiner et al. (2009)	1	0	0	All the results relating to the data set: matched panel #3 … were replicated using Stata 10 … and EViews 6. (Page 1054)
Grant (2009)	0	0	1	We ultimately conclude that financial incentives may influence…but, if so, the effect is much smaller than originally advertised. (2/7)
Greig (1992)	0	1	0	After controlling cross-sectional differences…no significant incremental predictive ability is attributed to Pr. The Pr measure is interpreted as a proxy for…rather than as new…(1/30)
Guindon & Contoyannis (2012)	0	1	0	In contrast with earlier findings, on the whole, no discernible relationship between spending on private or public pharmaceutical products and infant mortality or life expectancy at 65 is observed. (1/19)

Guthrie, Sokolowsky & Wan (2012)	0	1	0	CG estimate that CEO pay decreases 17% more…We document that 74% of this magnitude is attributed to two outliers…we find that the compensation committee independence requirement increases CEO total pay…our evidence casts doubt on the effectiveness of independent directors in constraining CEO pay...(2/21) ...the results are fragile...after excluding these two outliers...our results indicate that: board independence does not affect the level of CEO pay...(18/21) there is little evidence that the board reform have had any meaningful effect on the level of CEO pay. (19/21)
Hand, Pierson & Thompson (2016)	0	0	1	The empirical results reported in Gore's article are largely replicable and that its results are robust to substantial data extensions. Nevertheless, we believe that Gore reaches normative conclusions that municipalities hold "excess cash reserves", which are not justifies by her empirical results. (1/7) However, we believe that Gore fails to sustain this hypothesis...the positive residuals from her first model...do not necessarily indicate excess cash...(6/7)
Haab (1998)	0	1	0	CQ provide evidence against the use of the standard interval data model…An apparent miscalculation…overstates the poor performance of…When the miscalculation is corrected…interval-data models provide robust estimates…(1/5)
Han & Heydecker (2006)	0	1	0	LL showed by use of a counter-example that solution of the minimisation formulation is not necessarily consistent with…(4/19)…However, in the present paper, we have shown that the dynamic user equilibrium solution can be found by…In the case of the novel objective function...minimisation is required sequentially over the study period. Strictly speaking, LL's example just confirms that...However, we have established here two appropriate formulations that do yield satisfactory dynamic user...(18/19)
Hannum (2016)	0	1	0	A re-examination of their evidence does not support that conclusion. (Abstract)
Hansen, McMahon & Srisuma (2016)	1	0	0	We show numerically, via simulation and re-estimation of the US Supreme Court data, that the first-order interaction effects that appear…can have an important empirical implication. (1/10) This note proposes a simple way that can help improve their estimates...we propose including interaction terms in the first-stage estimation...a re-estimation of...supports the simulation results. (2/10) ...the inclusion of the interaction terms reduces justice heterogeneity both in terms of variances and ranges...display notably less heterogeneity...the recent two-step methodology proposed by IS provides a useful way to... (8/10)
Harrison & Marsh (1998)	0	1	0	Using error correction models the analysis provides an alternative account of…The findings indicate that the short-term impact of the economy is weaker than, and different from, that suggested by them. (2/18)

Harrison (2003)	1	0	0	Title = "Successful Replication of Thornton's (2000) JMCB Article"
Hartog et al. (2003)	1	0	0	We replicate estimates…find that the variance of earnings in an occupation affects individual wages positively while…negative effect…(1/11) Consistent with…we find that…
Hatzinikolaou (2010)	0	1	0	several econometric errors, omissions, and confusions render its results questionable. (Page 109)
Haupt, Schnurbus & Tschernig (2010)	0	1	0	P found that a nonparametric approach…is superior to formerly suggested parametric and semiparametric specifications…a previously proposed parametric specification does not have to be rejected…(2/9) The null hypothesis that the mean of the ASEP for the parametric specification is larger or equal than that for the nonparametric specification is rejected...the parametric benchmark model predicts best...(6/9)
Herrmann & Thöni (2009)	0	0	1	We find that the distribution of types is very similar across the four locations. The share of conditional co-operators…is comparable to the one found by F…However, the distribution of the other types differs from the one found in Switzerland. (1/6)
Hill (2008)	0	1	0	S claims that…the growth-maximizing size…was about 19% of GDP. However, if an error in the model specification is corrected…estimates of the growth-maximizing size…range between 9% and 29% of GDP. Further the model spuriously identifies…The model cannot address reliably the question it attempts to answer. (1/10)
Ho, Huynh & Jacho-Chávez (2016)	0	0	1	Our replication finds that the application of the nonparametric copulas…provides an alternative flexible specification for copulas. However, the overall cautionary message of the flexible-form copulas espoused in Zimmer remains. (1/8) We were only able to exactly replicate...but not the other ones...Nonetheless, the overall shapes of both asymmetric tails in most cases remain. (3/8) Estimates from the mixture copulas show that asymmetric tail dependence for the right rail is enhanced...however,...for the left tail is increased...the new results display tighter confidence intervals in the left tail. (7/8)
Holthause, Larcker & Sloan (1995)	0	0	1	Like H, we find evidence consistent with the hypothesis that…Unlike H, we find no evidence that…We demonstrate that H's results at the lower bound are likely to be induced by his methodology. (1/46) Our results provide support for…but only in…we do not find evidence...In addition, we provide evidence that...Further, we provide evidence that...As such, we conclude that the results in our paper are very similar to H's findings conditional on the observation that...(4/46)

Howe, He & Kao (1992)	0	1	0	LL's findings suggest an important role for…In contrast, we find the market's reaction to…is approximately the same for both high-Q and low-Q firms. (2/14) LL show that a Tobin's Q-ratio…Consistent with free cash-flow theory, they find that the return is significantly higher for low-Q firms...than high-Q firms...(3/14)
Hsu, Huang & Tang (2007)	1	0	0	We find that the support of the minimax hypothesis is stronger. The plays in our data pass all of the tests in WW and therefore are more consistent with the theory of equilibrium than those in WW…the two hypotheses implied by the equilibrium…are borne out in our data. (1/8)
Humphreys (2015)	0	0	1	I do not challenge the above estimates, which remain the best published estimates for the Cambodian education premium… Those estimates, however, are based on an error in how Lall and Sakellariou interpreted the dummy variable coefficients. (p. 340)
Hung & Plott (2001)	1	0	0	The AH results are replicated for the individualistic institution. Furthermore, their results are robust to changes…(9/20) The results of AH replicate. In our experiments we observe the phenomena they report…It follows that the AH discovery is robust to changes in these classes of variables. (11/20)
Huynh & Jacho-Chávez (2010)	0	0	1	We have revisited the results from GLUW and found that the results are somewhat fragile…(5/5)
Ibrahim & MacPhee (2003)	0	0	1	Feder formulated the first model with an explicit mechanism connecting international trade and economic growth … Comparisons of the results among countries suggest that the impact of exports on growth depends on population size, trade orientation, and the importance of manufacturing. (Abstract)
Ilic (2014)	1	0	0	They do not find evidence of racial prejudice; in my own analysis, I, too, do not find such evidence. The present critique, then, does not arrive at results about prejudice contrary to their results. (p. 250-251)
Imai (2005)	0	1	0	The reanalysis of Gerber and Green's field experiment shows that get-out-the-vote calls increase turnout rather than decrease it. (Page 299)
Isoni, Loomes & Sugden (2011)	0	0	1	As in PZ's experiment, we found no significant WTP-WTA gap for mugs, thereby adding weight to that particular result. However, we also observed a significant and persistent gap for lotteries of much the same kind found in PZ's unre ported lottery data, suggesting that the PZ procedure does not in general eliminate the WTP-WTA gap. (p. 994)
Iversen & Palmer-Jones (2008)	0	1	0	B presumably failed to observe that their sample of unmarried women included a majority of widows. (31/44) We find no convincing evidence in support of literacy sharing…Our results contradict those in B, so we did not find reliable support for…This, therefore, cannot be interpreted as supporting the idea...(32/44)

Iversen & Söderström (2014)	0	1	0	This comment shows, first, that correcting an error in one of Steinsson's models leads to substantially lower persistence and volatility of he real exchange rate; second, that S's models cannot match real exchange rate volatility relative to output; and, third, that reasonable variations of the model calibration or specification all lead to lower real exchange rate persistence and volatility (or both). (2/19) Our findings show that S's results are not robust to reasonable variations in the model economy. (3/19)
Iversen, Palmer-Jones & Sen (2013)	0	1	0	We correct a mis-interpretation of the land revenue system in Central Provinces…we find no evidence that…(1/17) …there is no longer support for BI' key proposition. (2/17)
Jegadeesh & Titman (2002)	0	1	0	This article shows that CK reach this conclusion because they do not take into account the small sample biases…Our unbiased empirical tests indicate that …explain little…(1/15)
Johnson, Moorman & Sorescu (2009)	0	1	1	Although we add evidence to a growing literature that finds an industry component in stock returns...we do not identify whether the source of return variation comes from industry-specific risk, or unexpected industry-specific performance. A deeper understanding of the role played by industries in the cross-section of stock returns is left for future research. (p.4785). ...The significant results found by Gompers, Ishii, and Metrick (2003) and by Bebchuk, Cohen, and Ferrell are artifacts of either asset pricing model misspecification or unexpected industry performance.
Johnston, Cox & Barilla (2000)	0	1	0	This study … finds results that support neither hypothesis. (Abstract)
Jones & Ziebarth (2016)	1	0	0	We were able to replicate L's findings almost exactly. We extend L by showing that the findings also hold for the years 2004-2011 despite changing driver characteristics and restraint use patterns.
Kachelmeier & Towry (2005)	0	1	0	Our findings do not support JJ's conclusion…Our ceteris paribus replication leaves us unable to offer any generalized conclusions…(1/13)
Kalemli-Ozcan & Turan (2011)	0	0	1	He assigns all the fertility observations before 1990 with…this appears to drive the significant negative effect found in his study. When one restricts…the effect of HIV prevalence on fertility turns to be positive for South Africa. Simulating Young's model utilizing these new estimates shows that the future generations of South Africa are worse off. (1/5) .. a significant negative effect...1961-1998, a significant positive effect...1990-1998, and a zero effect...1986-1998. As a result we cannot draw any generalized conclusion. (4/5)
Kamhöfer & Schmitz (2016)	1	0	0	We can confirm the previous result and also find zero returns for other compliers in higher track school. (1/8)

Katayama, Ponomareva & Sharma (2011)	0	1	0	Our replication study shows that their result is driven by errors in the data. With correct data, it can be found that central bank independence is positively and significantly correlated with the sacrifice ratio, even when the nature of the political regime is controlled for. (2/9)
Keane & Sauer (2009)	0	1	0	We soundly reject the hypothesis that fertility and nonlabor income are exogenous. This is in sharp contrast to main result in H. (17/18)
Ketcham, Kuminoff & Powers (2016)	0	1	0	Our replication of their analysis shows that just over two-thirds of this estimated welfare loss is due to AG's interpretation of the econometric error terms as consumer mistakes. (Page 3933)
Killewald & Bearak (2014)	0	1	0	we find, in contrast to Budig and Hodges 's claims, that the motherhood penalty is not largest for low-wage women (Abstract)
Klein, Powell & Vorotnikov (2012)	0	1	0	We find many problems, and some of the problems seem to be quite important. Also, we apply falsification tests to their findings and the results are damaging. (2/25) It would be reasonable to judge the points of this section as fatal to LM's article. If WA are correct, then the data...was incorrect. Such a major discrepancy raises doubts ...used by LM. (7/25) Based on these findings we conclude that, consistent with our other criticisms, the data used by LM in the analysis was of low quality, and the positive effects...most likely spurious correlations. (16/25)
Klein et al. (2009)	0	1	0	By rebuilding the study of A, we are able to detect mistakes in the empirical set-up. Based on these findings, we demonstrate how even minor flaws can have a crucial influence on the results of such studies…A find a statistically significant relationship. (1/9) We replicate the study by A and were able to detect several inconsistencies in their event study set-up. Based on these findings, we present typical mistakes found in such empirical studies...extremely susceptible to errors and assumptions. (2/9)...how formerly significant results disappear when robustness checks are applied. (5/9)
Klößner & Wagner (2014)	0	1	0	Using this new algorithm, we find that the true range of the spillover index can be up to three times as large as estimated by DY. (1/8) Applying the new algorithm to…estimating the spillover index's range by examining a small number of permutations comes at the cost of severely underestimating the true range. (6/8)
Krol & Svorny (2007)	0	1	0	However, L's finding is not robust to…Reestimation of L's regressions…reverses his result. (1/15) This conflicts with L's theory…This reversal of L's results calls into question…(12/15)
Kulaksizoglu (2015)	1	0	0	We obain results that are quite close to their results. (1/1) We are able to replicate their reults reasonably closely.
Kuosmanen & Kuosmanen (2009)	0	1	0	This paper critically examines FH's estimator for opportunity cost, and shows that the proposed estimator rests on a number of strong, unrealistic assumptions…the proposed estimator performs very poorly even under ideal conditions. (1/9)

Kurmann & Mertens (2014)	0	1	0	This comment shows that, when…the identification scheme does not have a unique solution…their identification scheme fails to determine TFP new. (2/8) The results reported in BP represent just one arbitrary choice among these solutions…The identification scheme and results presented in BP therefore do not shed light on the importance of TFP news shocks for business cycles. (3/8)
Lall (2016)	0	1	0	in almost half of the studies, key results “disappear” (by conventional statistical standards) when reanalyzed (Abstract)
Lampach & Morawetz (2016)	0	1	0	We find that the data do not support the hypothesis that there is, on average, an effect of membership of a Fair Trade certified cooperative on per capita income…consumption. (8/12) With the unconfoundedness and the common overlap assumption being questionable, the main results of "no significant effect" is hard to justify. If we disregard the uncertainties about the unconfoundedness assumption, the estimated treatment effects suggest that there is no significant difference in income between producers from certified and produces from non-certified cooperatives. (9/12)
Lee (2005)	0	0	1	"partially successful" from Replication Wiki
Leimer & Lesnoy (1982)	0	1	0	This paper presents new evidence that casts considerable doubt on F's conclusion…Simply correcting this error substantially changes…Adopting reasonable alternative assumptions leads to generally weaker estimates…the estimated relationship…is acutely sensitive to...(3/25)
Levy & Roll (2015)	0	0	1	In summary, while B&L’s results are insightful and perfectly correct, they do not at all imply that the sample parameters are inconsistent with positive efficient portfolios. In fact, the opposite is true. So don’t bury the CAPM just yet. P6 Like them, we find that the sample parameters lead to an impossible frontier. But we show that a slight modification of the parameters, well within their estimation error bounds, leads to a segment of positive portfolios on the frontier. Moreover, this segment can be quite large. Thus, the sample parameters are perfectly consistent with a possible frontier. p6
Levy (2009)	0	0	1	This note shows that while the lognormal distribution fits the empirical data extremely well for 99.4 percent of the size range, as convincingly argued by E, in the top 0.6 percent…the size distribution diverges dramatically and systematically from the lognormal distribution...(1/5)
Lévy-Garboua et al. (2012)	0	0	1	We chose to investigate the consistency of HL's measure of risk aversion and its sensitivity to framing…HL found however that a non-negligible part of subjects exhibited inconsistency…we investigate whether changing the order of the probabilities…might influence the level of inconsistency...(3/17)

Malikov (2011)	1	0	0	This paper describes a generally successful attempt to replicate results of DH…although there are some minor discrepancies. (1/10)
Mazza, van Ophem & Hartog (2013)	0	0	1	Our results deviate from Chen's in several respects and we find non uniform relationship between uncertainty and level of education. However, a key conclusion stands firmly, both in Chen's results and in our own estimates: the contribution to wage inequality...(2/16)
McCrary (2002)	0	1	0	This comment points out that a weighting error in L's estimation procedure led to incorrect inferences for the key results of the paper. (2/9)
McCullough (2003)	0	0	1	Title = “Partially Successful Replication of Brunner's 2000 JMCB Article”
McDonald, Crossley & Worswick (2001)	0	1	0	The evidence does not support the hypothesis that changes in the immigrant selection process over time have led to an increase in the receipt probabilities … The evidence also does not support the hypothesis that … represent a significant drain on the public purse compared with nonimmigrants (Page 395)
McGeary (2003)	1	0	0	Title = “Successful Replication of Wong's (2000) JMCB Article”
Mekasha & Tarp (2013)	0	1	0	We re-examine key hypotheses, and find that the effect of aid on growth is positive and statistically significant. This significant effect is genuine, and not an artefact of publication selection. DP conclude that the aid effectiveness literature has failed to show that the effect of development aid on growth is positive and statistically significant. (1/21)
Mercer & Reed (2015)	0	0	1	We are able to exactly replicate their findings…With one exception, our robustness checks fail to find strong evidence that economic development variables…(1/24)
Merriman (2015)	0	0	1	I replicate the most widely cited result in the original article. Other results are substantively but not quantitatively replicated…GM's results are sensitive to relatively arbitrary choices…I argue that the most cited result in the article does not come from the most preferred econometric specification and that when...GM's original article found no statistically significant evidence...I find no statistically significant evidence for this hypothesis...(1-2/21) Overall, my replication...was quite close...I am able to econometrically replicate the ley substantive findings...(7/21)
Midtgaard, Vadlamannati & de Soysa (2013)	0	1	0	We advance the debate…questioning their crucial assumptions…Using their data, we find signing on to an IMF program predicts the onset of a civil war negatively…the operationalization…simply capture the effect of ongoing conflct rather than the effects of liberalization...at no time does IMF involvement successfully predict the onset of a civil war. (1/18)
Mishkin (1990)	0	1	0	The evidence in this paper suggests that this conclusion is unwarranted (Page 24)

Moody (2001)	1	0	0	LM's basic conclusions are generally robust with respect to…(2/16) The results of the above analyses confirm and reinforce the basic findings of the original LM study. (15/16)
Mueller (2012)	0	1	0	This note discusses an important methodological shortcoming in CS. We highlight the fact that CS code civil wars differently than all other crisis in their study. This leads to a misrepresentation of the output response to civil war. (2/5) we show that the output response for civil war displayed in CS misrepresent their impact. (4/5)
Nakov (2010)	0	0	1	I replicate most of the results in A…point to a possible error in and re-estimate Model 3…I am unable to replicate the authors' Monte Carlo results for Model 3…(2/5)
Nekby & Pettersson-Lidbom (2017)	0	1	0	We find that their results are based on an unreliable and potentially invalid measure of…a mismeasurement…Correcting for any of these three problems reveals that there is no evidence of any relationship…(1/21)
Nightingale (1988)	0	1	0	His analysis appears to be flawed in a number of ways, and replication with data from the UK and Australia does not support the model. (Abstract)
Nonejad (2016)	1	0	0	To conclude: we are able to reproduce the results of Chan et al. (2013). (Abstract)
Norton & Patrick (1985)	0	1	0	His conclusion that…is not substantiated…the results are likely to be influenced by hypothetical bias. Other problems with…are also discussed. (1/5)
Ortmann, Fitzgerald & Boeing (2000)	1	0	0	Our re-examination of the well-known BDM results suggests that they are quite robust. (8/20)
Petersen & Winn (2014)	0	0	1	In our experiments we find no evidence of first-order money illusion, but we do find evidence of second-order money illusion. (3/17)
Reed & Sidek (2016)	1	0	0	We are able to exactly replicate their findings…Our analysis produces results that are qualitatively similar to NP, though few of our results are statistically significant. (2/9) …our results generally support NP;s original findings…(7/9)
Rees & Sabia (2010)	1	0	0	Our results are generally consistent with those of C. (1/4)
Rock, Sedo & Willenborg (2000)	0	1	0	In contrast with the original paper, our finding indicate…is inversely related with analyst following. We also provide…to support the preferred use of the negative binomial…(1/23)
Romo (2016)	0	1	0	"different results" from Replication Wiki
Roodman (2015)	0	1	0	I exactly replicate the estimation results of CRBB and then question these results with…Addressing this issue produces evidence of zero or negative Granger causation from aid/GDP to growth. (1/26)

Rothstein (2007)	0	1	0	These results turn out to be quite sensitive to…only with H's particular streams variables…smaller estimates…sample selection bias…H's positive estimated effect of…is not robust…does not support…(2/13)
Ruser & Smith (1991)	1	0	0	The sizes and patterns of coefficients that we obtain in our analyses…are consistent with those found earlier. (Abstract)
Rydval et al. (2009)	0	1	0	We find no evidence for IA violations and hence, for the uncertainty effect…It therefore seems that the uncertainty effect phenomenon is less robust than GLW's results suggest. (2/15)
Sanga (2009)	1	0	0	My results largely confirm theirs…(6/6)
Savva (2016)	1	0	0	The current study replicates their main results and performs a similar analysis…their findings are confirmed to a large extent. (1/5)
Schneider (2016)	0	1	0	Using data from their study and new data from eBay, I provide evidence that a key condition for identifying nonstandard behavior may not have been met, and that the observed over- bidding is not inconsistent with standard behavior. (Abstract)
Schober & Winter-Ebmer (2011)	0	1	0	We replicate the analysis using…and do not find any evidence that more discrimination might further economic growth - on the contrary: if anything the impact of gender inequality is negative for growth. (1/9)
Schulz (2016)	0	1	0	In contrast to the previous results by BBK, I show strong evidence that is supportive of an unconditionally flat term structure of equity risk premia (Page 3186)
Scott (1997)	1	0	0	The reasons are somewhat similar to those found by BS for the UK. (1/22)
Seidl & Moraes (2000)	1	0	0	The Brazilian Pantanal is implied by this study to be a global "hot spot"…(2/6)…the Brazilian Pantanal is a uniquely valuable watershed to the global value of ecosystem services…(5/6)
Seshamani & Gray (2004)	0	0	1	Z have previously proposed that proximity to death is a more important influence on health-care costs than age…Using…to find that neither age nor proximity to death have a significant effect on hospital costs…a two-part model shows both age and proximity to death to have significant effects on quarterly hospital costs. ZFM found age to be insignificant in...(1/12) A first replication of the ZFM model showed an insignificant relationship between...However, once the econometric weaknesses...time to death significantly affected...(9/12) A two-part model instead showed that both proximity to death and age have significant effects on cost. However, the effects of age are much smaller than those of proximity to death, providing compelling evidence that...(12/12)
Siskind (1977)	0	1	0	While attempting to replicate W's results …he had inadvertently…The result of this error was to lower…thereby lowering…the data error was likely to severely bias W's regression estimates…(1/4)

Sjoquist & Winters (2012)	0	0	1	We are able to replicate her results exactly…however, when…the coefficient estimates are considerably smaller…Further analysis reveals that the differences across the samples are mostly concentrated among women…the statistical significance levels in D are greatly overstated... (3/17) ...find much smaller effects. D's clustered standard errors are downwardly biased and lead to invalid inferences...Both procedures suggest statistically insignificant effects of merit programs on degree completion...(13/17)
Spamann (2009)	0	1	0	A thorough re-examination of the legal data, however, leads to corrections for…many empirical results established using the original index may not be replicable with corrected values…the corrected index fails to support…(1/21)
Spilker & Böhmelt (2013)	0	1	0	And indeed, unlike in Table 3 above or Hafner-Burton (2005a), PTA hard law is highly insignificant throughout Models 4–7. (Page 356)
Spindler (2016)	1	0	0	We replicate the simulation study of B in R…all three functions give identical results. (3/5) We can replicate the results in B and the only difference concerns the Ridge-based estimator…it is of minor importance for our comparison. (4/5)
Sun, Henderson & Kumbhakar (2011)	1	0	0	We illustrate this with a successful replication of MP…(2/8)
Takahashi (2014)	0	1	0	I show that these results arise from errors in their computational method. I resolve the model using a corrected method and find a strong positive correlation between hours and productivity…(2/16)
Taylor, Kreisel & Zimmerman (2010)	0	1	0	We find an effect of tenths of a cent per gallon, which is of little economic significance…Our empirical results cast doubt on whether…(2/9)
Temple (1999)	1	0	0	I was able to replicate the key results of BS. (2/4) This note demonstrates the point using data and specifications from BS. (4/4)
Thompson & Fox-Kean (2005)	0	0	1	Doing so eliminates evidence of strong intra-national localization effects at the state and metropolitan levels, but leaves largely unaffected evidence of international localization effects. (1/12) While we continue to find evidence of international localization effects similar in magnitude to those found by JTH, there is no evidence of the remarkably strong intra-national localization reported in JTH. (2/12)
Tsui & Ho (2004)	0	0	1	We have successfully replicated results based on T's yen-dollar series. There is evidence of…but no support for…Unlike T, however, we find evidence of…we find support for…In contrast, the evidence of…is rather mixed. We find stronger support for…(6/7)
Van de Sijpe (2013)	0	1	0	Allowing for the presence of off-budget aid indicates that the degree of fungibility of health aid is much more uncertain than at first blush appears…the conclusion of full fungibility is overturned…(1/10)

Van Ophem (2011)	1	0	0	I first replicate part of W's research. I then set out to analyse whether the zero correlation is actually true or comes from…My empirical analysis confirms the latter, but nevertheless also corroborates W's main conclusions on…
Wagner (2015)	0	1	0	Replication failed completely. The link found between…is never in line with the results from CE. (1/10)
Wagner (2015)	1	0	0	We first show that it replicates the empirical regularities recently unveiled by (Eaton et al., 2011) (Page 1207)
Wang (2006)	0	1	0	We find that an important programming error was made by HF…The empirical results…are subject to errors. (2/3)
Wells (2003)	0	1	0	This paper questions the numerical results presented there…we must conclude that F's 1978 model is not as poorly specified as suggested by the PV article. (2/4)
White (1988)	0	1	0	H therefore concluded that the monocentric urban model has little predictive value…(2/15) I find that only around 11 percent of the actual amount …is wasteful. Thus waste in fact appears to be only a minor factor in explaining…(3/15) The result presented here suggest that monocentric urban models are in better shape than H's gloomy diagnosis would imply...(14/15)
Wildman, Gravelle & Sutton (2003)	0	1	0	The study finds that his results do not hold for a more recent data set and it suggests that his method may not overcome the aggregation problem. (1/8)
Wolfers (2006)	0	1	0	This paper argues that these conclusions are somewhat misleading. I find that the divorce rate rose sharply following the adoption of unilateral divorce laws, but that this rise was reversed within about a decade. There is no evidence that this rise in divorce is persistent. (1/20)
Xu (2011)	0	1	0	The paper argues that the results in HPR are fragile to changes in sample and measures…(1/19)
Xun & Lubrano (2016)	0	1	0	We find that the empirical results reported in C are contingent on the specification of the model. (1/6) we could not reproduce C's results…(6/6)
Yamamura & Shin (2007)	0	1	0	We found that if we incorporate year dummy variables…is not negative but positive. These results are contrary to the assertion of KR…(1/8)
Zarkin et al. (1998)	0	1	0	Whereas FZ found that individuals…we found no evidence of…(1/16)
Zeileis & Kleiber (2009)	0	0	1	As for all other data sets, we are able to successfully replicate the plain OLS regression coefficients and…after omitting those observations indicated by Z. However, we encountered problems with…Second, we could not reproduce…Fortunately, the results are essentially identical...the final robust OLS regression is the same...(4/14)
Zhang & Ortmann (2014)	0	0	1	We find E's meta-study of…to be robust, with one important exception…While E reports this as having no statistically significant effect…we find an economically and statistically significant negative effect on… (1/7)

Zhou (2015)	0	1	0	When we consider the influence of…the predicted relation…in DKLM does not appear to be robust…(2/21) After controlling for…the findings in DKLM become weakened and unstable…(17/21)
Zhu, Ash & Pollin (2004)	0	1	0	We show that the LZ results are not robust to…when one properly controls for…stock market liquidity no longer exerts any statistically observable influence on GDP growth. (1/10)
Ziegelmeyer, March & Krügel (2013)	0	0	1	We report the results of regression analyses identical to those performed in W except that…We confirm that…but we find that…are statistically significantly smaller than in W. (2/11)
Ziegelmeyer, Schmelz & Ploner (2012)	0	0	1	We largely confirm the existence of hidden costs of control but, contrary to the original study, hidden costs of control are usually not substantial enough to significantly undermine the effectiveness of economic incentives. (p. 323)