DEPARTMENT OF ECONOMICS AND FINANCE

UC BUSINESS SCHOOL

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Do Negative Replications Affect Citations?

Tom Coupé
W. Robert Reed

WORKING PAPER

No. 14/2021

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

WORKING PAPER No. 14/2021

Do Negative Replications Affect Citations?

Tom Coupé[1†]
W. Robert Reed[1]

November 2021

Abstract: This study examines 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. Approximately half of the replication studies refuted key findings from the original studies, with the remaining half either providing a confirmation or a mixed conclusion. 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, we match each replicated study with multiple controls based on having comparable citation histories. Our main samples consist of 74, 103, and 142 replicated studies and 7,044, 7,552, and 11,202 matched control studies, respectively. We have two main findings. First, studies that are replicated receive somewhat 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.

Keywords: Replications, Citations, Matching, Hierarchical Linear Modelling, Quantile Regression

JEL Classifications: A11, B41, Z00

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

† Corresponding author: Tom Coupé. Email: tom.coupe@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. We start with the negative evidence.

In 2015, Open Science Collaboration published the results of a large-scale reproducibility project that focused on 100 highly-cited experiments in psychology (Open Science Collaboration, 2015). They reported that only 36% of the replicated experiments produced statistically significant estimates in the same direction as the originals. Two subsequent studies compared the citation performance of studies with unsuccessful replications to those with successful replications. Yang et al. (2020) found no difference between the two sets of studies, while Serra-Garcia & Gneezy (2021), using an expanded database that added replications from economics and general science journals, found that studies with unsuccessful replications were actually cited more. Unfortunately, both studies were hindered in their analysis by the short-time frame following the publication of Open Science Collaboration (2015). Yang et al. (2020)'s analysis period stopped in 2017. Serra-Garcia & Gneezy (2021)'s in 2019.[1]

A 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 studies in biomedicine received 65% fewer yearly citations over the post-retraction period compared to

[1] The set of replication studies examined by Yang et al. (2020) were taken from Open Science Collaboration (2015). Thus, Yang et al. (2020) only had two years of post-replication data. Serra-Garcia & Gneezy's (2021) sample included replication studies from Camerer et al. (2016) and Camerer et al. (2018). Thus, in some cases, their sample only had one year of post-replication data.

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".

In summary, while the results from the retraction literature suggest that science is self-correcting, the evidence from the replication literature is less favourable, albeit thinner. 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, of which there are many more than the egregious outliers that get retracted. Observing how the literature

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

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 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.

II. Matching Strategy

The "Treated". We obtained replication studies from two websites that collect data on replications in economics, the Replication Network and ReplicationWiki. Together with two research assistants, we then located the Scopus ID numbers for (i) the replication paper and (ii) the original paper that was replicated by the replication paper. We focus on published replications and excluded replication papers that replicated more than one original paper, and original papers that were replicated by more than one replication paper. This gave us pairs of a replication and an original paper that were not linked to any other replications or original

[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.

papers. We further excluded pairs for which we had (i) less than 3 years of post-replication citation data (i.e., papers published after 2016) and (ii) 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.

We read through each of the replications and classified them as (i) negative, (ii) positive, or (iii) mixed. In almost every case we took the replication authors' own assessment. TABLE 1 gives two examples from each category. In most cases, the replication authors' assessments were clearly stated in the abstract and/or conclusion of their papers. We took the authors' own assessment rather than attempt to make our own judgment because we felt what matters for the impact of a replication is how a study is perceived by its readers, and readers' perceptions are likely most influenced by the authors' own explicit assessment. 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 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.

[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.

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 the 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.

Selection of Controls: Stage 2. The goal of Stage 2 was to find control studies that closely matched the 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. Different intervening years meant different lengths of matching periods. Second, studies differed in how many citations they had, with some studies having only a few citations, and others have hundreds, or even thousands of citations.

FIGURE 1 here

Every treated study included 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 replicated in 2017, we compared citation histories 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 intervening years. Matching is based on the sum of absolute differences over the citation history. 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 citation record of the control exactly matches the treated's.

TABLE 2 here

It is difficult to get perfect matches. As shown in TABLE 2, there are only a total of 2,201 perfect matches out of 112,000 possible controls. As a result, we have to loosen the criterion for matching controls to treated if we want a larger pool of controls.

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 match to not be as close. Accordingly, we define a variable that counts 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

Citations among the 204 treated in our sample vary widely. FIGURE 3 plots a histogram of citations over various groupings of $TotOrigCites_K$. 63 (31%) of the treated in our sample 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.

We compare $TotAbsDiff_K$ of the potential control with $TotOrigCites_K$ of the treated and keep the control as a match if the total absolute difference in citations over the citation history is less than a given threshold value. The threshold value is specified as a function of the percent (*PCT*) of the number of citations of the treated study at the time the replication was published ($TotOrigCites_K$). Specifically, the decision rule matches a control with the treated if

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

where *PCT = 0%, 10%,* and *20%.* TABLE 3 shows the threshold values corresponding to different values of $TotOrigCites_K$ and PCT.

TABLE 3 here

When $PCT = 0\%$, the matching criterion states that the total absolute difference in citations between the treated and the control cannot be larger than 1 citation over the respective 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 shared a three-year citation history, where the treated study had a total of 20 citations. 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 the original study had 200 replications, 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". 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. However, results for all subsamples of *K* are included in the supplementary files that accompany this study.

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 an 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.

III. Results: The Effect of Replications on Citations

Before we estimate the effect of a negative replication, 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) $DIFF_{it,i\in K} = Citations_{it,i\in K}^{Treated} - Citations_{it,i\in K}^{Control}$

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. In the analysis that follows, we report 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.

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. While the differences are small in size, several are significant at the 5-percent level. For example, under the *PCT = 10%* matching regime, treated received 0.345 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 0.150 and 0.170 additional citations. The latter value is statistically significant at the 5-percent level.

Statistically significant differences in the pre-replication period raise concerns about the ability of our matching procedure to achieve balance in citations between treated and controls. They suggest that differences observed 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.

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? A reasonable conjecture 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. Or it could come in the form of readers of the replication updating the importance they attached to the original study. If a study is replicated and the replication is published, that could be taken as evidence that the replicated study must be important. For either or both reasons, if readers have greater awareness of a study, they are more likely to cite it.

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 in the same year they were 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 replicated studies having been circulated as working papers prior to publication. This would give time for readers to the replicated study to attract readers, have increased awareness of the replicated 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,

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

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. The β_0 and β_1 from this regression should closely approximate those from the individual year regressions, differing only because of the restriction that the errors have common variance across years.

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 results 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; none are statistically significant. 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 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.

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.

While it is only one estimate, the finding that a negative replication could have a positive impact on citations is puzzling.[6] While not reported here, we have occasionally seen this result in other of our regressions, though only for $t = 0$; some using 5-year post-replication periods, and some using alternative estimators such as panel random effects or OLS. Whenever the results were statistically significant, they were positive.

Why would a negative replication generate more citations than a positive or mixed replication? We can only conjecture. It may be that negative replications attract more attention than positive replications. As a result, more researchers become aware of the original study. Greater awareness of the original study may result in increased citations. It is also possible that the extra citations may appear in articles extolling the benefits of replication. Specifically, how replications can change our assessments of previous research, and that the studies with failed replications are given as examples.

TABLE 10 here

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. 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

[6] Serra-Garcia & Gneezy (2021) also report that negative replications are associated with increased citations.

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. 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.

On the other side, 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 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, we now find no significant differences for the *PCT* = 0% and *PCT* = 10% samples. In fact, the estimated median difference in citations during the pre-replication 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. 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 the 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.

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 samples consists of 74, 103, and 142 replicated studies (the "Treated") and 7.044, 7,552, and 11,202 matched control studies, respectively. We have two main findings: First, 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 compare 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. Replications appear to provide a positive lift to the citations of replicated studies, but the effect is small.

Second, 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 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.

Can our results be interpreted as evidence that science is not "self-correcting"? There are many reasons why replications may not diminish the influence of a study that has been "proven" wrong. One possibility is that researchers are unaware of the findings of replications. 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 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.

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.

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.

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.

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.

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.

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. *Scienc*e, 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.

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

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.

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.

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*

	Matching Criteria		
K	***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:**	$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{0.11}$
		t = 0.52
		p = 0.605
			$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{0.51}$
			t = 0.99
			p = 0.320
				$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{0.70}$
				t = 1.41
				p = 0.159

(9)	**Test of overall post-replication effect:**	$\sum_{t=1}^{5} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{4.01}$
		t = 1.37
		p = 0.170
			$\sum_{t=1}^{5} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{-1.48}$
			t = -0.29
			p = 0.775
				$\sum_{t=1}^{5} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{-2.11}$
				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:**	$\sum_{t=-3}^{-1} \hat{\beta}_{1t} = 0.04$
		t = 0.32
		p = 0.746
			$\sum_{t=-3}^{-1} \hat{\beta}_{1t} = 0.11$
			t = 0.61
			p = 0.540
				$\sum_{t=-3}^{-1} \hat{\beta}_{1t} = -0.06$
				t = -0.22
				p = 0.825

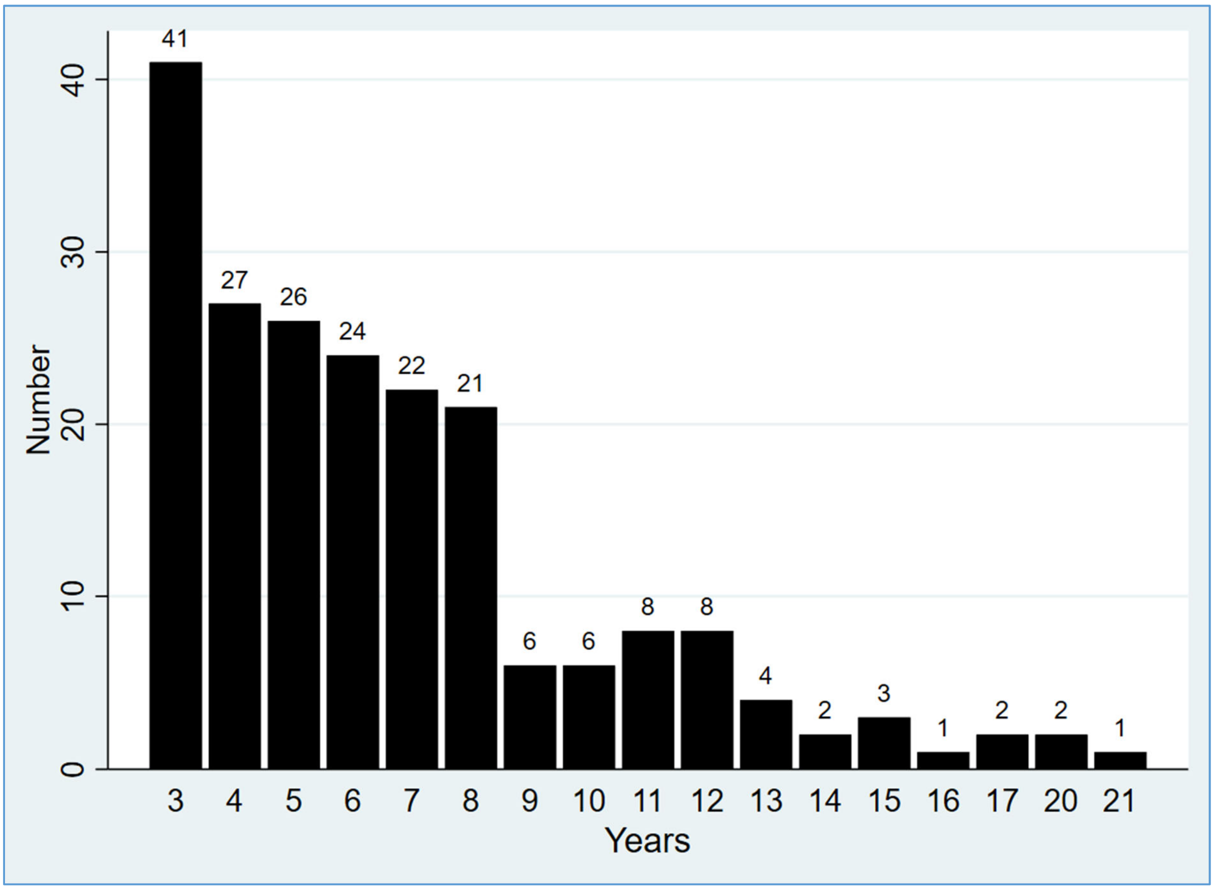
(9)	**Test of overall post-replication effect:**	$\sum_{t=1}^{5} \hat{\beta}_{1t} = 2.49^{*}$
		t = 1.92
		p = 0.055
			$\sum_{t=1}^{5} \hat{\beta}_{1t} = 1.27$
			t = 0.69
			p = 0.488
				$\sum_{t=1}^{5} \hat{\beta}_{1t} = 0.72$
				t = 0.43
				p = 0.666

(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.

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		$Citations_{T-2}^{Treated}$	$Citations_{T-1}^{Treated}$	

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

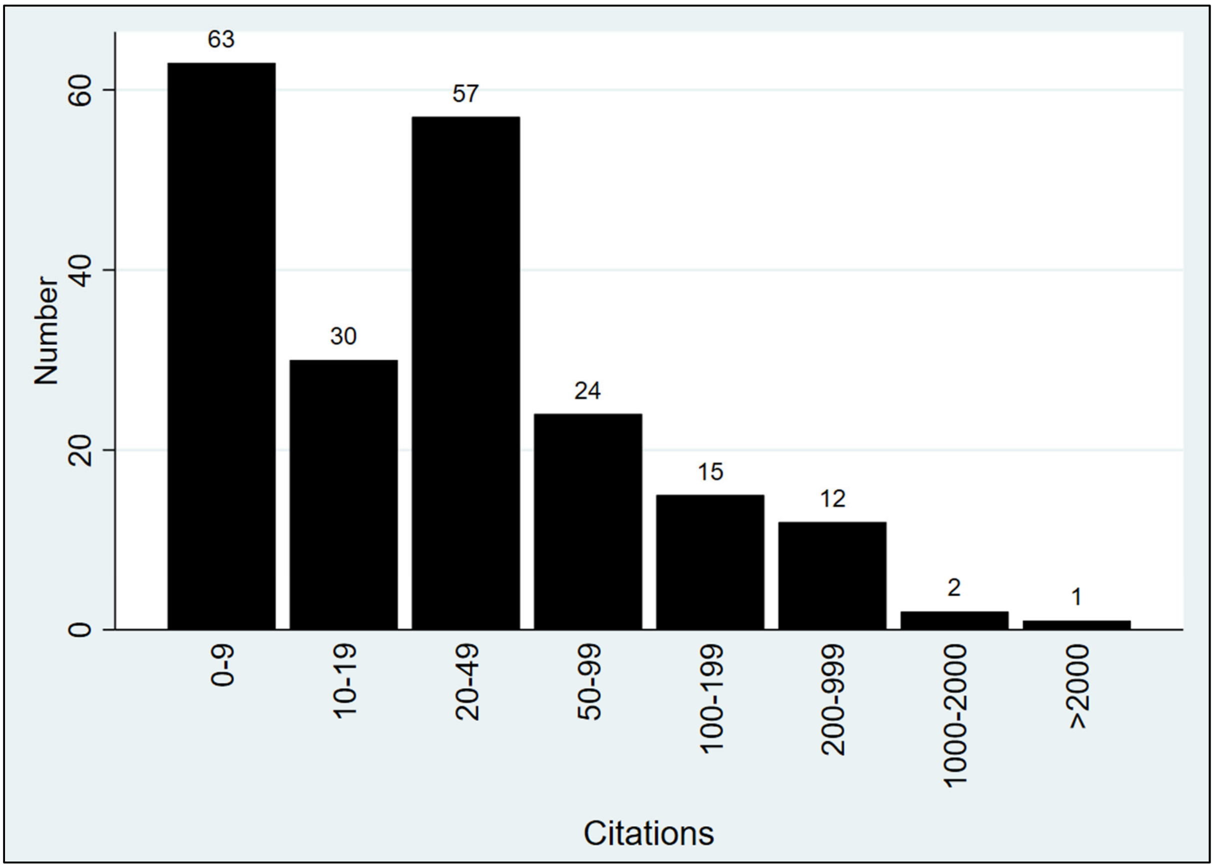
B. Four-year gap (K=4) between publication of original and replication

	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		$Citations_{T-3}^{Treated}$	$Citations_{T-2}^{Treated}$	$Citations_{T-1}^{Treated}$	

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

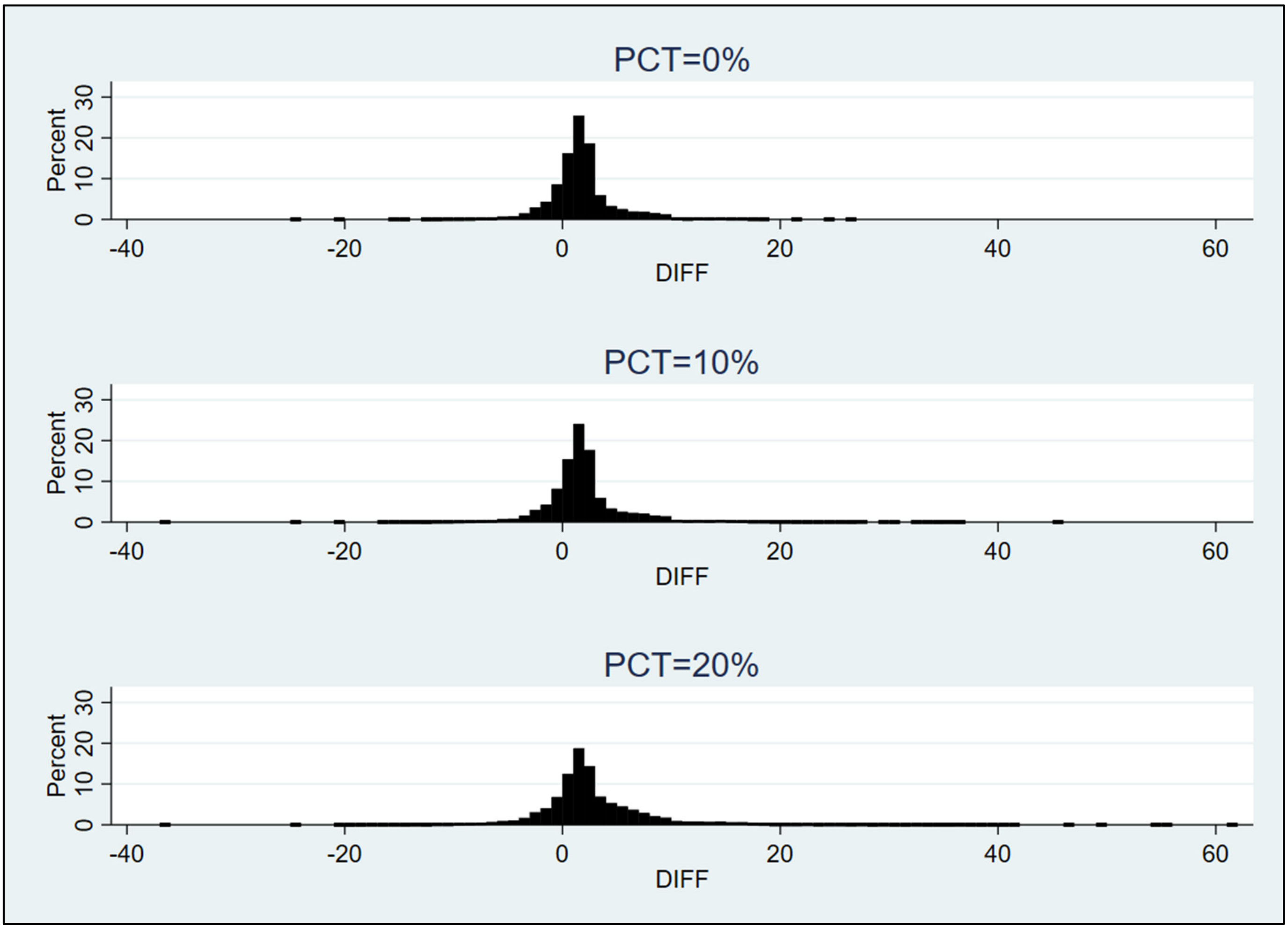
NOTE: This figure illustrates the relationship between years difference between when an original and its replication were published, and number of intervening years available to compare citation histories.

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).