DEPARTMENT OF ECONOMICS AND FINANCE

SCHOOL OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Paving the Road for Replications: Experimental Results from an Online Research Bibliography

Tom Coupé
W. Robert Reed
Christian Zimmermann

WORKING PAPER

No. 5/2022

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

WORKING PAPER No. 5/2022

Paving the Road for Replications: Experimental Results from an Online Research Bibliography

Note: This paper is a revision of WP 2021-9

Tom Coupé[1†]
W. Robert Reed[1]
Christian Zimmermann[2]

February 2022

Abstract: Are users of a bibliographic database interested in learning about replications? Can they be induced to learn? To answer these questions, we performed an experiment at the online research bibliography, RePEc (Research Papers in Economics). RePEc is the main research bibliography for pre-prints and published papers in economics. Using randomized stratification, we allocated 324 replications and their corresponding original studies to clusters. We then drew from those clusters to construct treatment and control groups. Brightly colored tabs were added to the relevant webpages to alert visitors to the existence of a replication study. We then monitored traffic over three phases lasting several months: a) no treatment, b) treatment on one group, c) treatment on both groups. Our estimates indicate that this intervention generated an average click-through-rate (CTR) of 1.6%, resulting in a 13% increase in the visits to replication webpages, though only the former estimate was statistically significant.

Keywords: Replications, RePEc, Experiment, Online Research Bibliography, Webpages, Click-throughs

JEL Classifications: A11, B41, Z00

Acknowledgements: The views expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Federal Reserve System, the Board of Governors, or the regional Federal Reserve Banks. We acknowledge helpful feedback from Andrew Chang, James Druckman, Uri Gneezy, Jan Höffler, Nicholas Huntington-Klein, Miklós Koren, Joshua Krieger, Felix Schafmeister, Marta Serra-Garcia, and participants at Metascience 2021 and aimos 2021 Conferences.

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

[2] Federal Reserve Bank of St. Louis, USA

† Corresponding author: Tom Coupé. Email: tom.coupe@canterbury.ac.nz

I. Introduction

There is accumulating evidence that many published articles in scientific disciplines do not reproduce well. Some have called this a "replication crisis" (Pashler & Wagenmakers, 2012). Probably the best-known test of replicability is the Reproducibility Project: Psychology (RPP), the results of which were published in *Science* (Open Science Collaboration, 2015). RPP attempted to replicate 100 prominent experiments in psychology.[1] While statistically significant estimates were reported for 97% of the original experiments, only 36% of the estimates from the replicated experiments were significant.

While the field of psychology provides some of the most prominent examples of failure to replicate, examples are readily found in other disciplines, such as medicine (Ioannidis, 2005; Begley & Ellis, 2012), economics (Camerer et al., 2016; Duvendack et al., 2015, 2017), management (Hubbard, 2015), and nutrition science (Sorkin et al., 2016). To be fair, there is debate about how to interpret these findings, with some arguing that seemingly low rates of replication do not necessarily indicate scientific failings (Gilbert et al., 2016). Nevertheless, there is a general consensus that replication is good for science and should be done more frequently (Fidler & Wilcox, 2018).

One of the reasons given for why journals don't publish more replications is that replications do not generate enough citations to make them sufficiently attractive for researchers to invest in (Duvendack et al, 2017). This claim, while widely believed, is difficult to prove. There is evidence, however, that the market for replicable results is not well-functioning.

Focusing on the 100 studies included in the RPP, Yang et al. (2020) and Schafmeister (2021) estimated that studies that failed replication were subsequently cited no less than studies

[1] In this paper, we use replication and replicability interchangeably with reproduction and reproducibility (Gernsbacher, 2021).

that passed replication. Serra-Garcia & Gneezy (2021), using a larger sample that included studies in psychology, economics, and general science journals, reported that failed replications were actually cited more than studies that were successfully replicated.

Relatedly, previous research has shown that retracted and discredited studies continue to be cited after the original claims have been repudiated (Budd et al., 1998; Bornemann-Cimenti et al., 2016; Tatsioni et al., 2007; Bar-Ilan & Halevi, 2018; Candal-Pedreira, 2020; Schneider et al., 2020; Fernández et al., 2021; Hardwicke et al., 2021; Hsiao & Schneider, 2021; Piller, 2021). This suggests that the problem may be lack of information: Researchers who cite flawed studies may be unaware of subsequent research that undermines the original findings.

In this paper, we describe the results of an online experiment aimed at providing information about the existence of replication studies. Motivating this experiment is the conjecture that greater awareness of replications could lead to greater citation rates. This, in turn, could open the publication market for replication studies. Thus, as a first step in increasing replications, we need to identify effective strategies for increasing awareness.

Our experiment consisted of an intervention in the online research bibliography, RePEc (Research Papers in Economics). RePEc is the main research bibliography for both pre-prints and published papers in economics, providing links to over 2,400,000 full-text articles. We modified some of the pages in IDEAS, which is one of the most popular RePEc websites, by inserting a prominent link alerting researchers to the existence of a replication.

We collected information on how often people visited the original page, how often they clicked on the link to the replication, and how frequently they visited the replication page. In this way, we were able to assess the efficacy of our webpage modification to direct attention to replications. If the problem is lack of awareness of replication studies, interventions such as this could solve that problem.

Our interventions produced a modest but significant effect. We estimate that our webpage modifications generated an average click-through-rate (CTR) of 1.6%, resulting in a 13% increase in the visits to replication webpages. However, only the former estimate was statistically significant. In section II of this paper, we explain our intervention. In section III, we discuss results. Section IV concludes.

II. The Treatment

Overall description. We obtained replication studies from two websites that collect data on published replications in economics, The Replication Network and ReplicationWiki (Höffler, 2017).[2] With the assistance of two research assistants, we then identified the RePEc handles for (i) the replication paper and (ii) the original paper replicated by the replication study. We 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 324 (original study-replication) pairs that were not associated with any other replications or original papers.

As described in detail below, we used a matching procedure to group pairs into 67 clusters. Pairs within each cluster were closely matched. The pairs in each cluster were then randomly assigned to one of two groups. Both groups received treatments, but at different stages of the intervention. The treatment consisted of a modification to the RePEc webpage of an original study that alerted the reader to the existence of its replication. We then used LogEc to obtain information on the number of visits to the respective papers' webpages.[3]

Our experiment was conducted over eighteen months in 2020 and 2021 and consisted of three stages. Stage I was a six-month, pre-treatment period from February 2020 to July 2020. Page visit statistics from this period were used to create clusters of papers based on the number

[2] The Replication Network draws most of its replications from ReplicationWiki. Its list of replications differ primarily because (i) it only lists replications that have been published, and (ii) it uses a stricter definition of what constitutes a replication. Specifically, it only counts studies as replication where the main purpose of the study is to determine whether one or more results from a previously published study are correct (Duvendack & Reed, n.d.).

[3] LogEc tries to weed out traffic from anything that seems automated (that is, traffic from webscrapers and bots such as Googlebot, Bingbot, and Baidu) as well as repeat visits.

of visits to replication and original papers. Stage II lasted from August 2020 to February 2021. In this stage, webpages of studies in the first group were treated, while webpages of studies in the second group were left unchanged. We recorded visits and clicks on all the associated RePEc webpages, allowing us to compare online traffic for the two groups. Note that we omitted August in the subsequent analysis because treatments were not all applied at the same time, but were spread throughout the month. As a result, not all the studies in the first group received the treatment for the full month in August.

In Stage III, the webpages of studies in the second group were also treated. This period lasted from March 2021 to July 2021. We again recorded visits and clicks on all the associated RePEc webpages. As in Stage II, we excluded the first month of treatment (March) because treatments were introduced piecemeal over the month.[4]

Assignment of studies to groups. To assign the pairs of original and replication papers to the two groups, we applied a procedure of stratified randomization. Original study-replication pairs were scored based on (i) the 6-month count of visits to the original study and (ii) the 6-month count of visits to the replication study during the pre-treatment period. In designing our experiment, we followed the advice of Athey and Imbens (2016, pages 2-3):

> We recommend using experimental design rather than analysis to adjust for covariates differences in experiments. Specifically, we recommend researchers to stratify the population into small strata and then randomize within the strata …We argue that this approach is preferred to model-based analyses applied after the randomization to adjust for differences in covariates. However, there are limits on how small the strata should be: we do not recommend to go as far as pairing the units, because it complicates the analysis due to the fact that variances cannot be estimated within pairs, whereas they can within strata with at least two treated and two control units.

Accordingly, we followed an iterative procedure to match similar pairs. The first step used hierarchical clustering (Euclidean distance, Ward variance) on the overall sample to find

[4] We exclude from the click counts the clicks that come from visitors clicking on the back button, that is, visitors who return immediately to the page they started from.

clusters that contained at least 4 pairs that had a maximum distance of zero (i.e., exact matches). We removed these clusters with exact matches from the overall sample and repeated hierarchical clustering on the remaining pairs, increasing the maximum distance between clusters in a stepwise manner until all study pairs were allocated to a cluster. This led to a total of 67 clusters, with an average of 5 pairs per cluster.[5]

Within each cluster, we randomly allocated half of the pairs to the first group and half to the second group. If a cluster had an uneven number of pairs, the odd pair was randomly assigned to a group. This resulted in 161 pairs assigned to Group 1 and 163 pairs to Group 2.[6] Webpages of Group 1 studies received the treatment during the month of August 2020. Webpages of Group 2 studies were unchanged until February 2021, but then they too received the treatment starting in March. As reported in TABLE 1 below, this assignment procedure resulted in a well-balanced sample as measured by pre-treatment page visits.

TABLE 1:
Balance in Pre-Treatment Page Visits between Groups 1 and 2

	Group 1	Group 2
Mean Monthly Page Visits of Replication Papers	10.2	10.0
Mean Monthly Page Visits of Original Papers	42.0	44.0
Median Monthly Page Visits of Replication Papers	5	5
Median Monthly Page Visits of Original Papers	17	17

As seen in the table, original studies received substantially more page visits than replications. The median monthly number of page visits to original studies in the 6-month period, February 2020 to July 2020 was 17, compared with only 5 page visits for their

[5] We also experimented with a purely random allocation and a k-nearest neighbor algorithm, but in our sample, the iterative hierarchical clustering method showed better matching results – see the annex. See also http://jmonlong.github.io/Hippocamplus/2018/06/09/cluster-same-size/

[6] One pair has been excluded from the analysis because LogEc visitor statistics are no longer available for the original paper in the pair.

replications. While the numbers in the table show close agreement in mean and median visits between Groups 1 and 2, a better assessment of balance is provided in FIGURE 1. The two panels in this figure display histograms of number of page visits for original (Panel A) and replication studies (Panel B). The two groups show very similar distributions of visits to both the original and replication studies' webpages.

Description of treatment. The treatment consisted of two changes to a paper's RePEc webpage. This is best understood via an illustration. Panel A of FIGURE 2 depicts a typical RePEc webpage. The page opens up to the first tab, which reports author and abstract information. Active tabs are indicated by bright blue coloring. Note that the second tab is colored bright red to alert the webpage visitor to the option of downloading the paper.

Panel B of FIGURE 2 depicts how the RePEc page was altered for papers in the treatment group. First, an asterisk was added to the title to indicate that *"* This paper has been replicated.*" Second, a bright yellow tab titled "Replication" was added, providing an additional option to the webpage visitor. Clicking on this Replication tab took the visitor to the RePEc webpage with bibliographic information about and a link to the RePEc page of the replication study (see FIGURE 3). We count how many people clicked on the link to the RePEc page of the replication paper.[7]

A similar modification was made to the webpages of replication papers. To those, a note was added that the paper replicated an original paper, and a bright yellow tab with a link to the original paper was added.

III. Results

Measure #1: Click-throughs as a percent of total page views (original study). For our first measure of treatment effect, we focused on monthly click-throughs from the webpages of the

[7] Such clicks are unlikely to be caused by robots. Note we did find some instances were LogEc did not record any hits even though clicks were recorded.

original studies to the webpages of the replications. The vertical axis of FIGURE 4 reports monthly click-throughs as a percent of total page views for each group (CTRs). The horizontal axis reports the calendar month. The three stages of the experiment are identified in the figure.

In Stage I, the pre-treatment stage (February 2020 to July 2020), page visitors to the original studies occasionally, but rarely, clicked through to the replication study. This would happen only if a reader found the replication study listed on the citations of the original study and clicked on the live link there.

Stage II ran from September 2020 until February 2021 and covered the first of our two interventions. Visitors to webpages of original studies in Group 1 were presented with an option to click on the link to the replication study. The black vertical bars show the CTRs for this group. Webpages for original studies in Group 2 were left unaltered. The associated CTRs for this group (if any) are represented by downward-pointing grey bars.

While the figure shows a clear treatment effect, the size of the effect is relatively modest in absolute terms. CTRs ranged from 0.8% to 2.6% per month for the treated pages. In total over this period, treated original papers were visited 6793 times from other sources, but the links to the replication pages were clicked only 107 times, giving an average monthly CTR of about 1.6%.[8]

Stage III (April 2021 to July 2021) covered the period of the second intervention. During this stage webpages for original studies in Group 2 also received the treatment. Similar to Stage II, there was an immediate increase in the number of click-throughs for the newly treated studies. However, also similar to Stage II, the size of the effect was relatively small in absolute terms. CTRs for the newly treated studies ranged between 1% and 2.1%.

[8] By ”from other sources” we mean that we subtract, from the total views to pages of original papers, the views related to the clicks from the replication page. Or, vice versa, we subtract, from the total views to pages of replication papers, the views related to the clicks from the original page.

Total click-throughs were quite similar for both groups of treated studies in Stage III. Group 1 studies were visited 3932 times, with 42 clicks on the link to the replication study, producing an average monthly CTR of 1.1%. Group 2 studies received 3995 visits, with 63 of those visits leading to clicks on the links to the replication, for a CTR of 1.6%.

It should be noted that while these CTRs may be modest in absolute size, they are relatively large by industry standards. Industry CTRs for display ads average 0.35%, with rates above 0.5% quite rare (Wordstream, n.d.). Industry CTRs for search ads, which follow users' targeted search for products and services via keywords, are higher; but even here, the average CTR across all industries is approximately 1.9%.

Measure #2: Clicks-throughs as a percent of total page views (replication study). Since many more people visit original studies than replications (cf. TABLE 1), even a small percent of click-throughs could represent a sizable impact on total page views for replication studies. For this reason, FIGURE 5 reports click-throughs as a percent of total page views of replication studies from other sources. As expected, the treatment effects using this alternative measure are larger.

For Group 1 during Stage II, click-throughs ranged from 3.8% to 11.3% as a share of total visits to replication pages from other sources. The treated replications received a total of 107 visits via the links from the original studies, compared with a total of 1587 visits from other sources. Assuming no crowding out (i.e., visitors who would have visited the replication even if they had not seen the link on the original study's webpage), this measure of effect indicates that our intervention increased the number of visits to replication pages by 6.7%.

During Stage III (April to July 2021), Group 1 replication studies received a total of 42 visits via links from treated original studies, compared to a total of 1102 visits from other sources, for an overall treatment effect of 3.8%. The treatment effect for Group 2, the newly treated group, was 6.2% (63/1019).

Measure #3: Total page visits to replication pages. A final measure of treatment effect compares total page visits to replication pages for Groups 1 and 2 across all three stages of our experiment. The associated results are reported in FIGURE 6. During the pre-treatment period (Stage I), both groups received roughly the same number of page visits, with Group 1 receiving slightly (1%) more. In Stage II, webpages for replication studies in Group 1 received 13.8% more visits than Group 2. In Stage III, during which both groups received treatments, page visits at first rose for both groups, then declined. The fall in page visits in the last months of our experiment is a reminder that unobserved factors were also at play. This highlights the importance of the matching procedure we employed to identify the respective treatment effects.

Estimates of treatment effects. The previous analysis reported monthly averages for the two groups of studies across the three stages of our experiment. In this section we estimate cluster-level treatment effects on the two main outcomes of interest: click-throughs and visits to replication webpages. Our estimation exploits the matching process used to assign papers to clusters.

Our analysis focuses on Stage II. In Stage I, none of the papers in our study were treated. In Stage III, all the papers were treated. In Stage II, each cluster contained both treated and untreated papers, matched on the basis of page visits during the pre-treatment period. We take the difference in the monthly averages of either (i) click-throughs or (ii) visits to replication pages between treated and untreated papers in the same cluster. Our estimated treatment effect consists of the mean/median of these average differences. There were 67 clusters and 6 months of observations, producing a total of 402 observations. TABLE 2 reports the results.

The top panel of TABLE 2 reports differences in average monthly click-throughs between treated and untreated original studies in the same cluster. Rows 1) and 3) show mean and median differences of 0.111 and 0 click-throughs, respectively. As a point of comparison, mean and median monthly page visits are 7.1 and 3 (cf. Rows 5 and 6). With respect to the mean difference, our estimated treatment effect amounts to a CTR of 1.6% and is statistically significant at the 5 percent level. Maybe the CTR wasn't larger because many or most of the original studies weren't visited? Row 7) addresses this concern. In fact, over 80% of the treated original observations received at least one visit a month during Stage II.

Our experimental design also allowed us to conduct an ex-post test of balance. In Stage III, all the papers in our experiment were treated. We tested whether there were any differences in the two groups during this third stage. If assignment to Groups 1 and 2 was random and treatment effects were homogenous, we would expect to see no difference in click-throughs between these two groups. Rows 2) and 4) confirm that we cannot reject this hypothesis. The mean difference in click throughs for the two groups is negligible at -0.038 and insignificant at the 5% level. This provides further evidence supporting the validity of our random assignment of studies to groups.

The bottom panel of TABLE 2 repeats the previous analysis, this time focusing on page visits to replication studies. We estimate cluster-level differences in average monthly page visits for treated and untreated replication pages in Stage II. The associated mean and median differences are 0.235 and 0.13 visits, respectively. These compare to mean and median values of overall monthly visits of 1.76 and 1. With respect to mean values , this represents an increase in page visits of about 13%, though this estimate is not statistically significant. We note that approximately 45% of the monthly observations of replication pages received no visits at all. As before, the balance test finds no significant difference between the two groups in Stage III, when all pairs were treated.

IV. Conclusions

This study investigated the effects of an online experiment. We modified the webpages of an online research bibliography to encourage visitors to the webpages of original studies to visit the webpages of the replications of those studies. We found positive impacts from our intervention, though the practical size of the estimated effects was modest. Using cluster-level data, we estimate that our treatments of the original studies' webpages generated an average click-through-rate (CTR) of 1.6%, resulting in a 13% increase in the visits to replication webpages. However, only the first of these estimates was statistically significant at the 5 percent level. The estimated effects compare favourably to CTRs for industry ad campaigns.

Possible explanations for our findings. One possible reason for the modest effect of our treatment was that it was not sufficiently visible to online visitors of the original studies (cf. FIGURE 2). We could have chosen more obtrusive indicators. Indeed, we discussed the possibility of using pop-ups to get the attention of page visitors. However, it was decided this would cross the line from informative to annoying. As a result, we chose not to implement more forceful interventions, as the intent of this experiment was to test an idea that, if successful, could be implemented on other research platforms.

Another possible reason is that visitors may have already been familiar with the replication studies and therefore had no need to click through to the replication study. This hypothesis, however, does not address why overall visits to originals are so much higher than visits to replications (cf. TABLE 1).

A third possibility is the obvious one: that researchers were not that interested in replications. Although they clearly showed some interest in the subject by clicking to the page of the original study, they may have not had enough interest to view the replication. This line of inquiry leads to many further conjectures. Online readers of the original studies may have merely wanted to get the "lay of the land" with respect to prominent papers in the field. They may have already decided to cite the paper in their own research, but wanted to be more familiar with it before doing so. Alternatively, they may not have had confidence, for any number of reasons, that there was much to learn from the replications. A last possibility, uncomfortable as it is to say, is that researchers may not have cared whether the published findings from the original papers were accurate.

Implications. Assuming our findings are externally valid for other research platforms, they suggest that online interventions to alert readers of replication studies -- while they may provide moderate benefits -- are unlikely to have major impacts. If so, what then? It really depends on the reasons why our interventions did not have larger effects.

If researchers are already familiar with the existence of replication studies, there is no reason to try to increase awareness. If researchers lack confidence in replications, then efforts to improve the quality/reliability of replications could be helpful. If most researchers don't care whether empirical findings from other studies are true, then changing academic incentives to induce them to care could be effective.

This last effort could be achieved, for example, if leading professional societies published journals solely devoted to replications or required researchers to publish replications to receive tenure or other promotions. Regardless of the underlying factors, one implication of this study is that it highlights the importance of finding out why researchers are not more motivated to learn about replications. We hope this study stimulates initiatives in this direction.

Limitations. Our sample size was ultimately limited by the number of replication studies we could identify: 324 matched (original studies/replication) pairs may seem like a large number, but the effective size was somewhat diminished by the fact that some of the webpages of the original studies were not heavily trafficked. As indicated in TABLE 2, about 17% of the originals did not get any visit in a specific month. Another limitation of the current study concerns external validity. It is uncertain how the results from this study would extrapolate to other online research repositories such as arXiv, bioRxiv, medRxiv, and PsyArXiv.

References

Athey, Susan and Guido W. Imbens (2016). The Econometrics of Randomized Experiments. ArXiv 1607.00698, https://arxiv.org/pdf/1607.00698.pdf

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

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

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

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391), 531-533.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.

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

Duvendack, M., Palmer-Jones, R. W., & Reed, W. R. (2015). Replications in economics: A progress report. *Econ Journal Watch*, 12(2), 164-191.

Duvendack, M., Palmer-Jones, R., & Reed, W. R. (2017). What is meant by" replication" and why does it encounter resistance in economics? *American Economic Review*, 107(5), 46-51.

Duvendack, M. & Reed, W.R. (n.d.). Why replications? https://replicationnetwork.com/why-replications/.

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

Fidler, F. & Wilcox, J. (2018). Reproducibility of scientific results. In Zalta, E. N. (Eds.) *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab Center for the Study of Language and Information Stanford University.

Gernsbacher, M.A. (2021). Replication (and replicability) versus reproduction (and reproducibility). https://online225.psych.wisc.edu/wp-content/uploads/225-Master/225-UnitPages/Unit-04/Replication_vs_Reproduction.pdf

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science,* 351(6277), 1037-1037.

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

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

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

Hubbard, R. (2015). *Corrupt research: The case for reconceptualizing empirical management and social science*. Sage Publications.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2), 218-228.

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528-530.

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

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

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

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

Sorkin, B. C., Kuszak, A. J., Williamson, J. S., Hopp, D. C., & Betz, J. M. (2016). The challenge of reproducibility and accuracy in nutrition research: resources and pitfalls. *Advances in Nutrition*, 7(2), 383-389.

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

Wordstream (n.d.) Click-Through Rate (CTR): Understanding Click-Through Rate for PPC. https://www.wordstream.com/click-through-rate

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 2:
Estimates of Treatment Effects

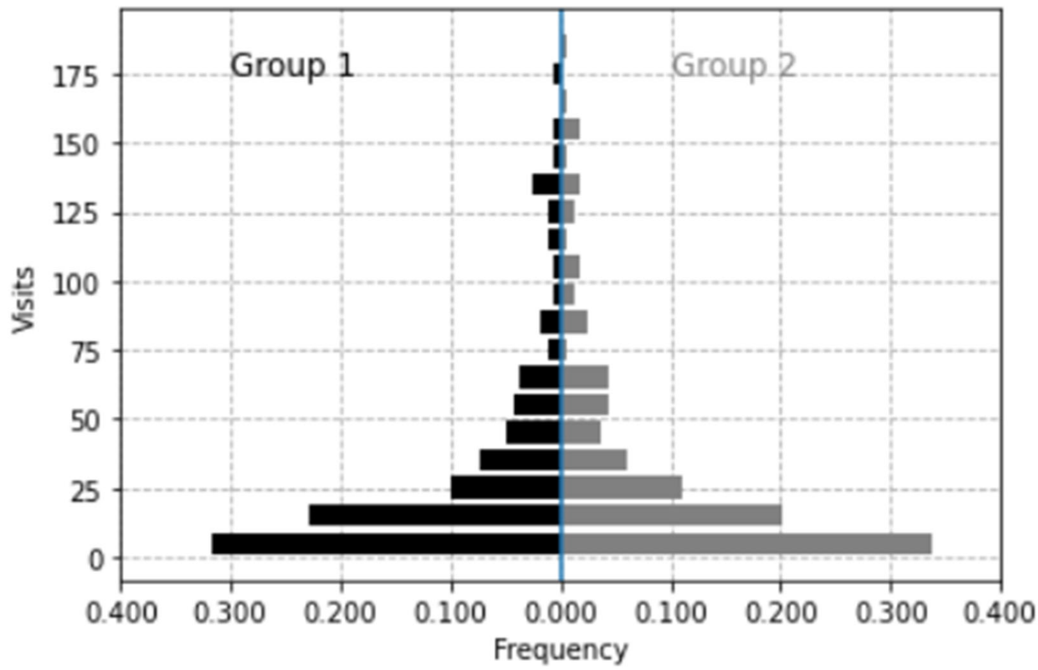
	Estimate	95% Confidence Interval
Estimate of Treatment Effect: Clicks[a] *(Dependent variable = Difference in Monthly Clicks)*		
1) ATE – *Mean Difference*	0.111	(0.086, 0.135)
2) Stage III Balance Test:[c] *Mean Difference*	-0.038	(-0.083, 0.007)
3) ATE – *Median Difference*	0	(0, 0)[d]
4) Stage III Balance Test:[c] *Median Difference*	0	(0, 0)[d]
5) Mean Number of Monthly Page Visits (Treated Originals/Stage II) = 7.1[e]		
6) Median Number of Monthly Page Visits (Treated Originals/Stage II) = 3[e]		
7) Percent of Treated Originals with Positive Page Visits (Stage II) =83.4%[e]		
Estimate of Treatment Effect: Page Visits to Replication Studies[b] *(Dependent variable = Difference in Monthly Page Visits)*		
8) ATET – *Mean Difference*	0.235	(-0.071, 0.541)
9) Stage III Balance Test:[c] *Mean Difference*	0.128	(-0.275, 0.531)
10) ATET – *Median Difference*	0.013	(0, 0.167)[d]
11) Stage III Balance Test:[c] *Median Difference*	-0.012	(-0.167,0)[d]
12) Mean Number of Monthly Page Visits (Treated Replications/Stage II) = 1.76[e]		
13) Median Number of Monthly Page Visits (Treated Replications/Stage II) = 1[e]		
14) Percent of Treated Replications with No Page Visits (Stage II) = 48.5%[e]		

NOTE: The data consist of cluster-level observations. There are a total of 402 observations (67 clusters × 6 months) in Stage II, and 268 observations in Stage III (67 clusters × 4 months). There is an average of approximately 5 studies in each cluster, with half allocated to Group 1 and half to Group 2. In Stage II, only Group 1 studies are treated. In Stage III, both groups are treated.

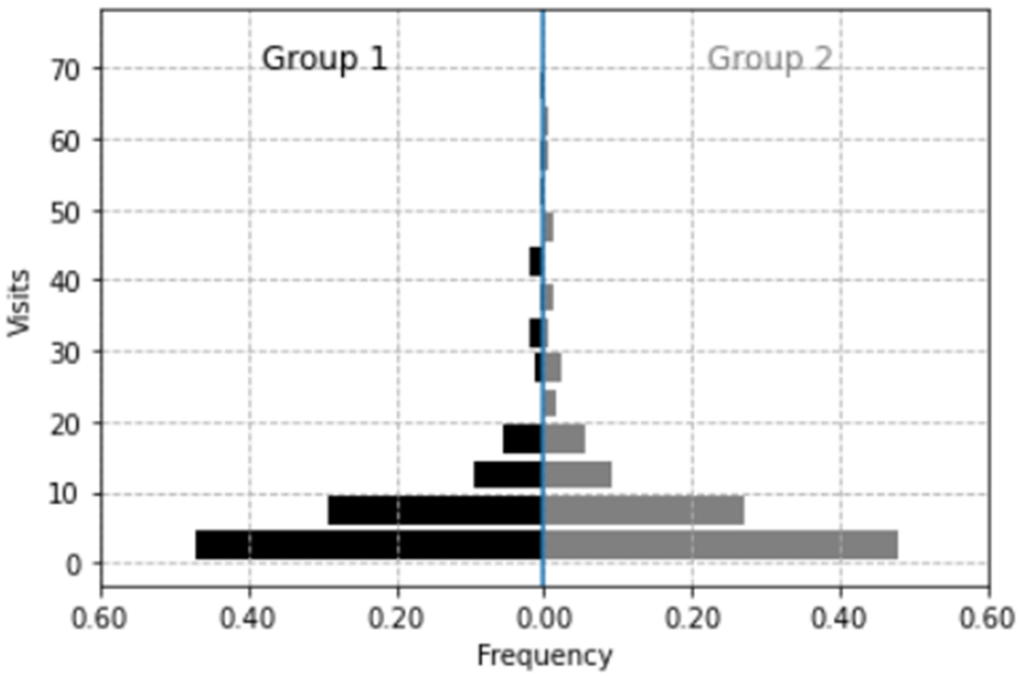
[a] For the top panel, an individual observation consists of the difference in the average number of clicks for treated studies minus the average number of clicks for untreated studies in the same cluster. Note that untreated, original studies can have a click-through if the researcher finds the replication included on the Citations tab of the original study. However, this rarely happens.
[b] For the bottom panel, an individual observation consists of the difference in the average number of page visits to treated replication studies minus the average number of page visits to untreated replications in the same cluster.
[c] In Stage III, all studies are treated. If our stratified randomization procedure is effective, there should be no difference between Group 1 and 2 studies with respect to clicks/page visits. The balancing test consists of performing the same statistical analysis for Stage III as we did for Stage II.
[d] A 95% confidence interval was constructed by bootstrapping the sample of differences 10,000 times, each time taking the median of the bootstrapped sample. The 10,000 bootstrapped, median values were then rank ordered, with the lower and upper bounds constructed by taking the 250th and 9750th values.
[e] Sample statistics are calculated by aggregating monthly values for all treated studies during the Stage II period

FIGURE 1:
Balance Check: Density Plots of Page Visits to Group 1 and Group 2 studies[9]

A. Page Visits to Original Studies



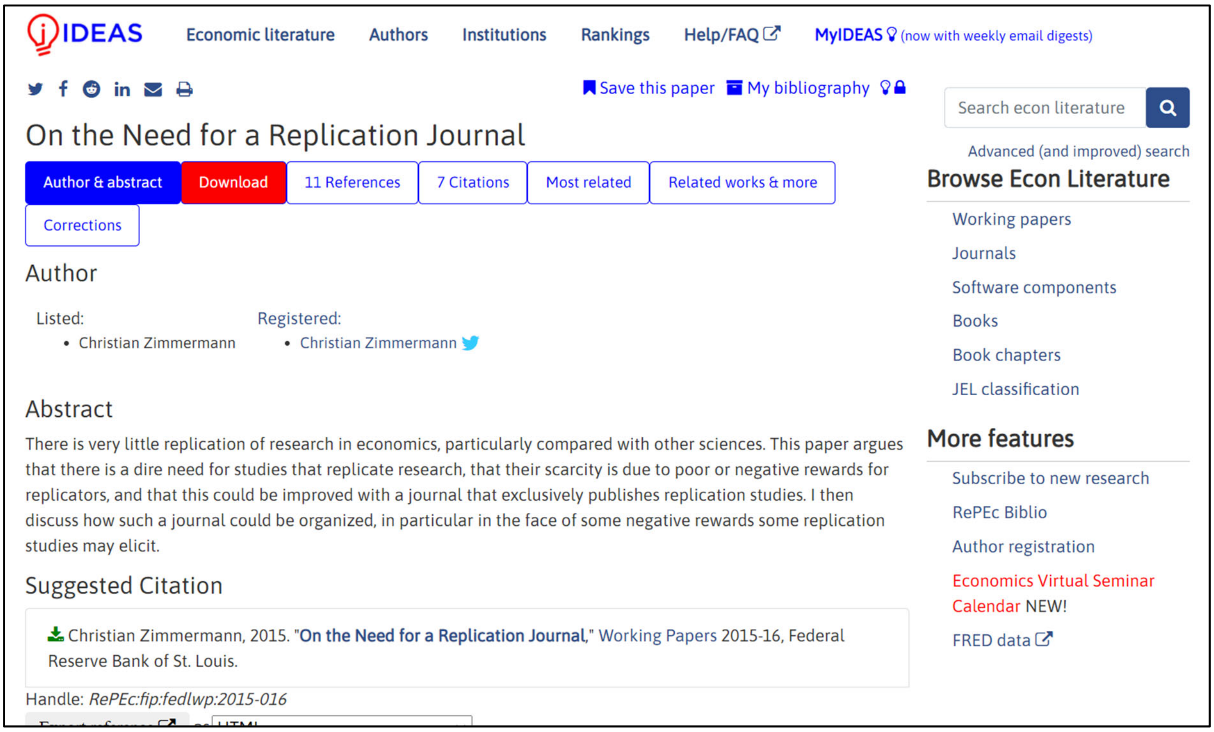
B. Page Visits to Replication Studies



[9] 9 original papers with more than 200 visits (1A) and 4 replications with more than 80 visits (1B) are not shown in the above graphs.

FIGURE 2:
Depiction of the Treatment (Original Study)

A. No Treatment



B. Treatment

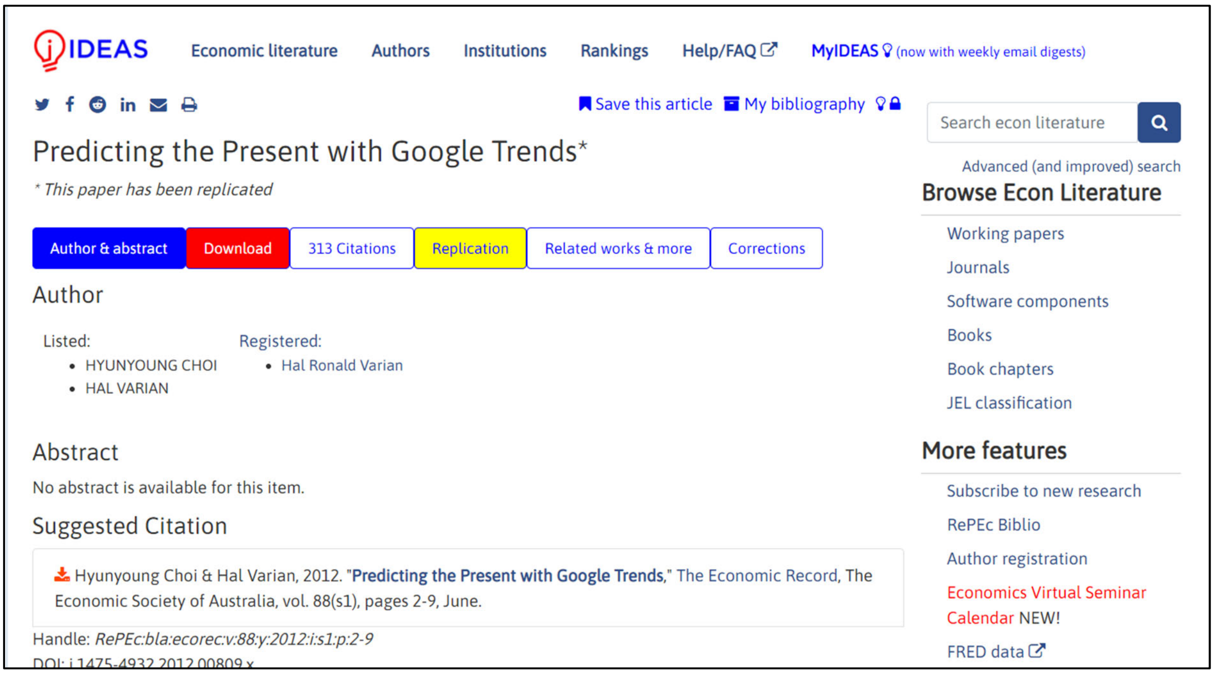


FIGURE 3:
Depiction of the Treatment (Replication Study)

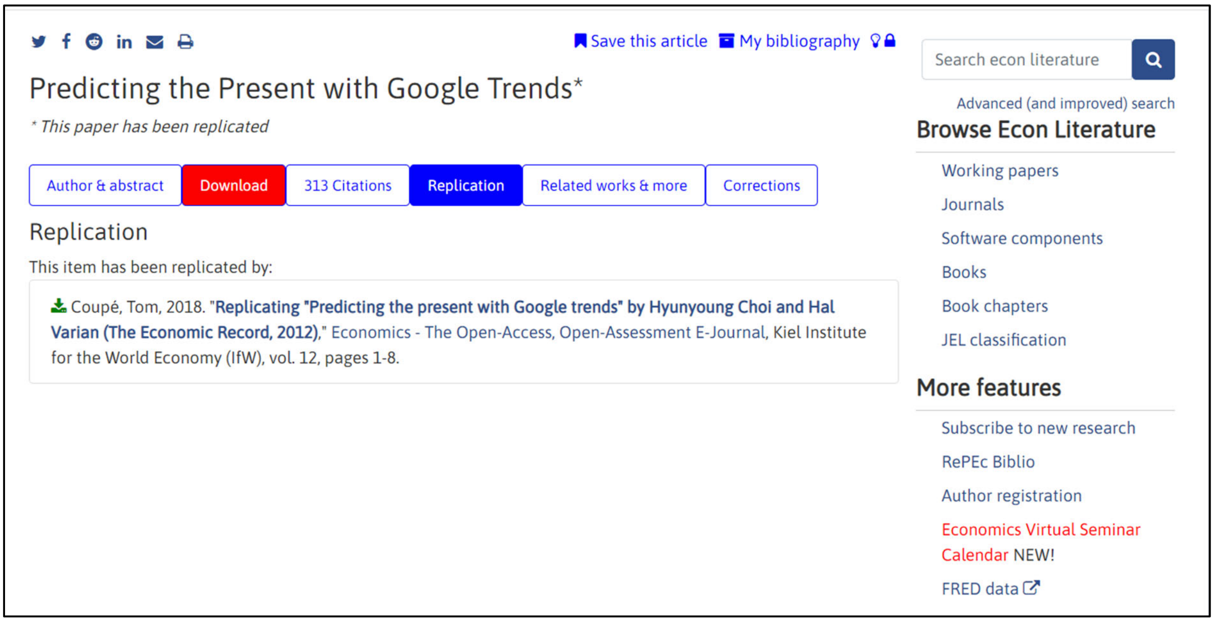


FIGURE 4:
Click-Throughs as a Percent of Total Page Visits (Original Study)

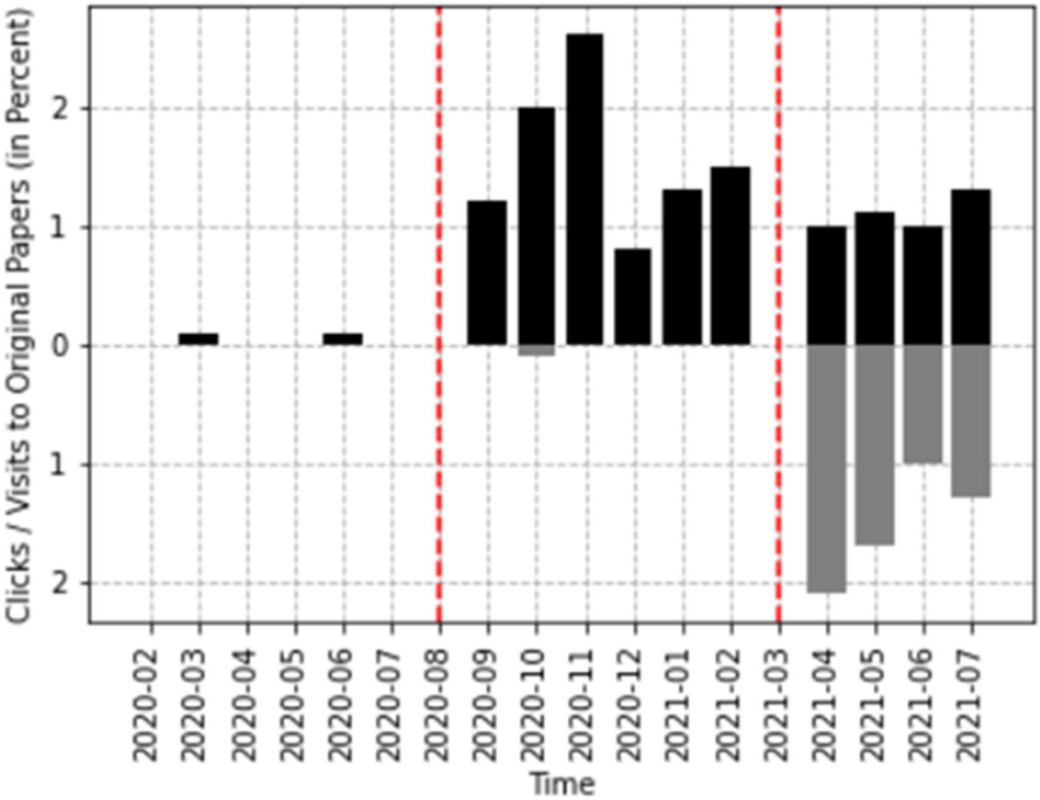


FIGURE 5:
Click-Throughs as a Percent of Total Page Visits (Replication Study)

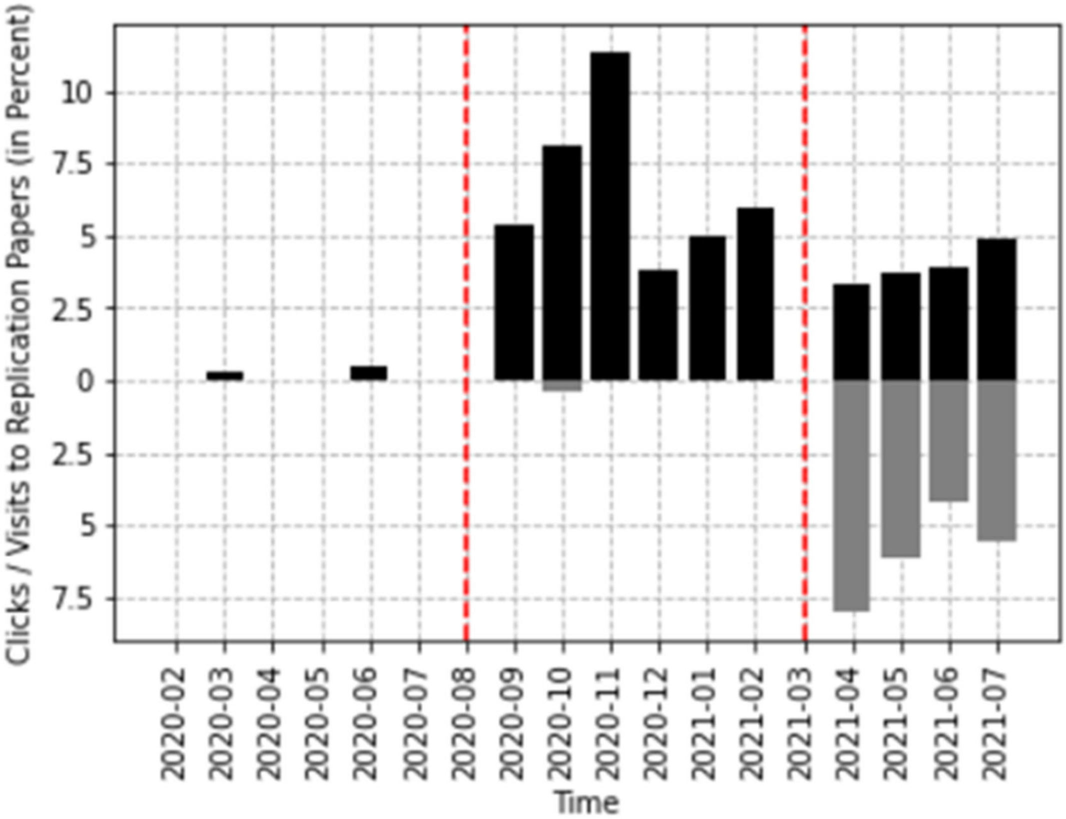


FIGURE 6:
Total Page Visits (Replication Studies)