Post-replication citation patterns in psychology: Four case studies

Tom E. Hardwicke[1,2], Dénes Szűcs[3], Robert T. Thibault[4,5], Sophia Crüwell[2], Olmo R. van den Akker[6], Michèle B. Nuijten[6], & John P. A. Ioannidis[2,7,8]

[1] Department of Psychology, University of Amsterdam

[2] Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Charité – Universitätsmedizin Berlin

[3] Department of Psychology, University of Cambridge, UK

[4] School of Psychological Science, University of Bristol

[5] MRC Integrative Epidemiology Unit at the University of Bristol

[6] Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University

[7] Meta-Research Innovation Center at Stanford (METRICS), Stanford University

[8] Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University

Author Note

Correspondence concerning this article should be addressed to Tom E. Hardwicke, Nieuwe Achtergracht 129B, Department of Psychology, University of Amsterdam, 1018 WS Amsterdam, The Netherlands. E-mail: tom.hardwicke@uva.nl

Abstract

Abstract here.

Keywords: Replication, self-correction, citations, citation bias, meta-research

Table 1

Progressive or regressive responses to strongly contradictory replication results and their expected impact on citation patterns for original studies.

Progressive responses	Regressive responses
1.00	1.00
2.00	2.00

Post-replication citation patterns in psychology: Four case studies

Methods

The study protocol (rationale, methods, and analysis plan) was pre-registered on April 7th 2018 (https://osf.io/eh5qd/). An amended protocol was registered part way through data collection on May 1st 2019, primarily because we extended the sampling frame to cover additional months (https://osf.io/pdvb5/). All deviations from these protocols are explicitly acknowledged in Supplementary Information A. All data exclusions and measures conducted during this study are reported.

Design

This was a retrospective observational study consisting of four case studies. Primary outcome variables were annual citation counts for original studies, citation valence (favourable, equivocal, unfavourable), co-citation of original and replication studies, frequency/type of counter-arguments.

Sample

We examined four case studies in which a prominent pre-registered and multi-laboratory replication study strongly contradicted and outweighed the findings of an original study (Table 2).

Table 2

Sample sizes and effect sizes for replication studies and original studies. d = Cohen's d; MD = mean difference; k = number of data collection sites; N = total number of participants; CI = confidence interval. Publication dates are earliest available (i.e., 'online first' where relevant).

Original Study	Replication study	Effect	Total citations to original*	Original sample size	Replication sample size	Original effect size [95% CI]	Replication effect size [95% CI]
Baumeister et al. (May, 1998)	Hagger et al. (July, 2016)**	Ego-depletion	1974	k = 1 N = 67	k = 23 N = 2,141	d = 2.05 [1.31, 2.79]	d = 0.04 [-0.07, 0.15]
Sripada et al. (April, 2014)	Hagger et al. (July, 2016)**	Ego-depletion	36	k = 1 N = 26	k = 23 N = 2,141	d = 0.68 [0.09, 1.27]	d = 0.04 [-0.07, 0.15]
Strack et al. (May, 1988)	Wagenmakers et al. (October, 2016)	Facial feedback	708	k = 1 N = 92	k = 17 N = 2,124	MD = 0.82 [-0.05, 1.69]	MD = 0.03 [-0.11, 0.16]
Caruso et al. (July, 2012)	Klein et al. (January, 2014)	Money priming	57	k = 1 N = 30	k = 36 N = 6,333	d = 0.8 [0.05. 1.54]	d = .01 [-.06, 0.09]
Carter et al. (July, 2011)	Klein et al. (January, 2014)	Flag priming	54	k = 1 N = 70	k = 36 N = 4,896	d = 0.50 [.01, .99]	d = .01 [-.07, 0.08]

* Total citations to the original study between the publication date and 31st December, 2019.

** For methodological reasons (see Hagger et al., 2016), the ego-depletion replication was aimed at a classic study in the field (Baumeister et al., 1998), but actually employed a modified computer-based version of the original paradigm (Sripada et al., 2014). We examined post-replication citation patterns for both studies.

Procedure

Annual citation counts. Citation histories (i.e., bibliographic records for all articles that cite the original study) from the publication date of each original study through to 31st December, 2019, were downloaded from Clarivate Analytics Web of Science Core Collection accessed via the Charité – Universitätsmedizin Berlin on 12th August, 2020. We also obtained citation histories for a reference class - all articles published in the same journal and the same year as each original study - from the same source. For example, for Baumeister et al. (1998) the reference class was all articles published in 1998 in the Journal of Personality and Social Psychology. Citation counts were standardized in each case study by setting the citation count in the replication year to the standardized value of "100" and then adjusting the counts in other years according to the same transformation ratio. For example, if the raw citation count in the replication year was 1000, citation counts in each year were standardized by dividing by 10. This computation was performed separately for the reference class and citations to the original article.

Qualitative assessment. Qualitative assessment of citation patterns was limited to a time period starting one year prior to the year of publication of the replication study up until 31st December, 2019, excluding the year in which the replication was published. We excluded the replication year because it may be unreasonable to expect citing articles already in the publication pipeline to cite the replication study. For the Baumeister case, the qualitative analysis was based on a random sample of 40% of citing articles from the pre-replication period and post-replication period due to the large number of citations to the original study (n = 1974; see Supplementary Information B for details).

For each citing article undergoing qualitative assessment, we attempted to retrieve the full text via at least two of the institutional libraries we are affiliated with. Inaccessible articles were excluded. For articles for which we could obtain the full text, we classified the research design according to the categories in Table 3 and recorded whether the replication study was cited after manual inspection of the reference section (see Table 1: citation balance/bias).

To examine the belief correction/perpetuation pattern (Table 1), the primary coder manually extracted the "citation context" of the original study and the replication study (i.e., all relevant verbatim text surrounding each in-text citation). The primary coder then classified the citation valence as "favourable", "equivocal", "unfavourable", or "unclassifiable". Favourable citations were those used to support a positive claim about the phenomenon of interest whereas unfavourable citations were used to support a negative claim about the phenomenon of interest. Citations were considered equivocal if the authors did not take a predominantly favourable or unfavourable position. Citations that did not endorse or oppose the phenomenon of interest (for example, simply referring to the procedures of the original study) were designated as "unclassifiable". Because this process was inherently subjective, the citation contexts and classifications were also examined by a secondary coder. Disagreements were resolved through discussion and a third coder arbitrated when necessary. Valence classifications by the primary coder were modified after discussion with the secondary coder in 30 (5%) cases.

To examine the explicit/absent defence pattern (Table 1), the primary coder flagged articles that co-cited the original and replication studies that also contained any explicit defence of the original study. Subsequently, two team members (ORA and SC) re-examined all of the flagged cases, extracted verbatim counter-arguments, and developed a post-hoc categorisation scheme that summarised them as concisely and informatively as possible (Table 5). Coding disagreements were resolved through discussion and a third coder (TEH) arbitrated when necessary.

In additional exploratory (not pre-registered) analyses, we examined overlap of authorship for articles that provided counter-arguments with (1) any of the authors of the original studies; and (2) any prior collaborators of the first authors of the original studies. These analyses are complicated by the fact that author names in bibliographic records do not always adhere to the same grammatical standards - for example, whether forenames are initialised or middle names are included - so it is not straightforward to isolate individual authors within bibliographic databases. In order to identify prior collaborators of the first authors of the original studies, we downloaded bibliographic records (on 2nd February, 2021) for all papers published by each of the original study first authors, according to their author record in the Web of Science Core Collection. These author records are automatically generated by an algorithm that attempts to identify all documents likely published by an individual author using several variations of their name (for example, "Hardwicke, Tom E.", "Hardwicke, Tom", "Hardwicke, T. E."), but errors can still occur and incomplete database coverage means that this method likely misses some of the authors' prior publications and thus some of their collaborators. Nevertheless, the method supports a reasonable lower-bound estimate of authorship overlap with articles providing counter-arguments. To identify authorship overlap, we used string manipulation tools in R to extract only author surnames from bibliographic records and then used string matching to automatically detect the presence of original author or collaborator surnames amongst the surnames of authors of articles that provided counter-arguments. When a match was detected, it was verified by manual examination of the authors' full names.

Results

In total, 2829 articles cited one of the original studies of which 632 articles (after taking a 40% random sample in the Baumeister case) fell within the time period designated for qualitative assessment. Of these 632 articles, we excluded 69 from the qualitative analysis because (1) we could not access the full text (n = 58); (2) they were non-English language (n

Table 3

Counts and percentages for research design classifications of articles included in qualitative analyses

Article type	Count (%)
Data synthesis - meta-analysis	11 (2)
No empirical data	163 (29)
Empirical data - case study	1 (0)
Empirical data - commentary including analysis	4 (1)
Empirical data - field study	39 (7)
Empirical data - laboratory study	245 (44)
Empirical data - multiple study types are reported	23 (4)
Empirical data - survey	77 (14)

= 6); (3) they included a citation to the original study in the reference section, but not in the main text (n = 4); or (4) manual inspection indicated that they did not actually appear to cite the original study at all (n = 1). Research design classifications for the remaining 563 articles included in the qualitative analysis are shown in Table 3.

Article characteristics

Annual citation counts and citation valence

Figure 1 shows standardized annual citation counts for each original study and the respective reference class (citations to all articles published in the same year and same journal as the original study), and classifications of citation valence (favourable, equivocal, unfavourable, unclassifiable or excluded). The data can also be viewed in tabular format in Supplementary Table B1. All counts (n) reported in the text and table are raw counts (i.e.,

not standardized).

After the replication was published, citations to the reference classes were continuing their trend to plateau (Baumeister case) or increase (other cases). By contrast, citations to the original study appeared to undergo a modest decline in the Strack case (decreasing from 56 to 41 between 2015 and 2019), and a small decline followed by a small increase in the Baumeister case (increasing from 191 to 199 between 2015 and 2019). In the other cases (Sripada, Carter, Caruso), the total citation counts were much lower and there was considerable variability in the post-replication citation patterns; nevertheless, there was no substantial change in annual citations from pre- to post- replication in these three cases (the maximum difference was +8 citations).

Prior to the replication, the vast majority of citations were favourable for all five articles (range 67% to 100%). In most cases (Strack, Sripada, Carter, and Caruso) there was a small post-replication increase in unfavourable citations and a small decrease in favourable citations, indicating a modest active correction pattern. However, the overall number of unfavourable citations was very low and there was still a substantial majority of favourable citations. For example, in the Strack case, unfavourable citations increased from 0% in the pre-replication period (2015) to 6% in the post-replication period, whilst favourable citations decreased from 82% to 71%. In the Baumeister case, the proportion of favourable citations remained stable from pre-replication (71%) to post-replication (73%), a pattern consistent with belief perpetuation. The very small number of unfavourable citations (2017: n = 7, 7.00%; 2018: n = 2, 2%; 2019: n = 2, 4%) suggests that this is largely an unchallenged belief perpetuation pattern (see Table 1).

Citation balance and citation bias

Figure 2 shows the proportion of citing articles that also cited or did not cite the replication study after it was published (excluding the publication year itself). The data can

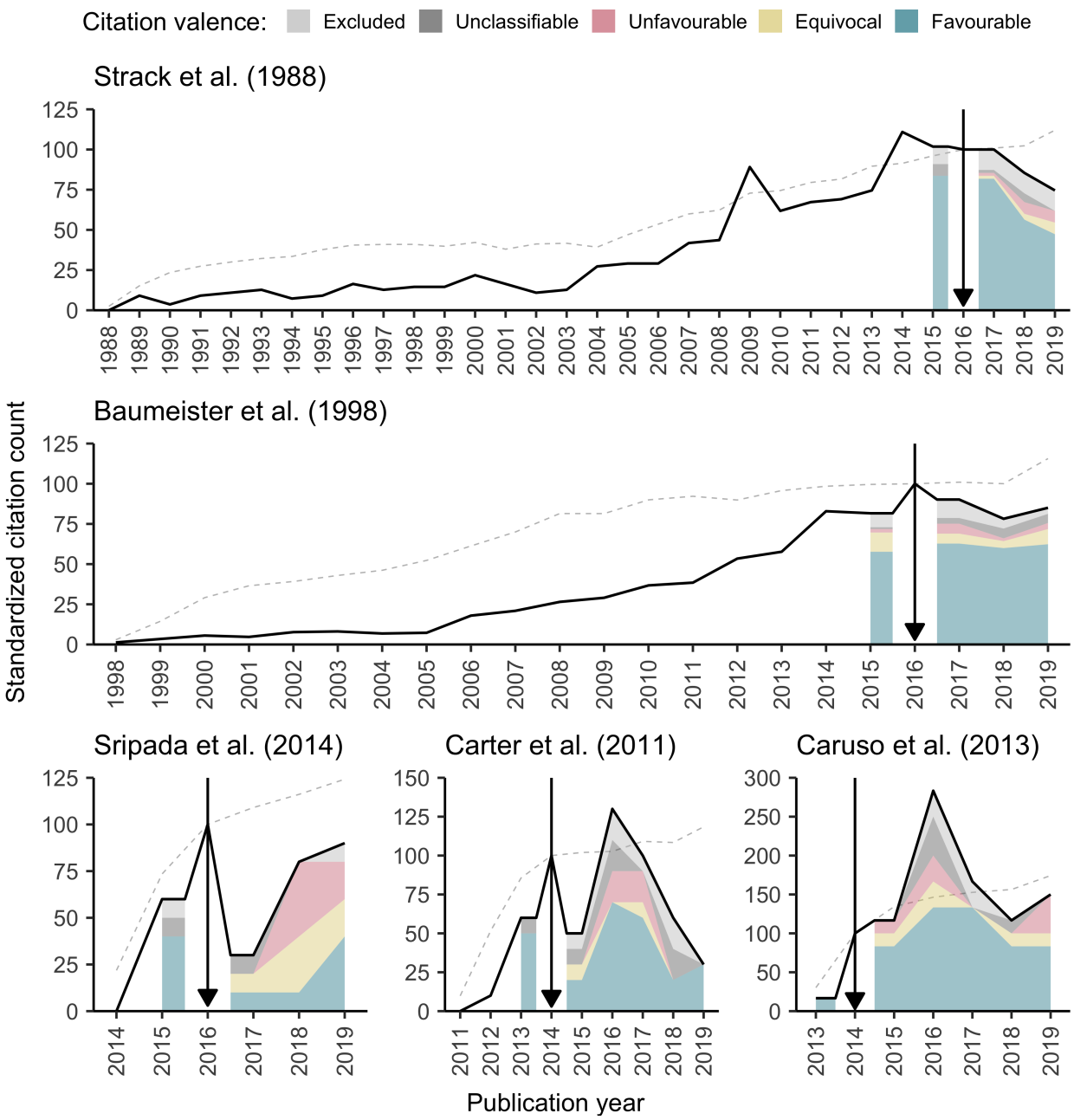

Figure 1. Standardized annual citation counts (solid line) for the five original studies with citation valence (favourable, equivocal, unfavourable, unclassifiable) illustrated by coloured areas in pre-replication and post-replication assessment periods. Dashed line depicts citations to the reference class (all articles published in the same journal and same year as the target article). Annual citation counts are standardized against the year in which the replication was published (citation counts in the replication year, indicated by a black arrow, are set at the standardized value of 100). Citation valence classifications for the Baumeister case are extrapolated to all articles in the assessment period based on a 40% random sample.

also be viewed in tabular format in Supplementary Table B1. In the Strack and Baumeister cases, a considerable majority of articles citing the original study did not cite the replication study, indicating substantial citation bias. In the Baumeister case the proportion of articles citing the replication study remained stable (20% in 2017, 16% in 2019). In the Strack case, the proportion increased from 13% to 39%. In the Carter and Caruso cases, the proportion never exceeded 40%, also consistent with substantial citation bias. In the Sripada case, it was much more common for the replication study to be cited (>88%) reflecting a balanced citation pattern.

Explicit defence and absent defence

Table 4 shows whether articles that cited the original study and replication study ("co-citing articles"), and the subset of co-citing articles that cited the original study favourably, provided any explicit counter-arguments to defend the credibility of the original finding (an explicit defence) or not (an absent defence). Overall, fewer than half of the 121 co-citing articles provided any counter-arguments. Of the 59 co-citing articles that cited the original study favourably, around half provided counter-arguments. We identified 57 discrete counter-arguments in 50 citing articles (45 of which were unique articles, as 5 of them were cited in two of the case studies) and allocated them to one of three categories (Table 5).

In additional exploratory analyses (not pre-registered) we examined other characteristics of the `d_contentAnalysis %>% filter(citesReplication==T,counterArguments==T) %>% distinct(doi) %>% nrow()` unique articles that contained counter-arguments. The articles were published in 34 individual journals, with *Frontiers in Psychology* publishing 7 of the articles, *Social Psychology* publishing 4 of the articles, and all other journals publishing only 1 or 2 of the articles. 17 of the articles did not involve empirical data, 3 involved reanalysis or meta-analysis of existing data, and 25 involved collection of novel data. The articles had 112 individual authors of whom all contributed to a single article except for 9 individuals who

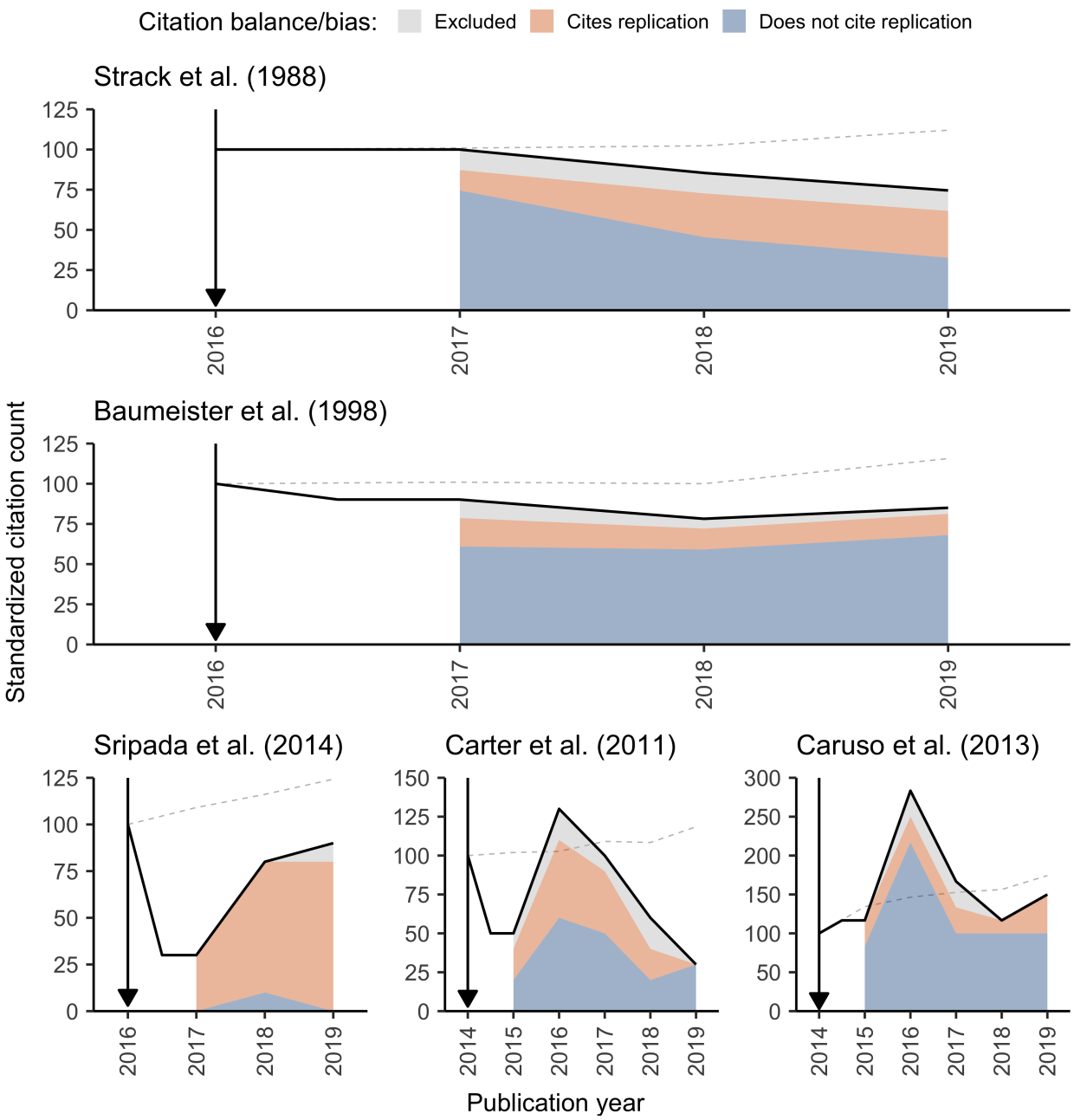

Figure 2. Standardized annual citation counts (solid line) for the five original studies with citation balance/bias (i.e., whether the replication is cited) illustrated by coloured areas in the post-replication assessment period. Dashed line depicts citations to the reference class (all articles published in the same journal and same year as the target article). Annual citation counts are standardized against the year in which the replication was published (citation counts in the replication year, indicated by a black arrow, are set at the standardized value of 100). Replication citation proportions for the Baumeister case are extrapolated to all articles in the assessment period based on a 40% random sample.

Table 4

Counts and percentages (in brackets) for whether articles that cited both the original study and replication study provided any explicit argumentation to defend the original study. Data are displayed for co-citing articles with any citation valence classification and the subset of co-citing articles with favourable citation valence classifications.

	All citation valences		Favourable citation valences	
case	No	Yes	No	Yes
baumeister	23 (0.55)	19 (0.45)	11 (0.46)	13 (0.54)
carter	10 (0.77)	3 (0.23)	3 (0.75)	1 (0.25)
caruso	6 (0.6)	4 (0.4)	1 (0.5)	1 (0.5)
sripada	9 (0.5)	9 (0.5)	2 (0.4)	3 (0.6)
strack	23 (0.61)	15 (0.39)	12 (0.5)	12 (0.5)
all cases	71 (0.59)	50 (0.41)	29 (0.49)	30 (0.51)

had (co)authored 2 articles. 3 articles were (co)authored by one of the original authors and 9 articles were (co)authored by at least one prior collaborator of one of the first authors of the original articles. 7 of these articles did not involve empirical data and 5 of them involved novel data collection.

Discussion

Open practices statement

The study protocol (hypotheses, methods, and analysis plan) was pre-registered on April 7th 2018 (https://osf.io/eh5qd/). An amended protocol was registered on May 1st 2019 (https://osf.io/pdvb5/). All deviations from these protocol or additional exploratory analyses are explicitly acknowledged. All data exclusions and measures conducted during

Table 5

Categorisation of counter-arguments provided to defend the original study in light of the contradictory replication result. 57 discrete counter-arguments were identified in 50 articles (45 unique articles across cases).

evidence	methods	expertise
11	44	2

this study are reported. All data, materials, and analysis scripts related to this study are publicly available on The Open Science Framework (https://osf.io/w8h2q/). To facilitate reproducibility this manuscript was written by interleaving regular prose and analysis code (TBA) using knitr (**???**) and papaja (**???**), and is available in a Code Ocean container (TBA) which re-creates the software environment in which the original analyses were performed,

Funding statement

Conflict of interest statement

The authors declare no conflicts of interest.

Author contributions:

References

\setlength{0.5in

Appendix A

Appendix A: Protocol Changes

Appendix B

Appendix B title

198 Tabular data for citation curves and citation valence

Table B1

Counts (percentages) for citations to the original study, exclusions, citation valence, and concomitant citations to replication study in the time periods subjected to qualitative analysis. Percentages for the Baumeister case are extrapolations based on 40% random samples drawn from the pre-replication and post-replication time periods.

year	citing articles	excluded	unclassifiable	favourable	equivocal	unfavourable	citesRep
Baumeister							
2015	191	8 (11%)	1 (1%)	53 (71%)	11 (15%)	2 (3%)	-
2017	211	13 (13%)	4 (4%)	71 (70%)	7 (7%)	7 (7%)	20 (20%)
2018	183	7 (8%)	7 (8%)	69 (77%)	5 (6%)	2 (2%)	15 (17%)
2019	199	2 (4%)	3 (7%)	33 (73%)	5 (11%)	2 (4%)	7 (16%)
Sripada							
2015	6	1 (17%)	1 (17%)	4 (67%)	0 (0%)	0 (0%)	-
2017	3	0 (0%)	1 (33%)	1 (33%)	1 (33%)	0 (0%)	3 (100%)
2018	8	0 (0%)	0 (0%)	1 (12%)	3 (38%)	4 (50%)	7 (88%)
2019	9	1 (11%)	0 (0%)	4 (44%)	2 (22%)	2 (22%)	8 (89%)
Strack							
2015	56	6 (11%)	4 (7%)	46 (82%)	0 (0%)	0 (0%)	-
2017	55	7 (13%)	1 (2%)	45 (82%)	1 (2%)	1 (2%)	7 (13%)
2018	47	7 (15%)	3 (6%)	31 (66%)	2 (4%)	4 (9%)	15 (32%)
2019	41	7 (17%)	0 (0%)	26 (63%)	4 (10%)	4 (10%)	16 (39%)
Carter							
2013	6	0 (0%)	1 (17%)	5 (83%)	0 (0%)	0 (0%)	-

2015	5	1 (20%)	1 (20%)	2 (40%)	1 (20%)	0 (0%)	2 (40%)
2016	13	2 (15%)	2 (15%)	7 (54%)	0 (0%)	2 (15%)	5 (38%)
2017	10	1 (10%)	0 (0%)	6 (60%)	1 (10%)	2 (20%)	4 (40%)
2018	6	2 (33%)	2 (33%)	2 (33%)	0 (0%)	0 (0%)	2 (33%)
2019	3	0 (0%)	0 (0%)	3 (100%)	0 (0%)	0 (0%)	0 (0%)
Caruso							
2013	1	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	-
2015	7	0 (0%)	0 (0%)	5 (71%)	1 (14%)	1 (14%)	2 (29%)
2016	17	2 (12%)	3 (18%)	8 (47%)	2 (12%)	2 (12%)	2 (12%)
2017	10	2 (20%)	0 (0%)	8 (80%)	0 (0%)	0 (0%)	2 (20%)
2018	7	0 (0%)	1 (14%)	5 (71%)	1 (14%)	0 (0%)	1 (14%)
2019	9	0 (0%)	0 (0%)	5 (56%)	1 (11%)	3 (33%)	3 (33%)

```
##         case             year citing\narticles  excluded unclassifiable
## 1 baumeister post-replication             593 22 (0.09)      14 (0.06)
## 2    sripada post-replication              20  1 (0.05)       1 (0.05)
## 3     strack post-replication             143 21 (0.15)       4 (0.03)
## 4     carter post-replication              37  6 (0.16)       5 (0.14)
## 5     caruso post-replication              50  4 (0.08)       4 (0.08)
##   favourable equivocal unfavourable  citesRep
## 1 173 (0.73) 17 (0.07)    11 (0.05) 42 (0.18)
## 2    6 (0.3)   6 (0.3)      6 (0.3)  18 (0.9)
## 3 102 (0.71)  7 (0.05)     9 (0.06) 38 (0.27)
## 4  20 (0.54)  2 (0.05)     4 (0.11) 13 (0.35)
## 5  31 (0.62)   5 (0.1)     6 (0.12)  10 (0.2)
```