The Effect of Replications on Citation Patterns: Evidence From a Large-Scale Reproducibility Project

Felix Schafmeister
Department of Economics, Stockholm School of Economics

Abstract
Replication of existing research is often referred to as one of the cornerstones of modern science. In this study, I tested whether the publication of independent replication attempts affects the citation patterns of the original studies. Investigating 95 replications conducted in the context of the Reproducibility Project: Psychology, I found little evidence for an adjustment of citation patterns in response to the publication of these independent replication attempts. This finding was robust to the choice of replication criterion, various model specifications, and the composition of the contrast group. I further present some suggestive evidence that shifts in the underlying composition of supporting and disputing citations have likely been small. I conclude with a review of the evidence in favor of the remaining explanations and discuss the potential consequences of these findings for the workings of the scientific process.

Keywords
scientific communication, statistical analysis, open data, preregistered

Replication of existing research is often referred to as one of the cornerstones of modern science. However, direct replications, defined as the same analyses being conducted on newly collected data using original materials, have historically been published infrequently. Recently, a number of systematic, large-scale replication attempts have been published in prominent scientific journals (e.g., see Camerer et al., 2016, 2018; Klein et al., 2018; Open Science Collaboration, 2015), and the question of replicability has received a substantial amount of attention in subsequent academic publications and media outlets.

The replication projects published thus far have undoubtedly succeeded in raising awareness of methodological shortcomings with regard to the power of research designs and the pitfalls of postdiction, culminating in the advent of Registered Reports (Nosek & Lakens, 2014) and a widespread adoption of preregistration of research designs in psychology (Nosek & Lindsay, 2018). However, little is known about how these replication attempts have shaped the way specific findings are perceived in the literature. In this study, I attempted to fill this gap by analyzing changes in yearly citation patterns of articles replicated in the context of the Reproducibility Project: Psychology (RP:P; Open Science Collaboration, 2015) after the publication of its results in Science in 2015.

A single replication attempt, especially if insufficiently powered, provides only limited information about whether a prior research result is robust. However, I worked under the assumption that replication attempts can shift beliefs about the validity of an existing research finding (McDiarmid et al., 2021). In particular, replications that produce evidence in line with the original findings should strengthen beliefs in their validity, whereas nonsupportive or contradicting replications should weaken these beliefs (Earp & Trafimow, 2019).
2015). This motivated my main research question: Does the publication of supportive versus nonsupportive replication attempts affect the frequency with which the underlying studies are cited? I believe that the answer to this question could provide important insights into the workings of the scientific process by illuminating the extent to which new insights replace or strengthen existing knowledge and thereby shape the way new research is conducted.

I found that neither supportive nor nonsupportive replications have had a statistically significant effect on the number of times articles are cited. This null result is robust to different replication criteria and various model specifications as well as to alternative compositions of the contrast group. In investigating potential explanations for my findings, I present some suggestive evidence that shifts in the underlying composition of supporting and disputing citations have likely been small and discuss how the remaining explanations fit with the evidence.

Although I acknowledge that a number of forces are at the heart of my results, some of the contending explanations have particularly daunting consequences and therefore deserve additional attention: If researchers either were unaware of replication results or, in spite of their awareness, chose to discount them and continued to cite the original results at face value, this could considerably limit the self-corrective ability of the scientific process because it would reduce the likelihood that research results that have been called into question will be phased out and replaced by new insights. This could be particularly problematic in light of evidence by Gneezy and Serra-Garcia (2021), who report that studies that failed to replicate within three publication markets were conducted) have higher citation counts than studies that replicated successfully.

Method

Setting

My empirical approach built on the sampling design of the RP:P (Open Science Collaboration, 2012, 2015). This project was a coordinated effort by the Center for Open Science that brought together a large and diverse group of more than 200 researchers and aimed to test the reproducibility of psychological science. Participating research teams chose studies to replicate from a pre-defined sampling frame and subsequently conducted independent replication attempts “using high-powered designs and original materials when available” (Open Science Collaboration, 2015, abstract). Ultimately, the results of 100 such replication attempts were included in the publication, which concluded that merely 36% of the replications had resulted in statistically significant results compared with 97% of the original studies.

Although a number of drawbacks in the design of the RP:P have been identified, it is one of the most comprehensive attempts to study replicability and followed a strict sampling protocol designed to minimize selection effects. This particular feature made it an appealing setting to study the present research question because it allowed for the definition of a contrast group against which the effect of replications can be evaluated.

Specifically, the sample of eligible studies for the RP:P was selected from the 2008 issues of the Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP), Journal of Personality and Social Psychology (JPSP), and Psychological Science (PS). Replication teams chose studies from among the first 30 articles published in each of these journals beginning with the first 2008 issue; additional articles were made available in sets of 10 in case of additional demand. This sampling strategy limits concerns regarding the selection of articles on the basis of observable or unobservable characteristics, which could complicate an analysis of citation patterns if those characteristics were correlated with both the likelihood of successful replication and citation patterns. Further, I expected that if the articles included in the RP:P sampling frame had not been selected for replication, they would have had similar citation patterns over time as studies published immediately prior or shortly after in the same journals.

The resulting variation therefore lends itself to be exploited in a generalized differences-in-differences

Statement of Relevance

A crucial feature of the scientific process is its ability to self-correct by identifying and promoting robust results and phasing out those that fail to hold up to further scrutiny. This process ensures that through the cumulative generation of research and systematic replications, a reliable body of knowledge is created. This article provides a test of this self-corrective ability “in the medium run” by analyzing articles replicated in the context of the Reproducibility Project: Psychology. I investigated changes in citation patterns around the time the replication results were published, and my findings do not support the hypothesis that citation rates changed in response to independent replications. These results thus emphasize the need for replication on a large scale and the effective communication of replication results.
design. This design compares changes in yearly citations after the publication of the replication results between the replicated and the contrast sample. Under the assumption of parallel trends—in the absence of the RP:P replications, the different sets of studies would have followed similar citation trajectories—this allowed me to identify the effect of interest.

Sample selection

The definition of a contrast group is complicated by two notable deviations of the RP:P from its intended sampling frame. First, because of constraints regarding available instruments, samples, and knowledge among replicators, not all studies in the sampling frame could be replicated. Second, the sampling frame was not always followed, resulting in some studies being replicated despite being published in an issue that had not been made available for selection.

This introduces some degree of selection into the sample of replicated studies. The first deviation implies that studies with certain features were infrequently subjected to replications. This becomes an issue if these types of studies are also subject to different citation trends, thus potentially violating the common-trends assumption that underlies my identification of the effect of interest. Similarly, the second deviation suggests that some studies might have been selected by groups of researchers particularly interested in their results. Although the exact drivers of these choices are unclear, this interest might correlate with factors that affect citation trends and could therefore present a challenge to the identifying assumption.

To deal with these concerns, I implemented the following sample-selection protocol. I collected data for the 95 unique articles that had significant original findings and for which a replication was reported in the RP:P, thus excluding the three studies with insignificant original findings as well as the two duplicate replications. These data were combined with information from articles that were published in the same journals in adjacent months. This initial selection of eligible articles is detailed in Table 1. Within the sampling frame, all original research articles were considered eligible. Other material, such as editorials, commentaries, corrigenda, and book reviews, were excluded.

I further implemented a set of exclusion criteria intended to increase the comparability of the contrast and treatment groups. These exclusion criteria were designed to mimic the constraints faced by the replication teams in the RP:P, eliminating studies that use hard-to-access samples or that require specialized instruments that are not readily available in most laboratories. Further, I eliminated articles that did not include any experimental results that can be represented by a single statistical inference test or effect size.4 I first coded the articles according to these criteria. To ensure the fidelity of the coding, a research assistant blind to the research question additionally coded a random subsample of 169 studies. The two classifications aligned in 97% of cases.

The sampling frame laid out in Table 1 contains a total of 476 articles, of which 202 were published in PS, 144 in JPSP, and 130 in JEP. Applying the exclusion criteria left 329 articles, of which 117 were published in PS, 112 in JPSP, and 100 in JEP. This implies journal shares ranging from 30% for JEP to 36% for PS, which closely matches the percentages of articles replicated in the RP:P (29% for JEP to 41% for PS).

Variable definitions

My main outcome variable was the number of citations that a publication received in a specific year. I collected these data from the Web of Science Core Collection between 2010 and 2019. This time range was chosen with the aim of constructing a balanced panel data set despite the fact that the sampling frame contains articles published between 2007 and 2009. The fact that the coverage of the Web of Science is limited within the social sciences (Kousha & Thelwall, 2007) was unlikely to represent a problem for my empirical strategy because there should not be any reason to expect that coverage of citing articles would have changed over time for different studies in my sample. In the main analyses, I applied an inverse hyperbolic sine (IHS) transformation to account for the nonnormality of the data. This transformation is conceptually similar to the
approach of adding 1 unit to each observation and taking the natural logarithm recommended by Thelwall and Wilson (2014) but is generally considered less arbitrary (Bellemare & Wichman, 2020).

Further, to address the research question, I required a measure to evaluate the degree to which a replication attempt supports the original finding. An intuitive criterion relies on the binary decision of whether the replication produces evidence that is statistically significant at the 5% level in the same direction as the original finding. The main point of contention in using this criterion is its inability to quantify the strength of the evidence contained in a replication coded as failed or successful. For example, a failed replication attempt might not be able to detect a true effect if it is itself underpowered, whereas a high-powered replication might uncover an effect that is orders of magnitude smaller than the original effect. The “small-telescopes” approach of Simonsohn (2015) is based on the idea of detectability in that it tests whether the effect obtained in a replication study is smaller than an effect size that the original study had only 33% power to detect. In comparison, the Bayesian approach applied by Etz (2015) is based on the calculation of Bayes factors (BFs) comparing the hypothesis of no effect with an alternative of the effect size found in the original publication, originally proposed by Verhagen and Wagenmakers (2014).

In the main specification, I employed the intuitive binary criterion, which has been used as the main replication criterion in the RP:P and has been widely communicated. In a secondary analysis, I instead used an alternative definition of failed replications based on the Bayesian approach outlined above. To ensure the comparability of my analyses across the different criteria, I collapsed the continuous BF into three categories. Specifically, I coded a replication as successful if it yielded a BF of 3 or more and as unsuccessful if it yielded a BF of one third or less; all remaining replications were coded as inconclusive. In Section S1 in the Supplementary Material available online, I further implemented similar analyses on the basis of the small-telescopes approach.

**Summary statistics**

The main criterion classified 35 replication attempts as successful and the remaining 60 as unsuccessful. The Bayesian criterion considered 31 replications successful, 34 unsuccessful, and 26 inconclusive. The discrepancy in the number of articles coded is due to three studies missing effect-size estimates for either the original analysis or the replication attempt in the RP:P files. For one further study, the BF could not be calculated, resulting in a total of 91 replicated studies for which I had a valid classification.

Regarding the distribution of yearly citations, the average study in my estimation sample received 7.34 (SD = 7.91) citations, and the share of article years with no citations was 0.068. Further, yearly citations have been on an increasing trajectory even several years after the initial publication: On average, there were 6.46 (SD = 5.37) citations per year between 2010 and 2014 and 8.22 (SD = 8.81) yearly citations from 2015 to 2019. This reflects the fact that the articles under consideration are among the most influential in the field and thus have a long citation half-life (Walters, 2011).6

**Empirical strategy**

The core of the estimation strategy lies in the assumption that in the absence of the RP:P replications, the studies included in the contrast group would have been subject to the same citation trends as those in the RP:P sample. Under this assumption, I could identify the effect of interest in a generalized differences-in-differences estimation framework by comparing changes in citation trends. I employed two closely related operationalizations: The first assumed that treatment effects are constant over time, whereas the second allowed for time-varying treatment effects.

The first model was

$$Y_{it} = \alpha_i + \beta_t + \delta_i \text{Successful} + \delta_i \text{Failed} + \rho + \Gamma_{it} + \epsilon_{it}, \tag{1}$$

where $Y_{it}$ is the IHS-transformed number of citations for article $i$ in year $t$, with $t$ running from 2010 to 2019, and $\alpha_i$ and $\beta_t$ indicate article and year fixed effects, respectively. $\delta_i \text{Successful}$ and $\delta_i \text{Failed}$ are indicator variables taking a value of 1 for observations from 2015 or later for which the replication attempt has been successful or unsuccessful, respectively. Finally, the matrix $\Gamma_{it}$ indicates time-varying, article-specific controls that are described in more detail below. The model is estimated with ordinary least squares (OLS), and standard errors allow for clustering at the article level.

The two sets of fixed effects $\alpha_i$ and $\beta_t$ control for time-invariant, article-specific variation as well as article-invariant, year-specific effects. In particular, assuming that beliefs about an article’s replicability as well as their effect on citations are time invariant, the article fixed effects account for initial heterogeneity in beliefs about the likelihood of replication success. This is important because considerable heterogeneity likely exists in beliefs about the likelihood of replication success in the contrast group, and no survey measures are available to control for these beliefs. Dreber et al.
(2015) have shown that prediction markets, in which experts bet on the outcome of the replications, can be successful at predicting the results of replications, as can machine-learning algorithms (Altmejd et al., 2019; Yang et al., 2020). This suggests that researchers have meaningful priors about the likelihood of replication, which should correlate with the number of citations (Gneezy & Serra-Garcia, 2021). Hence, these fixed effects alleviate concerns about articles that were replicated successfully receiving on average higher (or lower) citations than those in the contrast group because the latter is a mixture of studies that would not.

The second model differs from the first only in that it allows for time-varying treatment coefficients. It is given by

\[ Y_{ijt} = \alpha_i + \beta_j + \sum_{j \in S} D_{ijt}^{\text{Successful}} \delta_j + D_{ijt}^{\text{Failed}} \rho_j + X_{ijt}' \gamma + \epsilon_{ijt}, \]

with $S$ equal to \{2010, . . . , 2019\}\{2014\}. The variables $D_{ijt}^{\text{Successful}}$ take a value of 1 if an article was replicated successfully and the observation is from year $j$, and equivalently for $D_{ijt}^{\text{Failed}}$. The 2014 coefficients are normalized to 0 and thus serve as the reference categories against which the posttreatment coefficients will be compared.

This specification has two advantages. First, it allows one to gauge whether any effects of replications on citation patterns arise immediately or only with a lag as well as whether these effects persist over time. Second, the pre-2014 coefficients allow for an investigation of the identifying assumption: If the estimated coefficients prior to the publication of the RP:P results were significantly different from zero, this would call into question the assumption that replicated and nonreplicated articles would have followed similar citation trends in the absence of the RP:P.

It is important to note that the above exposition deviates from my preregistered analysis plan in a number of ways. Most importantly, I specified Equation 2 as my main model and proposed to test my main hypothesis by testing for the joint significance of the estimated effects and depends strongly on the choice of base year. The results of this test and further explanations are provided in Section S1.

### Results

Table 2 presents four specifications of Equation 1 that differ in terms of which set of time-varying control variables were included in the model. Row 2 presents the prespecified model equation including fixed effects for the number of years since publication. Rows 3 and 4 represent further attempts at accounting for potential deviations from the common-trends assumption by including separate year-fixed effects for each journal (Row 3) or allowing for issue-specific linear trends (Row 4). In all specifications, successful replications were positively associated with citation rates, $\delta = 0.053$, 95% confidence interval (CI) = [−0.09, 0.195], $t(423) = 0.73$, $p = 0.468$, and failed replications were negatively associated with citation rates, $\rho = -0.035$, 95% CI = [−0.159, 0.089], $t(423) = -0.56$, $p = 0.576$, but neither of these effects

---

### Table 2. Differences-in-Differences Estimates of the Effect of Replications on Yearly Citations: Main Criterion

| Model | Successful replication | Failed replication | Fixed effects |
|-------|------------------------|--------------------|--------------|
|       | $b$ | $p$     | $b$ | $p$ | Article | Year | Years since publication | Number of observations |
| Baseline | 0.037 | .609 | -0.051 | .412 | Yes | Yes | No | 4,240 |
| Baseline + fixed effects for number of years since publication | 0.053 | .468 | -0.055 | .576 | Yes | Yes | Yes | 4,240 |
| Baseline + separate year-fixed effects for each journal | 0.082 | .260 | -0.050 | .424 | Yes | Yes | Yes | 4,240 |
| Baseline allowing for issue-specific linear trends | 0.102 | .319 | -0.010 | .906 | Yes | Yes | Yes | 4,240 |

Note: The table presents ordinary least squares estimates of Equation 1 with different sets of time-varying control variables. The outcome variable (yearly citations) has been transformed using an inverse hyperbolic sine to account for the nonnormality of the data. Successful replication is an indicator variable that takes the value of 1 if the article has been successfully replicated and $t$ is greater than 2014. A replication is considered successful if it yields an effect in the same direction as the original effect and the replication $p$ value is below .05. Standard errors allow for clustering at the article level and are presented in parentheses; $p$ values were obtained from two-sided $t$ tests.

---

\[ Y_{ijt} = \alpha_i + \beta_j + \sum_{j \in S} D_{ijt}^{\text{Successful}} \delta_j + D_{ijt}^{\text{Failed}} \rho_j + X_{ijt}' \gamma + \epsilon_{ijt}, \]
reached statistical significance. Under the approximation that \( \ln(y + \sqrt{y^2 + 1}) \approx \ln(2y) \) for large values of \( y \), the coefficient estimates obtained in the main specification, taken at face value, suggest that successful replications led to an increase in yearly citations of around 5% and that unsuccessful replications led to a decrease in yearly citations of around 4%. For the average article in my sample, which has roughly eight citations per year, this would imply a change of \( \pm 1 \) citation every 2 to 3 years.

**Potential explanations**

An obvious question is whether my inability to reject the null hypothesis is the result of the study being inadequately powered. On the basis of the estimated standard errors in the main specification, I had 80% power to detect an effect of around 0.2 in IHS units for successful replications and around 0.18 IHS units for failed replications at an \( \alpha \) of .05, corresponding to 0.19 and 0.17 standard deviations, respectively. I thus conclude that my results are informative about effect sizes in this range but cannot make meaningful statements about the presence or absence of subtler effects. Although this important qualification should be taken into account in interpreting my results, I nonetheless emphasize that, as of yet, the RP:P represents the largest available sample to test my main hypothesis. It is my hope that future systematic replications will allow for the effect to be analyzed in even larger samples, thus enabling the investigation of smaller effect sizes.

Another possible explanation lies in the relatively short postreplication time window I had at my disposal. The citation count in the Web of Science is largely based on articles published in peer-reviewed journals. Hence, in order to be represented in the citation count, an article has to be written and pass through peer review, two processes that take a considerable amount of time. This could call into question my assumption of time-invariant treatment effects, a violation that could bias my estimates toward zero and hence drive the null result.

To investigate this concern, I present OLS estimates of Equation 2 in Figure 1, including the additional set of fixed effects for years since publication. The plot reveals two main patterns. First, the estimated coefficients in the years up to the publication of the RP:P do not differ substantially from zero for either the successful or the unsuccessful replications, which increases confidence in the common-trends assumption underlying my previous estimation results. Second, the plot provides scant evidence for treatment effects arising in later years. Although some of the estimated yearly coefficients are significantly different from zero at the 5% level on the basis of a two-sided \( t \) test, no consistent pattern arises. Further, these effects are dependent on the normalization with regard to the base year 2014.
This is reflected in the fact that already prior to 2014, all estimated coefficients for successful replications are positive, and all estimated coefficients for unsuccessful replications are negative. As a result, the coefficients would appear considerably more muted if compared with those of the average prereplication year. Overall, the figure provides little evidence that time-varying effects were driving the null result.

A third candidate explanation for the absence of a statistically significant effect is that the replication criterion did not account for the strength of the evidence obtained in the replication. This issue is particularly important because the power calculations underlying the RP:P replications were based on the original effect sizes. In the presence of a type-M error (Gelman & Carlin, 2014), this practice is likely to result in replication attempts that are themselves not sufficiently powered to detect the effects of interest and hence could render some of the evidence too weak to shift researchers’ beliefs. In line with this argument, the Bayesian analysis by Etzion and Vandekerckhove (2016) suggests that many original studies and their replications did not provide strong evidence for either the null or the alternative hypothesis, implying a limited need for belief updating in many cases. Thus, my measure of the effects of failed replications represents a weighted average of inconclusive and refuting information, which might shift the estimated coefficient toward zero.

I addressed this concern by classifying replication attempts according to the Bayesian criterion, thereby introducing an explicit distinction between failed and inconclusive replication attempts (see Table 3). The results were largely in line with those obtained previously: Neither successful, \( \hat{\delta} = 0.057, 95\% \text{ CI} = [-0.091, 0.204], t(419) = 0.76, p = .45; \) inconclusive, \( \Phi = -0.127, 95\% \text{ CI} = [-0.319, 0.066], t(419) = -1.30, p = .196; \) nor failed, \( \rho = -0.002, 95\% \text{ CI} = [-0.146, 0.143], t(419) = -0.02, p = .98, \) replication attempts displayed a statistically significant effect. Although one has to be cautious in interpreting insignificant differences between coefficients, it is worth noting that the estimates in Table 3 do not support my previous conjecture. If anything, the coefficient on inconclusive replications was more negative than that on failed replications. The corresponding time-varying coefficients are presented in Figure 2 and closely mirror the coefficients obtained in Figure 1.

### Robustness

In this subsection, I address some remaining concerns regarding the validity of my estimates along two dimensions.

First, as described in more detail in Section S2 in the Supplemental Material, I acknowledge that the inclusion criteria at best resulted in a rough approximation of the selection process of studies to be replicated in the RP:P. To ensure that my sample selection did not drive the results, I included all nonretracted studies in the contrast group in Row 1 of Table 4. This expansion of my estimation sample resulted in small gains in the precision of estimates and both successful, \( \hat{\delta} = 0.064, 95\% \text{ CI} = [-0.074, 0.202], t(567) = 0.91, p = .362, \) and unsuccessful, \( \rho = -0.024, 95\% \text{ CI} = [-0.143, 0.095], t(567) = -0.39, p = .695, \) replications continued to yield insignificant coefficient estimates.
Second, I considered whether my results were sensitive to the choice of transformation applied to the outcome variable. Thus far, all regressions employed the IHS transformation. An alternative approach that has frequently been implemented when working with count data is adding a value of 1 to each observation and subsequently taking the logarithm. Although this approach has been criticized by some researchers for being inherently arbitrary (Burbidge et al., 1988; Campbell & Mau, 2020), in Table 4, I present estimates of Equation 1 for both untransformed citation counts (Row 2) and the logarithmic transformation (Row 3). Although the coefficients are not directly comparable across models, the broad patterns are very similar, with both successful and unsuccessful replications resulting in small and statistically insignificant coefficient estimates—successful untransformed: $\hat{b} = 0.056$, 95% CI $= [-1.158, 1.27]$, $t(423) = 0.09, p = .928$; successful logarithmic: $\hat{b} = 0.044$.

**Table 4. Robustness Checks: Sample and Outcome Variables**

| Specification                          | Successful replication | Failed replication | Fixed effects | Number of observations |
|----------------------------------------|------------------------|--------------------|---------------|------------------------|
| All articles using IHS-transformed citation counts | $b = 0.064, p = .362$ | $b = -0.024, p = .695$ | Yes, Yes, Yes | 5,680 |
| Main sample using raw citation counts  | $b = 0.056, p = .928$ | $b = -0.480, p = .393$ | Yes, Yes, Yes | 4,240 |
| Main sample using logarithmically transformed citation counts: ln(citations + 1) | $b = 0.044, p = .475$ | $b = -0.032, p = .548$ | Yes, Yes, Yes | 4,240 |

Note: The table presents ordinary least squares estimates of Equation 1. Row 1 uses the full sample for the contrast group, whereas Row 2 and Row 3 use the sample subjected to the exclusion criteria outlined in the text. The outcome variable (yearly citations) in Row 1 has been transformed using the inverse hyperbolic sine (IHS), whereas the values in Row 2 were obtained using raw citation counts, and the values in Row 3 were obtained by adding a value of 1 to each observation and taking the natural logarithm. Successful replication is an indicator variable that takes the value of 1 if the article has been successfully replicated and $t$ is greater than 2014. A replication is considered successful if it yields an effect in the same direction as the original effect and the replication $p$ value is below .05. Standard errors allow for clustering at the article level and are presented in parentheses; $p$ values were obtained from two-sided $t$ tests.
95% CI = [−0.076, 0.163], \( t(423) = 0.72 \), \( p = .475 \); unsuccessful untransformed: \( \rho = −0.48 \), 95% CI = [−1.585, 0.625], \( t(423) = −0.85 \), \( p = .393 \); unsuccessful logarithmic: \( \rho = −0.032 \), 95% CI = [−0.135, 0.072], \( t(423) = −0.60 \), \( p = .548 \).

**Discussion**

The failure of my analyses to reject the null hypothesis that there was no effect of RP:P replications on yearly citation counts ran counter to my hypothesis that citation patterns should change as researchers adjust their beliefs about the validity of an existing research result. In the following, I outline a number of contending explanations for this null result and discuss the extent to which they are in line with the data.

First, a necessary condition for belief updating in response to replication attempts is researchers’ awareness of the replication results. Previous findings of Simkin and Roychowdhury (2005) suggest that a large number of citations are merely copied from existing reference lists and not actually read by the citing authors, making it likely that at least some researchers remain unaware of existing replications for the studies they cite.

Such inattention is likely exacerbated by the general difficulty of acquiring information about replication results. Unpublished replications are often difficult to find, but even if replication results are published, finding and evaluating them requires a substantial time investment from citing researchers. This concern carries particular weight in my setting because the RP:P was designed with the intention to draw conclusions about replicability on an aggregate level rather than to scrutinize individual research results. As a consequence, the outcomes of individual replication attempts were neither discussed in detail by the Open Science Collaboration (2015), nor were citations to the original studies included in their article, requiring researchers interested in the results of individual replication attempts to delve into the supplemental materials.

This factor substantially qualifies the external validity of my findings because other replication studies might discuss individual replication outcomes in more detail and are more easily picked up by search engines if they are similar in title and include direct references to the original study. This increased visibility has the potential to alter the citation impact of a replication attempt compared with the effects that I uncovered in the context of the RP:P; indeed, the case studies by Hardwicke et al. (2021) suggest that somewhat more marked effects might arise in other settings.

Second, even among researchers aware of the replication attempts, belief updating might have been limited. Although McDiarmid et al. (2021) show that researchers updated their beliefs about the strength of a research finding in reaction to replications conducted in a number of large-scale replication projects (not including the RP:P), it is unclear to what extent these findings can be extrapolated to my setting. In particular, the authors note the possibility that experimenter demand and observer effects could have resulted in inflated estimates of researchers’ true belief updating. Moreover, some authors of original studies that were replicated in the RP:P voiced concerns regarding the fidelity of the replication attempts (e.g., Bressan, 2019; Gilbert et al., 2016; and replies to the RP:P published on OSF by the original authors). Although Ebersole et al. (2020) show that the results of the RP:P replications were not sensitive to using peer-reviewed protocols, if citing researchers were nonetheless convinced that the replication attempts were not true to the original study, this might have weakened belief updating.

Other potential explanations could lie in articles gaining additional citations by being cited in the context of replications rather than for their content or in the citation count not taking into account citation content. Regarding the first argument, if this factor were to play a large role, one would expect to find an increase in citation rates for successful and inconclusive replications. In particular, because inconclusive replications were largely considered failures by the main criterion, these replications were likely among the most controversial and thus should have received the largest number of citations through this channel, a hypothesis that is not borne out by the present results.

Further, the second concern suggests that even if one cannot detect changes in total citation counts, the composition of supporting and disputing citations might have shifted. The analyses above are unable to directly shed light on the importance of this explanation because I am missing a reliable measure of citation content. Recently, a large-scale source of citation content classifications has become available through the website scite.ai, which uses deep learning to determine whether a citation supports, disputes, or merely mentions an existing research result. However, at the time of writing, the service is still in its beta stage and has only limited coverage. Hence, rather than subjecting these noisy measures to a formal statistical analysis, I present some suggestive evidence on the role of this channel.

According to the scite.ai classifications, only a small minority of citations are disputing or supporting existing findings. In the 10 years between 2010 and 2019, the average article in the RP:P sample has been subject to merely 0.83 disputing and 4.39 supporting citations, and 46% of the sample was never disputed. Moreover, investigating the timing of citations, I found little evidence that the frequency of disputing citations has been affected by the replication results. When the main
replication criterion was used, studies that were replicated successfully received on average 0.4 disputing citations between 2015 and 2019, compared with 0.66 in the 5 years prior to replication, and studies that were replicated unsuccessfully received on average 0.38 disputing citations between 2015 and 2019, compared with 0.32 in the 5 years prior to replication. These numbers suggest that even if the RP:P replications shifted citation content, the size of these effects would likely be small.

In conclusion, my analyses fail to support the hypothesis that citation patterns adjust in response to the release of replication results. Among the potential reasons underlying these findings, a lack of attention to and the limited communication of replication results stand out as particularly important. These factors therefore have the potential to slow down the self-corrective ability of the scientific process and addressing them could represent an important step in maximizing the impact of recent advances to improve the quality and reliability of academic research. I am hopeful that technological advances such as scite.ai, with their potential to greatly improve the accessibility of the body of knowledge, can help to alleviate these issues in the future.

Transparency

Action Editor: Marc J. Buehner
Editor: Patricia J. Bauer

Author Contributions

F. Schafmeister is the sole author of this article and is responsible for its content.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser; Grant Nos. P18 0073 and P21 0091) and the Knut and Alice Wallenberg Foundation. Further support for this research was provided by a grant to F. Schafmeister’s PhD supervisor, Anna Dreber Almenberg, from the Marianne and Marcus Wallenberg Foundation.

Open Practices

All data and analysis code have been made publicly available via OSF and can be accessed at https://osf.io/8vgm2. The design and analysis plan for this study were preregistered at https://osf.io/get5x. A number of deviations were made from the preregistration (for more information, see Note 7). This article has received the badges for Open Data and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicscience.org/publications/badges.

ORCID iD

Felix Schafmeister https://orcid.org/0000-0003-0033-576X

Acknowledgments

I am thankful to Anna Dreber Almenberg, Oliver Engist, Magnus Johansson, Brian Nosek, Robert Ostling, Joakim Semb, and Domenico Viganola for their helpful comments. Vera Wellander Lindén provided excellent research assistance.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976211005767

Notes

1. Makel et al. (2012) showed that among all articles published in the top 100 psychology journals since 1900, roughly 1% contain a direct replication of an original result. Makel and Plucker (2014), Makel et al. (2016), Eanschitzky et al. (2007), and Hubbard and Armstrong (1994) present similar results for educational sciences, special education, and marketing.
2. Tahamtan and Bornmann (2018) review the manifold reasons underlying citations that prior research has identified. My research design assumed that at least some citations accrue from sources with the intent to build on the prior study.
3. A more detailed description of the sampling frame can be found in the article by the Open Science Collaboration (2012, 2015).
4. A detailed description of the exclusions is presented in Section S2 in the Supplemental Material available online.
5. For one of the two articles that were replicated twice, the replication outcome according to the Bayesian criterion was ambiguous; one replication attempt was coded as failed and one as inconclusive. I coded this replication as inconclusive.
6. Section S2 presents more detailed summary statistics.
7. Because the observational nature of the data led to some unexpected complications, a number of deviations from the preregistration protocol were made. The Supplemental Material presents an explicit mapping of all changes from the preregistered analysis plan to the final article. The Supplemental Material further contains all analyses that were preregistered but not included in the present article, as well as some of the robustness tests. The main article and Supplemental Material in combination contain all preregistered analyses.

The most important change was to the main estimation model. I prespecified the use of a model with time-varying coefficients but chose to present the coefficients from a time-invariant model as my main analysis instead. The reason for this was that (a) it was noticed that the coefficient estimates were highly dependent on the year that was chosen as the base period, rendering the use of the time-invariant model more robust, and (b) the test I specified based on the time-varying model did not take into account the direction of the effect, which was crucial for the purpose of my study. The time-varying estimates are nonetheless presented in Figures 1 and 2.
Further deviations from the preregistration were due to unanticipated reductions in sample size as a result of missing information. For example, I had not anticipated that my sampling frame contained studies that were subsequently retracted and hence did not have citation data available; in addition, the RP:P files contained a few missing values.

8. For concision, only the coefficients and $p$ values from the preregistered specifications are noted in the text. Unless otherwise stated, reported $p$ values are based on two-sided $t$ tests. Following the recommendation of Wilkinson (1999), I report effect sizes as unstandardized regression coefficients.

References

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., & Camerer, C. (2019). Predicting the replicability of social science lab experiments. PLOS ONE, 14(12), Article e0225826. https://doi.org/10.1371/journal.pone.0225826

Bellemare, M. F., & Wichman, C. J. (2020). Elasticities and the inverse hyperbolic sine transformation. Oxford Bulletin of Economics and Statistics, 82(1), 50–61.

Bressan, P. (2019). Confounds in “failed” replications. Frontiers in Psychology, 10, Article 1884. https://doi.org/10.3389/fpsyg.2019.01884

Burbidge, J. B., Magee, L., & Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. Journal of the American Statistical Association, 83(401), 123–127.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. Science, 351(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 2(9), 637–644.

Campbell, D. L., & Mau, K. (2020). On trade induced technical change: The impact of Chinese imports on innovation, IT and productivity (Working Paper w0264). New Economic School.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. Proceedings of the National Academy of Sciences, USA, 112(50), 15343–15347.

Earp, B. D., & Trafnimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. Frontiers in Psychology, 6, Article 621. https://doi.org/10.3389/fpsyg.2015.00621

Ebersole, C. R., Nosek, B. A., Kidwell, M. C., Buttrick, N., Baranski, E., & Hartshorne, J. (2020). Many Labs 5: Testing pre-data collection peer review as an intervention to increase replicability. Advances in Methods and Practices in Psychological Science, 3(3), 309–311.

Etz, A. (2015). The Bayesian reproducibility project. https://alexanderetz.com/2015/08/30/the-bayesian-reproducibility-project/

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. PLOS ONE, 11(2), Article e0149794. https://doi.org/10.1371/journal.pone.0149794

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research’s disturbing trend. Journal of Business Research, 60(4), 411–415.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, 9(6), 641–651.

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

Gneezy, U., & Serra-Garcia, M. (2021). Nonreplicable publications are cited more than replicable ones. Science Advances, 7(21), Article eab1705. https://doi.org/10.1126/sciadv.abc1705

Hardwicke, T. E., Szucs, D., Thibault, R. T., Crüwell, S., Van den Akker, O., Nuijten, M. B., & Ioannidis, J. (2021). Citation patterns following a strongly contradictory replication result: Four case studies from psychology. MetaArXiv. https://doi.org/10.31222/osf.io/wt5ny

Hubbard, R., & Armstrong, J. S. (1994). Replications and extensions in marketing: Rarely published but quite contrary. International Journal of Research in Marketing, 11(3), 235–248.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahnik, S., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science, 1(4), 443–490.

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. Journal of the American Society for Information Science and Technology, 58(7), 1055–1065.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. Educational Researcher, 43(6), 304–316.

Makel, M. C., Plucker, J. A., Freeman, J., Lombardi, A., Simonsen, B., & Coyne, M. (2016). Replication of special education research: Necessary but far too rare. Remedial and Special Education, 37(4), 205–212.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? Perspectives on Psychological Science, 7(6), 537–542.

McDiarmid, A. D., Tullett, A. M., Whitt, C. M., Vazire, S., Smaldino, P. E., & Stephens, E. E. (2021). Self-correction in psychological science: How do psychologists update their beliefs in response to replications? PsyArXiv. https://doi.org/10.31234/osf.io/hjcm4

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. Social Psychology, 45, 137–141.
Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer, 31*(3), 19–21.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Simkin, M. V., & Roychowdhury, V. P. (2005). Stochastic modeling of citation slips. *Scientometrics, 62*(3), 367–384.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569.

Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics, 12*(1), 203–216.

Thelwall, M., & Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *Journal of Informetrics, 8*(4), 963–971.

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143*(4), 1457–1475.

Walters, G. D. (2011). The citation life cycle of articles published in 13 American Psychological Association journals: A 25-year longitudinal analysis. *Journal of the American Society for Information Science and Technology, 62*(8), 1629–1636.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

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, USA, 117*(20), 10762–10768.