DEPARTMENT OF ECONOMICS AND FINANCE

SCHOOL OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Do Replications Make a Difference?

NOTE: This paper is a revision of University of Canterbury WP No. 2022/22

Tom Coupé
W. Robert Reed

WORKING PAPER

No. 10/2023

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

WORKING PAPER No. 10/2023

Do Replications Make a Difference?

Tom Coupé[1]
W. Robert Reed[1†]

August 2023

Abstract: This study examines the effect of negative replications on the citation rates of replicated studies. It makes three contributions. First, we explain why previous research has not adequately addressed this subject. Second, we develop a matched difference-in-difference (DID) procedure that does not assume parallel trends (PT). Previous research has shown that studies that fail to replicate have different trends prior to replication than studies that successfully replicate. Given this difference, imposing the assumption of PT biases estimation of citation effects. Our DID procedure avoids this bias. Lastly, we study a set of 204 replicated studies and investigate whether there is a citation penalty associated with negative replications. Replicated studies are matched with non-replicated studies, with the matched controls being used to "predict" the counterfactual citation performance of replicated studies. Our preferred estimates indicate that studies that fail to replicate receive more citations than studies that have positive or mixed replications. Our less preferred estimates, based on looser matching criteria, find evidence of a citation penalty for negative replications, but the estimated effects are small and statistically insignificant. We conclude that replications have not been correcting the scientific record in the manner that proponents might have hoped.

JEL Classification: A11, A14, B41, C18

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

Data Availability Policy: All the programs to reproduce the results in this paper, along with much of the output, is publicly available here: https://osf.io/m9dpr/. The associated data files, being approximately 17 MB each, are too large to store on OSF and are available upon request.

[1] Department of Economics and Finance, University of Canterbury, NEW ZEALAND
† Corresponding author: W. Robert Reed. Email: bob.reed@canterbury.ac.nz

1. Introduction

In recent years the field of economics has seen increased initiatives promoting reproducibility and replication. In 2018, the American Economic Association (AEA) appointed a data editor whose mandate was to "promote reproducible research" (Duflo & Hoynes, 2018). Several online platforms, such as *The Replication Wiki*, *The Replication Network*, and the *Institute for Replication* have emerged to foster and facilitate replication studies. And numerous journals, such as the *Journal of Applied Econometrics*, the *American Economic Review*, *Critical Finance Review*, *Econ Journal Watch*, and the *Journal of Comments and Replications in Economics* now have dedicated replication sections and/or frequently publish replication-based papers. Others, like *Energy Economics*, *Public Finance Review*, *Applied Economic Perspectives and Policy*, *Economics*, the *Journal of Development Studies*, the *Canadian Journal of Economics*, and *Economic Inquiry*, have published or are publishing special issues featuring replications.

There are many potential benefits to replication. Replications, along with supporting efforts to make research reproducible, can help establish the validity of previous research, discourage questionable research practices and p-hacking, promote best practices, and encourage the standardization of research methods. But do they actually make a difference?

Quantifying the impact of replications is difficult. In this study, we focus on citations. Citations are the most commonly used measure of academic influence. If replications are making a difference, we would expect to see them having an impact on the citations of the studies they replicate. Accordingly, we investigate whether studies that fail to replicate suffer a citation penalty. Ours is not the first study to attempt to do this. However, as we argue below, the question of impact remains largely open given the shortcomings of previous research.

Previous research.[1] Three recent papers attempt to estimate the causal impact of failed replications (Serra-Garcia & Gneezy, 2021; Schafmeister, 2021; and von Hippel, 2022).[2] All conclude that studies do not suffer a citation penalty when a replication fails to confirm their findings. All three studies relied either heavily or exclusively on a widely cited set of replication experiments that were published in Open Science Collaboration (2015).[3] OSC drew their data from multiple labs performing replications of 100 experiments in psychology. Their finding that a large percentage of experiments failed to replicate garnered widespread attention.

Serra-Garcia & Gneezy (2021), Schafmeister (2021), and von Hippel (2022) tracked the studies included in OSC before and after its publication in 2015 to see whether studies that failed to replicate suffered a citation penalty. Underlying these citation analyses is the assumption that once OSC was published, readers were immediately made aware of the non-replicability of the respective studies. However, that assumption is a significant leap.

It is quite challenging to identify the specific studies included in OSC.[4] A list of the experiments selected for replication by OSC is not supplied in their paper. Nor is it included in the published article's 26-page supplementary documentation. To identify them, one needs to search the spreadsheet posted at Open Science Framework that contains their data. Even if one makes it that far, identifying which studies failed replication is difficult. The information in the spreadsheet is voluminous and contains multiple performance assessments using different measures of replication "success". This is easily confirmed here: https://osf.io/fgjvw/.

[1] The material for this section draws heavily from Reed (2023, April 5) and Reed (2023, April 16).

[2] Ankel-Peters, Fiala, & Neubauer (2023) also study the effect of replications on citations, but they do not attempt to estimate a causal effect.

[3] Serra-Garcia & Gneezy (2021) also included results from Camerer et al. (2016) and Camerer et al. (2018), though these constituted a minority of their analyses.

[4] Von Hippel (2022) explicitly acknowledges this in his study: "*Our findings are limited to the 100 studies replicated by the OSC, and we should acknowledge that the purpose of those replication studies was unusual. The goal of the OSC replication project was to estimate the reproducibility of psychological research in general—not to call out specific findings that were or were not successfully replicated. Although the results of all 100 replication studies were made publicly available, there was no effort to draw attention to individual results*" (page 1562).

In addition, shortly after OSC published their findings, Etz & Vandekerckhove (2016) concluded that there was insufficient power to claim that the original studies had failed to replicate: "*We revisit the results of the recent Reproducibility Project: Psychology by the Open Science Collaboration…The majority of the studies (64%) did not provide strong evidence for either the null or the alternative hypothesis in either the original or the replication, and no replication attempts provided strong evidence in favor of the null*." Even if one was able to determine that a particular study had failed to replicate, Etz & Vandekerckhove would have had the effect of raising doubts among researchers that RPP's results were sufficiently powerful to be convincing.

Thus an alternative interpretation of Serra-Garcia & Gneezy's (2021), Schafmeister's (2021), and von Hippel's (2022) null finding is that researchers were either unaware that the respective studies had failed to replicate, or unconvinced by the strength of the evidence. In contrast, as we discuss below, the set of replications that we analyse are clearly identified, have conclusions that are (mostly) unambiguous, and have each passed through a peer review process in which objections to the findings had to be addressed.

A framework for estimating the citation impact of failed replications and the importance of the assumption of parallel trends. While the problems above are sufficient to highlight the need for additional study, there is a further issue that serves as a methodological introduction to our study. Serra-Garcia & Gneezy (2021), Schafmeister (2021), and von Hippel (2022) all employ a Difference-in-Differences (DID) identification strategy that relies on the assumption of parallel trends (PT).[5] This is evident from the specifications below.

[5] Von Hippel (2022) also estimated a "lagged model" that does not assume parallel trends: $Y_{it} = \beta_0 + \beta_1 Y_{it-1} + \beta_3 Failure_i + \beta_4 Failure_i \times Y_{it-1}$, where the last term allows the pre-treatment period, citation trends of unsuccessfully replicated studies to be different from those that are successfully replicated.

Serra-Garcia & Gneezy:

$$Y_{it} = \beta_{0i} + \beta_1 Success \times AfterReplication_{it} + \beta_2 Success_{it} + \beta_3 AfterReplication_{it} + \boldsymbol{Year\ Fixed\ Effects} + Control\ Variables.$$

Schafmeister:

$$Y_{it} = \beta_{0i} + \beta_1 Successful_{it} + \beta_2 Failed_{it} + \boldsymbol{Year\ Fixed\ Effects} + Control\ Variables.$$

Von Hippel:

$$Y_{it} = \beta_{0i} + \beta_1 AfterFailure_{it} + \boldsymbol{Year\ Fixed\ Effects} + Control\ Variables.$$

In the above, Y_{it} represents a measure of the citations of study *i* in year *t*. The estimate of the citation penalty for failed versus successful replications is given respectively by β_1, $(\beta_2 - \beta_1)$, and β_1. While the three specifications have differences, all three include a common time trend for both failed and successful replications, represented by "Year Fixed Effects". This imposes the PT assumption on the estimating equations.

Interestingly, Figure 3 in Serra-Garcia & Gneezy (2021) undermines the PT assumption that forms the basis of that study's results. We reproduce this figure in our FIGURE 1.

FIGURE 1 here

The vertical axis reports "yearly citation count". The horizontal axis reports year, and the vertical line shows the year that OSC published their replication findings (2015). Of particular interest is Panel C, which plots citations for the OSC studies. There is a clear difference in the pre-treatment citation trends of studies that were not successfully replicated (black line) and those that were (blue line). Studies that failed to be replicated were cited at a higher rate during the pre-treatment period and, importantly, the difference in citation rates was growing even before the treatment. In fact, this is the main point of Serra-Garcia & Gneezy's article.

FIGURE 2 illustrates the bias that results from incorrectly assuming PT when estimating a DID regression. It highlights the case where a failed replication negatively affects

a study's citations and a successful replication results in more citations. As above, the vertical axis represents annual citations (Y) and the horizontal axis shows time in years (T). T* is the year that the outcome of the replications is revealed (failed replication, successful replication).

FIGURE 2 here

The black lines (blue lines) represent the citation trends for studies that had failed (had successful) replications. Solid lines represent actual, observed citations. Dotted lines represent the counterfactual where the researcher assumes that the pre-treatment trends would have continued had the studies not been replicated.

The top panel identifies the treatment effect we want to estimate. For studies with failed replications, the effect is given by the difference in slopes between the actual trend line and the counterfactual trend line, represented by A, $A < 0$. The citation effect for successful replications is given by B, $B > 0$. The total citation effect of a failed replication versus a successful replication is thus (A-B).

The middle panel represents the case where the researcher assumes that the citation trends have the same slope. The bottom panel then shows how estimation of the citation effect is affected when this common slope is assumed for the counterfactuals. When the averaged, common trend is used to establish the respective counterfactuals, both |A| and |B| are underestimated, so that the total citation effect of a failed replication versus a successful replication is underestimated. As a result, conventional DID will produce biased estimates of the citation effect of failed replications.

The problem with conventional DID lies in constructing an appropriate counterfactual. The set of replicated studies consists of two treated subsamples: those that receive the "treatment" of a failed replication, and those that receive the "treatment" of a successful replication. The identifying assumption is that the two subsamples are identical in their citation behaviour in the period before the "treatment" is applied. As Serra-Garcia & Gneezy

convincingly demonstrated, this is not true for the OSC papers and is likely not true for other sets of replication studies.

To properly estimate the citation effect, we need to construct counterfactuals that allow different pre-treatment citation behaviors for the two subsamples. Our approach is to use non-replicated studies whose citation behaviors before "treatment" match those of the replicated studies. For every study in our sample that has had a replication, we locate other studies that have not been replicated but have near identical, pre-treatment citation histories. We illustrate our identification strategy in FIGURE 2.

Consider first original studies with failed replications. In the context of FIGURE 2, one can think of two citation trends that lie on top of each other in the pre-treatment period, one each for the replicated and matched, non-replicated studies. Accordingly, the solid black line in FIGURE 2 in the period $T < T^*$ now represents citation trends for both the original studies and their matched controls.

Once we enter the post-treatment period, the two citation trends diverge. The dotted black line is the observed citation trend of the matched controls, which serves as the counterfactual for the original studies. The difference in slopes represents the treatment effect of a failed replication relative to its matched controls. The same story applies to studies with successful replications, which are now represented by the blue lines in FIGURE 2.

In both cases, the matched controls are being used to predict what the citation trends of studies with failed and successful replications would have been in the absence of the original studies being replicated. The identification strategy relies on the assumption that the post-treatment differences between matched treated and controls for studies with failed replications are not systematically different from the post-treatment differences between matched treated and controls for studies with successful replications. Or, in other words, the predictions of the

matched controls for studies with failed replications are not more or less biased than the predictions of the matched controls for studies with successful replications.

Outline of paper. We proceed as follows. Section 2 derives a matched DID estimator that does not require the assumption of PT. Section 3 describes the process by which we assembled our data and some of the key features of our sample. Section 4 specifies the regression equations that we estimate and the hypotheses we test. Section 5 presents the results. Section 6 summarizes and concludes our study. Our consistent finding is that there is no evidence of a citation penalty for studies that failed to replicate.

2. A Matched Difference-in-Difference (DID) estimator that does not require PT

In this section we derive a two-period difference-in-difference (DID) estimator applied to matched data that does not require the PT assumption.[6] Define:

(1) $Citations_{it}^{Treated} = Citation\ of\ treated\ article\ i\ at\ time\ t$

(2) $Citations_{it}^{Control} =$
$Citation\ of\ contol\ article\ \ matched\ with\ treated\ article\ i\ at\ time\ t,$

where t indicates either the pre- or post-treatment period. Let $P = \{Original\ Studies\ with\ Positive/Mixed\ Replications\}$, N_P the number of original studies with positive or mixed replications, and define Δ^{Pos} as the DID estimator of the treatment effect for positive/mixed replications relative to nonreplicated, matched controls.

(3) $$\Delta^{Pos} = \sum_{i\in P}\left(Citations_{i,Post}^{Treated} - Citations_{i,Pre}^{Treated}\right)/N_p - \sum_{i\in P}\left(Citations_{i,Post}^{Control} - Citations_{i,Pre}^{Control}\right)/N_p$$

We can rewrite the above as

(4) $$\Delta^{Pos} = \sum_{i\in P}\left(Citations_{i,Post}^{Treated} - Citations_{i,Post}^{Control}\right)/N_p - \sum_{i\in P}\left(Citations_{i,Pre}^{Treated} - Citations_{i,Pre}^{Control}\right)/N_p$$

This can be simplified as:

[6] For a general analysis of the benefits and costs of matching on pre-treatment outcomes prior to DID, see Ham & Miratrix (2022).

(5) $\Delta^{Pos} = \sum_{i \in P} \frac{Diff_{i,Post}}{N_P} + \epsilon_{Pre}^{Pos}$

where $Diff_{i,Post} = \left(Citations_{i,Post}^{Treated} - Citations_{i,Post}^{Control}\right)$ and

$\epsilon_{Pre}^{Pos} = \sum_{i \in P}\left(Citations_{i,Pre}^{Treated} - Citations_{I,Pre}^{Control}\right)/N_P$.

Note that $\epsilon_{Pre}^{Pos} \cong 0$ since Treated and Controls are matched in the pre-treatment period.

Similarly, for studies with negative replications:

(6) $\Delta^{Neg} = \sum_{i \in N} \frac{Diff_{i,Post}}{N_N} + \epsilon_{Pre}^{Neg}$

where $Diff_{i,Post}$, N_N and ϵ_{Pre}^{Neg} are defined as above except the respective studies belong to $N = \{Original\ Studies\ with\ Negative\ Replications\}$. Similarly, $\epsilon_{Pre}^{Neg} \cong 0$ because of pre-treatment matching.

We can estimate the respective DID estimators in Equations (5) and (6) from the following OLS regression equation:

(7) $Diff_{i,Post} = \beta_0 + \beta_1 \cdot Negative_i + \varepsilon_i$

where $i \in \{All\ Original\ Studies\ with\ Replications\}$, β_0 estimates Δ^{Pos}, β_1 estimates $(\Delta^{Neg} - \Delta^{Pos})$, and $Negative_i$ = a dummy variable that takes the value 1 for studies that have negative replications. Underlying the estimation of β_0 as a reliable measure of Δ^{Pos} is the assumption that $E(\varepsilon_i) = 0$. In words, this implies that the mean difference between the citations of treated and controls in the pre-treatment period is expected to equal zero.

Underlying the estimation of $\hat{\beta}_1$ as a measure of $(\Delta^{Neg} - \Delta^{Pos})$ is the assumption that $Cov(Negative_i, \varepsilon_i) = 0$. This implies that the expected difference between the citations of treated and controls in the pre-treatment period for studies with negative replications is the same as for studies with positive/mixed replications. It is difficult to imagine scenarios where citation differences between matched treated and controls are systematically different for studies with negative replications compared to studies with positive/mixed replications.

It is good to review our identification strategy in the context of the estimation procedure above. In order for Δ^{Pos} to identify the causal, citation effect of positive/mixed replications, we

require the matched controls to be unbiased predictors of the citations of the treated studies in the absence of replication. In contrast, in order for $(\Delta^{Neg} - \Delta^{Pos})$ to identify the causal effect of the difference in citation effects of negative versus positive/mixed replications, all that we require is that the predictions of the matched controls for studies with negative replications are no more or less biased than the predictions of the matched controls for studies with positive/mixed replications. Our analysis focuses on the estimation of β_1 in Equation (7).

3. Data

The "Treated". In assembling our data, the first issue we need to address in is how to define a "replication". The term is used in many different ways. It can mean anything from verifying that 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 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 of the replication 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. This makes it more likely that the replication will be known by the authors of the papers that cite the original.

We then paired 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; that is, a replication that does not successfully replicate the original. 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 be the pertinent criterion. When estimating the effect of a policy intervention, effect size may be what's important. An important issue in deciding upon a criterion for replication "success" is that it be one that readers can easily identify.

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. As noted above, OSC used multiple criteria to determine replication "success". Even if a reader were given a list of studies included in OSC, different readers might have come to different conclusions depending on which measure of replication "success" they adopted. We avoid this ambiguity by using the author's own assessment.

A further benefit of this approach is that the author's assessment is often prominently reported in the abstract and/or conclusion where readers are sure to see it. When it comes to impact on citations, 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.

For all of these reasons -- (i) clarity of assessment, (ii) prominent placement of assessment, and (iii) certification by the peer review process -- we argue that using authors' own assessments provides the information that is most likely to impact a study's citations.

Authors' assessments were categorized into three categories: negative, positive, and mixed. 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 the Appendix. 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").[7] With this number, we were able to extract their corresponding year-by-year citation histories from Elsevier's API. All of the citations come from published sources. From the same source we also extracted information about their year of publication, the journal in which they were published, and their volume and

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

issue number[8]. 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 original 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. Some 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 whose citations closely matched those of the treated studies prior to the replication being published ("the pre-treatment period").

Selection of Controls: Stage 2. For each treated study, 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. Specifically, we take the absolute value of the difference in citations between the treated and the control for each year and sum over the citation history ("$TotAbsDiff$"). 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

[8] 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.

period, but B more closely matches the year-by-year citations of the treated study ($TotAbsDiff_A = 4; TotAbsDiff_B = 2$).

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.

FIGURE 3 here

FIGURE 3 shows the distribution of the difference in years between publication of the original and publication of the replication for the 204 treated studies in our sample. We note that a minimum of three years' difference was required so that we had a sufficient pre-treatment period on which to match. The number of treated studies in our sample decreases monotonically with the length of the pre-treatment, citation histories (K). After eight years, it falls off substantially. We note that the longer the pre-treatment period, the more difficult it is to find close matches.

FIGURE 4 here

A challenge in matching citation histories is that some studies have only a few citations, and others have hundreds. FIGURE 4 reports the distribution of citations for the treated studies at the time their respective replications were published. The median number of citations is 23, with a minimum of 0 and a maximum of 2239 citations. As with longer pre-treatment periods, it is harder to find close matches for studies with many citations. Even with our large pool of potential control studies (112,000), it was difficult to get perfect matches. Only 53 of the 204 treated articles had a perfect match. As a result, if we want a larger pool of matched studies, we have to loosen the criterion for matching controls to treated.

Our approach is to use a "sliding scale" matching criterion. For a treated study with just a few citations at the time the replication 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. Accordingly, we develop three different matching criteria: *PCT=0%, PCT=10%,* and *PCT=20%.* Define $TotOrigCites$ as the total number of citations for the treated study up to the year the replication was published. The corresponding matching criteria are functions of $TotOrigCites$, with $TotOrigCites$ being multiplied by 0%, 10%, and 20%, respectively. TABLE 2 shows the maximum $TotAbsDiff$ allowed for matching treated with controls for different values of $TotOrigCites$ and PCT.

TABLE 2 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 over the entire 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 and, hence, we find more matches.

When $PCT = 10\%$, the matching threshold increases with $TotOrigCites$. Consider a treated and potential control that share a three-year citation history, where the treated study has a total of 20 citations. According to TABLE 2, in order to be a successful match, the total, year-by-year, absolute difference cannot be more than 3. For example, if the treated's citations were (5,7,8), potential controls with citations (4,7,9) and (5,5,9) would satisfy the matching criterion, but one with citations (4,6,10) would not. If, instead, the original study had 200 citations, $TotAbsDiff$ could differ by up to 21. For $PCT = 20\%$, the threshold values are slightly less than twice as large compared to $PCT = 10\%$.

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. The subsequent analysis focuses on studies that have at least 3 years of post-replication citation data and for which K = 3 to 8. However, as a robustness check, we also report results for studies having at least 5 years of post-replication citation data.

TABLE 3 here

TABLE 3 displays the number of treated and controls for different matching criteria and different lengths of post-replication periods. For example, when *PCT = 10%* and there are three or more years of post-replication citation data available, our sample consists of 103 treated studies and 7,552 matched controls. Requiring closer matches produces fewer treated and controls. Likewise, requiring a longer post-replication period (5-year versus 3-year) also results in fewer treated and controls.

TABLE 4 here

TABLE 4 reports 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 come 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 5 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 5. The first column reports the distribution 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%). Note that the respective quantile values are all smaller. For example, for *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 (compare to 8, 23, and 54.5).

Note also that the quantile values for the controls are less than the treated. This illustrates that treated studies with fewer citations have more controls matched to them. Thus, if all the controls are included as individual observations our results will represent the behaviour of studies with disproportionately few citations. This might be different from the average effect across all the studies in the respective sample. To mitigate this, when calculating $Diff_{i.Post}$ (see Equation 7), we take the difference in citations between the treated article and the *average* of the control articles that are matched to it.

4. Empirical Framework

As presented above, our identification strategy leads to the following regression equation,

(8) $$Diff_{it} = \beta_0 + \beta_1 Negative_i + \varepsilon_{it} ,$$

where $i \in PCT$, and PCT indicates which of three samples we are using, $PCT = 0\%,\ 10\%,\ 20\%$. We estimate separate regressions for each year of a seven-year period, $t = -3, -2, -1, 0, 1, 2, 3$; encompassing the three years before the replication was published, the year the replication was published, and the three years after the replication was published. β_1 measures the treatment effect. In words, β_1 measures the difference in citations at time *t* between treated and controls for treated studies with negative replications minus the difference in citations at time *t* between treated and controls for treated studies with positive/mixed replications. If negative replications adversely affect a study's citations, $\beta_1 < 0$ for $t > 0$.

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

(9) $Diff_{it} = \sum_{t=-3}^{3} \beta_{0t} \times T(t)_{it} + \sum_{t=-3}^{3} \beta_{1t} Negative_i \times T(t)_{it} + \varepsilon_{it}$,

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

With Equation (9) we can test for an overall, post-replication treatment effect by testing the null hypothesis:

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

Equation (9) is also useful for testing for an overall, pre-replication "treatment effect" by testing the null hypothesis:

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

Specifically, it allows us to test whether the matching process during the pre-treatment period produced systematically different citation outcomes for studies depending on whether they were successfully or unsuccessfully replicated. If the matched controls are able to equally "predict" citations for studies with negative and positive/mixed replications during the pre-

treatment period, that provides confidence that they should predict equally well during the post-treatment period.

A final issue arises because some of the control studies are used for more than one replicated 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 error independence. To address this problem, we bootstrap the standard errors.

5. Results

TABLE 6 reports the results of (i) estimating β_1 in Equations (8) and (9) and (ii) testing the hypotheses in Equations (10) and (11). Rows (1)-(3) provide a test of the effectiveness of our matching procedure. If matched controls do an equally good job of predicting citations for negative and positive/mixed replications, then we shouldn't see any differences between them in the pre-treatment period. As a result, β_1 should be indistinguishable from zero. Alternatively, if β_1 is significantly different from zero, this suggests that post-treatment differences might simply be carryovers from the pre-treatment period.

TABLE 6 here

The estimates in TABLE 6 are consistent with $\beta_1 = 0$ in the pre-treatment period.. All of the estimated coefficients are small in size. For example, when *PCT = 0%* and t = -3, we estimate a mean difference of 0.061 citations between studies with negative replications and studies with positive/mixed replications. Of the nine estimated coefficients in the first three rows, all are statistically insignificant.

Row (8) presents the results of a test of an overall, pre-treatment, citation difference between negative and positive/mixed replications. The cumulative sum of $\hat{\beta}_{1t}$ over the pre-treatment period is 0.091 citations for the *PCT = 0%* sample, and 0.541 and 0.712 citations for the more loosely matched, *PCT = 10%* and *20%* samples. In all three cases, we fail to reject

the hypothesis that the sum of the estimated differences is equal to zero. These results are consistent with the pre-treatment matching procedure performing equally well for studies with negative and positive/mixed replications. However, the fact that the pre-treatment difference is substantially smaller for the *PCT = 0%* sample indicates that the controls do a particularly good job of predicting equally well for this sample. As a result, the associated estimates should be the most reliable.

Rows (5)-(7) report estimates of the annual effect of negative replications on citations in the years immediately following publication of the replication. For the most closely matched sample (*PCT = 0%*), we estimate that studies with negative replications received 2.483 more citations than studies with positive replications in the year following publication of the replication (*t=1*). While the estimate is not statistically significant at the 5-percent level, it is opposite in sign of what would be expected if there was a citation penalty to failed replications.

Likewise for the second and third years after publication (*t=2,3*): These are smaller in size, 1.117 and 0.429 citations, respectively, but still positive. For the cumulative three-year period following replication, $\sum_{t=1}^{3} \hat{\beta}_{1t} = 4.028$, though it is, again, insignificant at the 5 percent level. On the other hand, the results for the more loosely matched, though larger samples, show a cumulative penalty for negative replications. For the *PCT = 10%* and *20%* samples, $\sum_{t=1}^{3} \hat{\beta}_{1t}$ equals -1.458 and -2.080 citations over the three-year, post-replication period. Every one of the estimates in TABLE 6 are statistically insignificant. This raises the issue of statistical power. We address this further below.

The samples underlying TABLE 6 were constructed based on having at least 3 years of post-replication observations. We can extend the post-replication period to 5 years, but at the cost of smaller samples. For example, when we require replicated studies to have at least 3 years of post-treatment observations, there are 74, 103, and 142 replicated studies that meet our criteria for the *PCT = 0%, 10%,* and *20%* samples. When using a post-treatment period of

5 years, there are 55, 79, and 112 replicated studies that meet our criteria. TABLE 7 repeats the analysis of TABLE 6 to see how extending the post-treatment period affects our results.

TABLE 7 here

The results are similar to those of TABLE 6: Specifically, (i) we cannot reject the hypothesis of no differences during the pre-treatment period; and (ii) the cumulative estimates of the effect of negative replications on citations is positive for the *PCT = 0%* sample, negative for the *PCT = 10%* and *20%* samples, and (iii) the estimated citation effects in the individual year regressions (t = -3, -2, …, 5) are everywhere insignificant. The main difference is that the estimated, cumulative, 5-year, post-treatment citation effect is relatively large (9.142) and statistically significant at the 1 percent level, though the sign goes against what would be expected if there was a citation penalty to failed replications.

We now address the issue of statistical power. Typical power analyses are constructed around the probability of rejecting the null hypothesis that the effect of interest equals zero. In our case, we are not very interested in determining whether negative replications produce a null effect. Rather, we want to know how likely it is that we would observe the estimates in TABLES 6 and 7 if there was, in fact, a citation penalty from negative replications. This is related to statistical power because both depend in part on the respective standard errors. If effects are estimated imprecisely, then small citation penalties, and even positive citation effects, will be observed relatively frequently even if the true citation penalty isn't small. TABLE 8 addresses this issue.

TABLE 8 here

Suppose the true, 3-year, cumulative citation effect associated with a negative replication versus a positive/mixed replication was either -3 citations (i.e., a citation penalty of one citation a year) or -6 citations (two a year). These seem to be the kinds of effect sizes one might reasonably expect given that the average number of citations for the treated studies at

the time the replications were published was 7.8 (*PCT=0%* sample*),* 19.5 (*PCT=10%* sample*),* and 27.8 (*PCT=20%* sample).

The table asks this question: What is the probability of estimating an effect size that is equal to or greater than the values in the left-hand column ($\widehat{\boldsymbol{ES}}$). For example, given the standard errors corresponding to the estimates in the *PCT=0%* regression of TABLE 6, if the true effect was -3, then the probability of estimating a cumulative, 3-year citation effect greater than or equal to -2 would be 32.6%. The probability of estimating a cumulative effect greater than or equal to -1 would be 18.4%. The associated probabilities for the *PCT=10%* and *PCT=20%* regressions are roughly similar. In contrast, if the true effect was -6, then the probabilities of estimating citations effects greater than -2 and -1 would 3.6% and 1.2%, respectively.

We can use TABLE 8 to gauge the likelihood of observing the estimates on Row (11) in TABLE 6. For the *PCT = 0%* regression, the estimated, cumulative 3-year effect is $\sum_{t=1}^{3}\hat{\beta}_{1t} = 4.028$. If the true cumulative, 3-year effect is -3 (-6), the likelihood of estimating an effect that is greater than or equal to this value is less than 0.1% (0.001%). Thus, for the "closest fit" sample we can easily reject the hypothesis of even a relatively small citation penalty of 1 citation per year.

For the *PCT = 10%* regression, $\sum_{t=1}^{3}\hat{\beta}_{1t} = -1.458$. If the true cumulative, 3-year effect were -3, there would be a 33.2% probability of observing an effect ≥ -1.458. On the other hand, if the true effect was -6, there would only be 10.0% probability of observing an effect this large or larger. Finally, using the "loosest fit" sample of *PCT = 20%,* if the true effect was -3, the probability of observing an estimated effect of -2.080 would be 37.9%. If the true effect was -6, this would fall to 9.4%. Similar results obtain if we use the estimates from TABLE 7, though with the smaller sample sizes come larger standard errors. Even so, using the *PCT = 0%* sample, we again reject the hypothesis of even a relatively small citation penalty.

In summary, when we use the strictest matching criterion of *PCT = 0%*, we find no evidence of a citation penalty. Indeed, we find evidence that studies that have negative replications are actually cited more in the post-treatment period than studies with positive/mixed replications. Further, we can strongly reject the hypothesis of even a relatively small citation penalty of a 1-citation difference between negative and positive/mixed studies per year. When we loosen the matching criteria to *PCT = 10%* and *PCT = 20%*, we find evidence of a small citation penalty, but even here our estimates indicate a citation penalty of less than 1 citation, and the estimates are not statistically significant from zero.

6. Conclusion

This study examines the effect of negative replications on the citation rates of replicated studies. We study a set of 204 replicated studies and investigate whether there is a citation penalty associated with negative replications. To do that, we employ an estimation strategy that matches replicated studies with controls, with the matched controls being used to "predict" the counterfactual citation performance of replicated studies.

Our study makes three contributions. First, we explain why there is a need for our study. While previous studies have estimated the effect of negative replications on citations (Serra-Garcia & Gneezy, 2021; Schafmeister, 2021; and von Hippel, 2022), these studies have relied on Open Science Collaboration (2015). Unfortunately, OSC is far from ideal for addressing this research question.

Second, in contrast to previous research, we develop a matched, difference-in-differences (DID) estimator that does not assume parallel trends (PT). Serra-Garcia & Gneezy (2021) demonstrate that studies that fail to replicate have different trends prior to replication than studies that successfully replicate. We show how assuming PT biases estimation of the citation effect of negative replications. Our estimation procedure avoids this bias.

Lastly, we present a variety of estimates of the citation effect of negative replications. Our preferred estimates indicate that studies that fail to replicate receive more citations than studies that have positive or mixed replications. Our less preferred estimates, based on looser matching criteria, find evidence of a citation penalty for negative replications, but the estimated effects are small and statistically insignificant.

Our analysis leaves us with questions: If replications currently do not play a self-correcting role in economics, can anything be done to enhance their impact? If not, what is their scientific benefit? Should the economics discipline allocate more, or less, resources in replications? And why? These are uncomfortable, but necessary, questions that should be addressed in future research.

References

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.

Ankel-Peters, J., Fiala, N., & Neubauer, F. (2023). Is economics self-correcting? Replications in the American Economic Review (No. 1005). Ruhr Economic Papers.

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.

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.

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.

Duflo, E., & Hoynes, H. (2018). Report of the Search Committee to Appoint a Data Editor for the AEA. *AEA Papers and Proceedings*, 108, 745–745. https://www.jstor.org/stable/26452832.

Ham, D. W., & Miratrix, L. (2022). Benefits and costs of matching prior to a difference in differences analysis when parallel trends does not hold. arXiv preprint arXiv:2205.08644.

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.

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

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.

Reed, W.R. (2023, April 5). Is science self-correcting? Evidence from 5 recent papers on the effect of replications on citations. *The Replication Network*. https://replicationnetwork.com/2023/04/05/reed-is-science-self-correcting-evidence-from-5-recent-papers-on-the-effect-of-replications-on-citations/

Reed, W.R. (2023, April 16). More on self-correcting science and replications: A critical review. *The Replication Network*. https://replicationnetwork.com/2023/04/16/reed-more-on-self-correcting-science-and-replications-a-critical-review/

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

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

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.

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
Threshold Values for $TotAbsDiff$ for Various Combinations of $TotOrigCites$ 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 indicate the maximum allowable value of the sum of annual, absolute differences between treated and control citations over the length of the pre-treatment, citation history period. The threshold value is determined using the following formula: $TotAbsDiff_K \leq ceil(PCT \times TotOrigCites_K + 0.001)$, where $TotAbsDiff_K = \sum_{k=1}^{K-1} \left| Citations_{T-k}^{Control} - Citations_{T-k}^{Treated} \right|$, $TotOrigCites_K = \sum_{k=1}^{K-1} Citations_{T-k}^{Treated}$, and K is the number of years between the publication of the originl study and its replication.

TABLE 3
Number of Treated and Controls for Different Matching Criteria and Length of Post-Treatment Period

	Matching Criteria		
	PCT = 0%	***PCT = 10%***	***PCT = 20%***
	3-year Post-Treatment Period		
Treated	74	103	142
Controls	7,044	7,552	11,202
	5-Year Post-Treatment Period		
Treated	55	79	112
Controls	6,171	6,587	8,689

NOTE: The values in the table report the number of matched treated and controls for the samples corresponding to each of the three matching criteria (*PCT=0%, PCT=10%, PCT=20%*) and two different lengths of post-treatment periods (3-year and 5-year).

TABLE 4
Distribution of $TotAbsDiff$ 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 difference in citations between treated and controls for the three samples (*PCT* = 0%, 10%, and 30%), where $TotAbsDiff_K = \sum_{k=1}^{K-1} \left| Citations_{T-k}^{Control} - Citations_{T-k}^{Treated} \right|$ and *K* is the number of years between the publication of the treated and its replication.

TABLE 5
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: the full sample of 204 treated studies, and the three analysis samples defined by (*PCT=0%, PCT=10%, PCT=20%*). The difference between the Max values for *PCT=0%* sample is greater than 1 because these citation numbers include citations received in the year the paper was published

TABLE 6
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***	***0.061*** [0.177] (0.729)	***0.422*** [0.413] (0.307)	***0.315*** [0.297] (0.290)
(2)	***t = -2***	***-0.010*** [0.067] (0.883)	***0.154*** [0.210] (0.463)	***0.084*** [0.286] (0.769)
(3)	***t = -1***	***0.040*** [0.052] (0.447)	***-0.035*** [0.202] (0.863)	***0.313*** [0.360] (0.384)
(4)	***t = 0***	***2.210***** [0.991] (0.026)	***1.799*** [1.142] (0.115)	***1.444*** [1.073] (0.178)
(5)	***t = 1***	***2.483**** [1.490] (0.096)	***0.341*** [1.488] (0.819)	***0.559*** [1.297] (0.666)
(6)	***t = 2***	***1.117*** [1.002] (0.265)	***-0.231*** [1.688] (0.891)	***-0.353*** [1.675] (0.833)
(7)	***t = 3***	***0.429*** [1.150] (0.709)	***-1.568*** [2.218] (0.480)	***-2.286*** [2.050] (0.265)
(8)	**Test of overall pre-replication effect:**	$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{0.091}$ s.e. = 0.196 p = 0.642	$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{0.541}$ s.e. = 0.420 p = 0.197	$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{0.712}$ s.e. = 0.582 p = 0.221
(9)	**Test of overall post-replication effect:**	$\sum_{t=1}^{3} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{4.028}^{*}$ s.e. = 2.220 p = 0.070	$\sum_{t=1}^{3} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{-1.458}$ s.e. = 3.549 p = 0.681	$\sum_{t=1}^{3} \widehat{\boldsymbol{\beta}}_{1t} = \mathbf{-2.080}$ s.e. = 2.981 p = 0.485
(10)	**Mean Total Cites (t=0)**	7.8	19.5	27.8
(11)	**Median Total Cites (t=0)**	4.8	8.4	14.8
(12)	**Observations**	74	103	142

NOTE: The table reports the results of estimating β_1 in Equation (8) for three different samples (*PCT=0%, PCT=10%, PCT=20%*). 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 the average of their matched, non-replicated control studies. Estimates in brackets are bootstrapped standard errors. Estimates in parentheses are z-based *p*-values.

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 all years and testing whether the cumulative effect during the pre- and post-treatment periods, respectively, equals zero. 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 7
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***	***-0.048*** [0.165] (0.770)	***0.397*** [0.455] (0.383)	***0.275*** [0.405] (0.497)
(2)	***t = -2***	***0.002*** [0.080] (0.981)	***0.240*** [0.249] (0.334)	***0.233*** [0.351] (0.507)
(3)	***t = -1***	***0.036*** [0.057] (0.526)	***0.053*** [0.240] (0.826)	***0.449*** [0.420] (0.285)
(4)	***t = 0***	***1.919**** [1.036] (0.064)	***2.026*** [1.341] (0.131)	***1.679*** [1.088] (0.123)
(5)	***t = 1***	***2.402*** [1.682] (0.153)	***0.611*** [2.046] (0.765)	***0.781*** [1.758] (0.657)
(6)	***t = 2***	***0.922*** [1.251] (0.461)	***0.193*** [1.902] (0.919)	***-0.335*** [2.032] (0.869)
(7)	***t = 3***	***1.317*** [0.858] (0.125)	***-0.608*** [3.360] (0.856)	***-1.599*** [2.515] (0.525)
(8)	***t = 4***	***1.313*** [1.359] (0.334)	***-0.354*** [3.474] (0.919)	***-0.431*** [2.972] (0.885)
(9)	***t = 5***	***3.188*** [2.133] (0.135)	***-0.464*** [5.832] (0.937)	***-0.023*** [4.309] (0.996)
(10)	**Test of overall pre-replication effect:**	$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = -0.010$ s.e. = 0.181 p = 0.954	$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = 0.690$ s.e. = 0.647 p = 0.286	$\sum_{t=-3}^{-1} \widehat{\boldsymbol{\beta}}_{1t} = 0.958$ s.e. = 0.675 p = 0.156
(11)	**Test of overall post-replication effect:**	$\sum_{t=1}^{5} \widehat{\boldsymbol{\beta}}_{1t} = 9.142$ *** s.e. = 3.491 p = 0.009	$\sum_{t=1}^{5} \widehat{\boldsymbol{\beta}}_{1t} = -0.622$ s.e. = 6.813 p = 0.927	$\sum_{t=1}^{5} \widehat{\boldsymbol{\beta}}_{1t} = -1.607$ s.e. = 6.541 p = 0.806

		Matching Criteria		
		PCT = 0%	***PCT = 10%***	***PCT = 20%***
(12)	**Mean Total Cites (t=0)**	7.3	21.0	30.1
(13)	**Median Total Cites (t=0)**	4.1	8.0	18.7
(15)	**Observations**	55	79	112

NOTE: This table reports the same information as TABLE 6, 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 6).

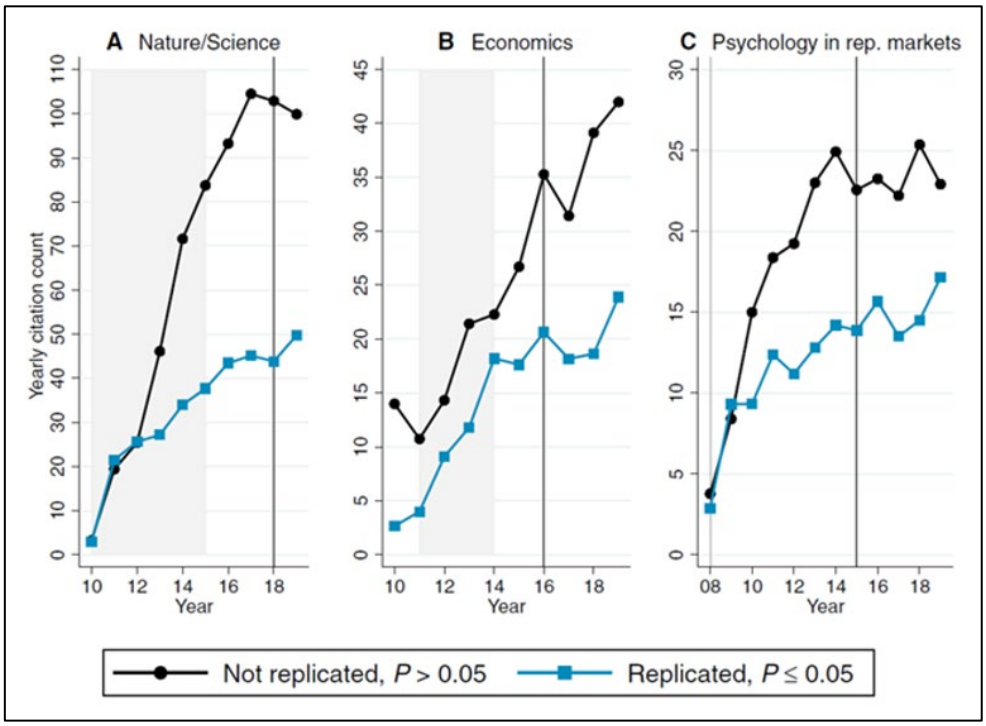
*, **, and *** indicate statistical significance at the 10-, 5-, and 1-percent levels.

TABLE 8
Conditional Probabilities of Observing Selected Estimated Effects ($\widehat{ES}$)

$\widehat{ES}$	*TRUE EFFECT SIZE = -3 citations over 3 years*			*TRUE EFFECT SIZE = -6 citations over 3 years*		
	$Prob\left(\sum_{t=1}^{3}\widehat{\beta}_{1t} \geq ES \mid \sum_{t=1}^{3}\beta_{1t} = -3\right)$			$Prob\left(\sum_{t=1}^{3}\widehat{\beta}_{1t} \geq ES \mid \sum_{t=1}^{3}\beta_{1t} = -6\right)$		
	PCT = 0%	***PCT = 10%***	***PCT = 20%***	***PCT = 0%***	***PCT = 10%***	***PCT = 20%***
-3	50.0%	50.0%	50.0%	8.8%	19.9%	15.7%
-2	32.6%	38.9%	36.9%	3.6%	13.0%	9.0%
-1	18.4%	28.7%	25.1%	1.2%	7.9%	4.7%
0	8.8%	19.9%	15.7%	0.3%	4.5%	2.2%
1	3.6%	13.0%	9.0%	0.1%	2.4%	0.9%
2	1.2%	7.9%	4.7%	0.0%	1.2%	0.4%
3	0.3%	4.5%	2.2%	0.0%	0.6%	0.1%
4	0.1%	2.4%	0.9%	0.0%	0.2%	0.0%
5	0.0%	1.2%	0.4%	0.0%	0.1%	0.0%

NOTE: Values in the table are probabilities based on the estimates in TABLE 6. They can be interpreted as the probability of observing a given estimated effect size, $\widehat{ES} = \sum_{t=1}^{3} \hat{\beta}_{1t}$, given that the true effect size is either -3 or -6 citations over the three-year, post-treatment period. Standard errors while not reported in TABLE 6, can be obtained from $\left(\sum_{t=1}^{3} \hat{\beta}_{1t}/t\right)$. Probabilities are calculated assuming a z-distribution given that the underlying standard errors are bootstrapped.

FIGURE 1
Yearly Citation Count by Replicability

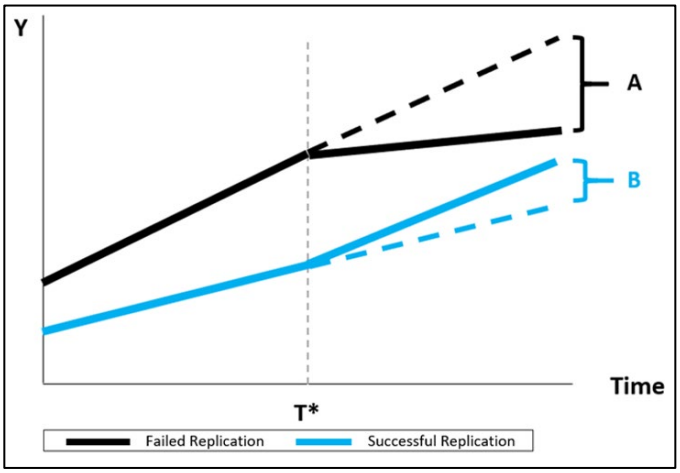


SOURCE: Serra-Garcia & Gneezy (2021, Figure 3)

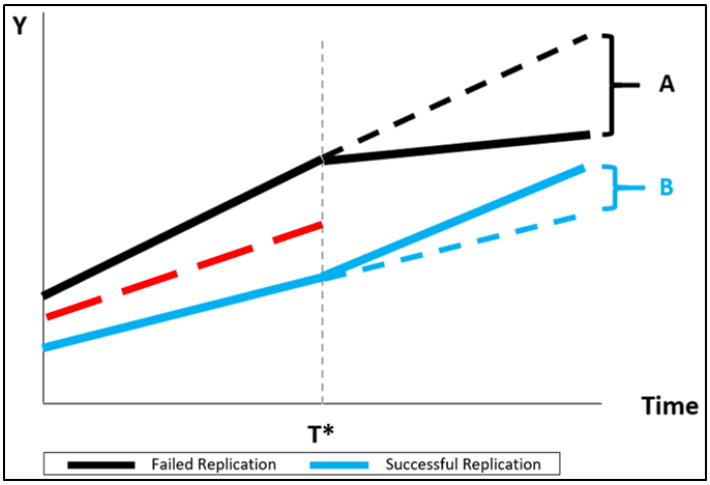
NOTE: The vertical axis reports "yearly citation count". The horizontal axis reports year, and the vertical line shows the year that OSC published their replication findings (2015). Of particular interest is Panel C, which plots citations for the OSC studies. There is a clear difference in the pre-treatment citation trends of studies that were not successfully replicated (black line) and those that were (blue line). Studies that failed to be replicated were cited at a higher rate during the pre-treatment period.

FIGURE 2
How Assuming Parallel Trends Biases Estimation of the Citation Penalty from Failed Replications

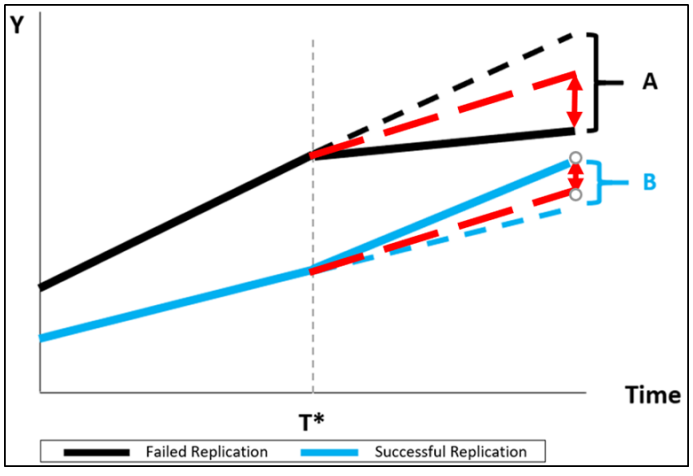
A. The Citation Effect of Failed versus Successful Replications



B. The Assumption of Parallel Trends

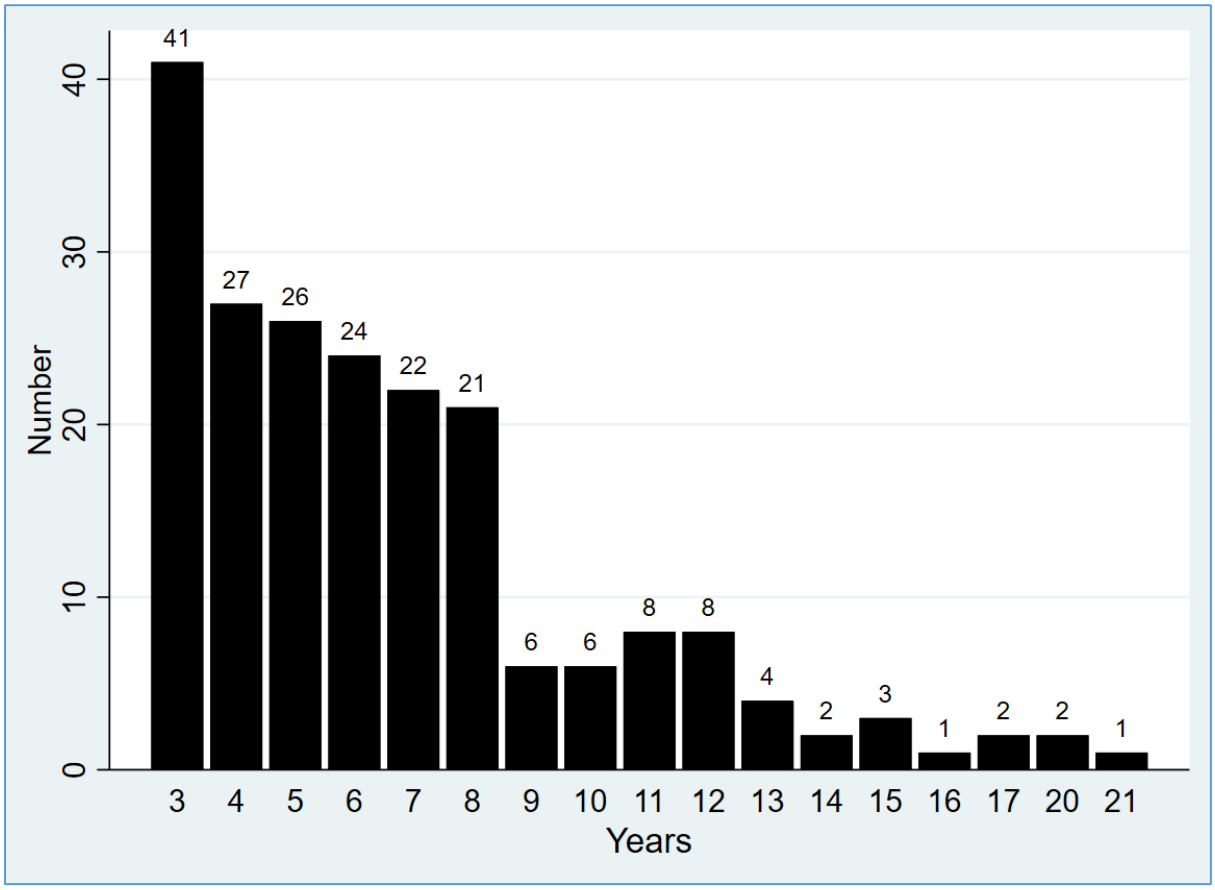


C. Assuming Parallel Trends Biases Estimate of the Treatment Effect



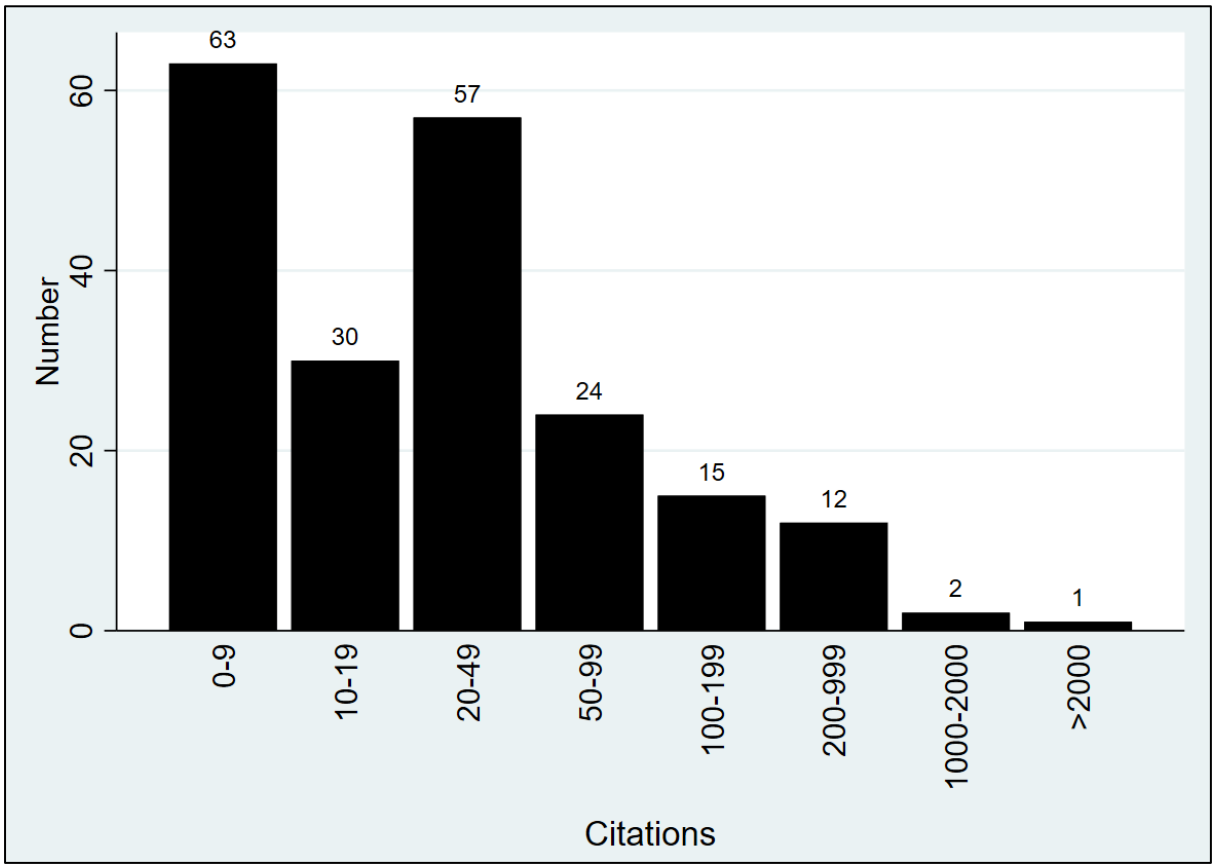
NOTE: The three panels highlight the bias that arises from estimating the citation effect associated with failed replications when one incorrectly assumes parallel trends. The vertical axis measures annual citations and the horizontal axis measures years, where T* indicates the year when the replication was published. The black (blue) lines identify citation trends for studies with failed (successful) replications. Solid lines indicate observed citations, dotted line indicate the estimated counterfactual had there been no replication. “A” measures the negative citation effect associated with a failed replication. “B” measures the positive citation effect associated with a successful replication. The total citation effect of a failed versus a successful replication is given by A-B. The red line indicates a parallel trend calculated as the average trend for studies with failed and successful replications. Extending this trend into the post-treatment period causes the researcher to underestimate the true effect. A more detailed discussion is given in the text.

FIGURE 3
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 4
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.