Ankel-Peters, Jörg; Fiala, Nathan; Neubauer, Florian

Working Paper

Is Economics Self-Correcting? Replications in the American Economic Review

I4R Discussion Paper Series, No. 68

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Ankel-Peters, Jörg; Fiala, Nathan; Neubauer, Florian (2023) : Is Economics Self-Correcting? Replications in the American Economic Review, I4R Discussion Paper Series, No. 68, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/276962

INSTITUTE for REPLICATION

No. 68
I4R DISCUSSION PAPER SERIES

Is Economics Self-Correcting? Replications in the *American Economic Review*

Jörg Ankel-Peters

Nathan Fiala

Florian Neubauer

September 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 68

Is Economics Self-Correcting? Replications in the *American Economic Review*

Jörg Ankel-Peters[1], Nathan Fiala[2], Florian Neubauer[2]

[1]RWI – Leibniz Institute for Economic Research, Essen/Germany and University of Passau/Germany
[2]University of Connecticut, Storrs/USA and RWI – Leibniz Institute for Economic Research, Essen/Germany

SEPTEMBER 2023

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Is Economics Self-Correcting? Replications in the *American Economic Review*

BY JÖRG ANKEL-PETERS, NATHAN FIALA, & FLORIAN NEUBAUER*

September 2023

This paper reviews the impact of replications published as comments in the *American Economic Review* between 2010 and 2020. We examine their citations and influence on the original papers' subsequent citations. Our results show that comments are barely cited, and they do not affect the original paper's citations – even if the comment diagnoses substantive problems. Furthermore, we conduct an opinion survey among replicators and authors and find that there often is no consensus on whether the original paper's contribution sustains. We conclude that the economics literature does not self-correct, and that robustness and replicability are hard to define in economics.

Keywords: Replication, citations, meta-science

JEL Classifications: A11, A14, B40

* Ankel-Peters (corresponding author): RWI - Leibniz Institute for Economic Research and University of Passau (e-mail: peters@rwi-essen.de); Fiala: University of Connecticut and RWI - Leibniz Institute for Economic Research (e-mail: nathan.fiala@uconn.edu); Neubauer: University of Connecticut and RWI - Leibniz Institute for Economic Research (e-mail: florian.neubauer@uconn.edu). Funding: This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) [Grant No. 3473/1-1 within the DFG Priority Program META-REP (SPP 2317)]. We are grateful for valuable comments and suggestions by Krisztina Kis-Katos. We also thank research seminar and conference participants at University of Wageningen, University of Siegen, the German Development Economics conference in Dresden and the annual congress of the European Association of Agricultural Economics in Rennes. We are especially grateful to all AER editors since 1985 for responding to our short survey: Orley Ashenfelter, Ben Bernanke, Robert Moffitt, Pinelopi Goldberg, Esther Duflo, and Erzo Luttmer. We further acknowledge the participation of the following people in our survey of authors of original papers, comments, and replies in our AER sample: Armon Rezai, Art Fraas, Brock Smith, Christoph March, Frank Venmans, Gary Bolton, Jens Iversen, Jérôme Adda, Johannes Pfeifer, John William Hatfield, Jonathan Ketcham, Jordi Galí, Juan Rubio-Ramírez, Julian Wright, Karel Mertens, Kathryn Zeiler, Kurt G. Lunsford, Marcel Timmer, Matteo Cervellati, Miklós Koren, Paul Milgrom, Rafael Di Tella, Samuel Lundstrom, Sebastian Claro, Sweta Chaman Saxena, Uwe Sunde, Yuval Heller, Zhi Wang, Zhouxiang Shen as well as 90 authors who participated in our survey but preferred not to be acknowledged. Caitlin Chan, Rita Reischl, Youmna Eid, and Martin Buchner provided valuable research assistance.

1. Introduction

Karl Popper's falsificationism is the prevailing view on the scientific method in economics and replication is at the core of his philosophy of science (Popper, 1959). According to Merton (1973, p. 276), it is replication and "*rigorous[ly] policing*" each other's work that keeps scientists truthful and disinterested – often referred to as scientific self-correction (Peterson and Panofsky, 2021). In this paper, we examine whether replications in economics lead to a reappraisal of the replicated papers' influence on the literature. We do so by investigating whether replications are cited and whether they change the citations of replicated papers. The self-correction narrative would arguably imply changing citation patterns in response to an unsuccessful replication.

The need for replications in economics has been debated over the past decades (Clemens, 2017; H. M. Collins, 1991; Dewald et al., 1986; Hamermesh, 2007; Leamer, 1983; Mirowski and Sklivas, 1991; Whaples, 2006). More recently, the profession has experienced noteworthy improvements in terms of preregistration and data-sharing policies (Christensen and Miguel, 2018; Miguel, 2021), but replications are still rare (Ankel-Peters et al., 2023a; Mueller-Langer et al., 2019; Sukhtankar, 2017).[1] At the same time, new meta-evidence indicates various forms of replicability problems and thereby emphasizes the need for more replications (Askarov et al., 2022; Brodeur et al., 2016, 2020, 2023; Camerer et al., 2016; Chang and Li, 2022; Dahal and Fiala, 2020; Ferraro and Shukla, 2020; Ioannidis et al., 2017; Peters et al., 2018; Vivalt, 2020).

We focus on replications published as comments in the *American Economic Review* (*AER*), one of the profession's flagship journals. Comments in the *AER* challenge papers that were previously published in the journal, mostly based on robustness replications of that original paper (OP). The *AER* has always had a leadership role in

[1] Berry et al. (2017) provides a more optimistic replication rate. Low replication rates have also been discussed in other disciplines such as psychology and medical science (Hensel, 2021; Maxwell et al., 2015; Open Science Collaboration, 2015).

the profession by publishing a relatively large number of comments and, more recently, by rigorously applying data-sharing policies (AEA, 2023). We therefore start by tracing the publication of comments over time and by showing the results of a short survey we conducted among the *AER* editors since 1985.

In the main part of our paper, we investigate all comments published between 2010 and 2020 and find 56 comments, of which 37 received a reply by the original authors. We use Google Scholar (GS) citations to examine how often OPs and comments are cited and whether the citation trend of the replicated OP has changed after the comment was published. Our underlying assumption is that citations reflect the priors in the research community (Rubin and Rubin, 2021) and the community's appreciation of the publication (Bornmann and Daniel, 2008; Card and DellaVigna, 2020; Hamermesh, 2018; Siler et al., 2015). As we will show, some of the OPs in our sample are cited heavily. This suggests that they have had a strong influence on the thinking, hence the prior, in the literature (Teplitskiy et al., 2022).

For the self-correction claim to hold, we hypothesize that a comment should lead to a *strong* reaction of the literature, especially for a comment raising substantive concerns about an OP. If it does not respond strongly, we argue, the prior in the literature sustains. We look at two facets of a strong response: 1) Citations of the comment relative to citations to the OP after comment publication (henceforth: citation ratio), and 2) Whether the comment affects the OP's annual citations. In our main analysis, we do not conduct formal statistical testing and rather subject these two indicators to a descriptive analysis. Effective self-correction should either lead to a very high citation ratio because the comment is cited most of the time when the OP is cited (Coffman et al., 2017; Hardwicke et al., 2021) or to a clear and discernable effect on the OP's annual citations.

For this, we visually inspect the OP's annual citations before and after the publication of the comment. If there is no visibly discernable decline in citations, especially for substantive comments, we reject the hypothesis of self-correction. Note that whenever

we use causal expressions such as 'effect', 'impact', or 'influence', we refer to this visual inspection, not a quantitative analysis. We believe this approach does justice to our very generic research question about scientific self-correction, as well as to the nature of our sample, which is small but contains very influential OPs and highly published comments. We argue that if there is no discernable effect on the literature for comments that made it into the *AER*, it is unlikely to materialize for other replications.

This part of our paper is similar to and complements Coupé and Reed (2022) who examine the effect of 204 replications in economics on citation patterns of replicated papers, using econometric analysis. Coupé and Reed include peer-reviewed replications from The Replication Network (The Replication Network, 2023). They quantitatively estimate a counterfactual citation trend in the absence of a replication by matching replicated and non-replicated papers, most importantly based on the pre-replication citation trends. We complement Coupé and Reed (2022) by probing deeper into a smaller number of prominent replications, while their work benefits from a larger and arguably more representative sample.

We find that *AER* comments do not affect the OP's citations and hence their influence on the literature. We observe an average citation ratio of 14%. Comments are cited on average seven times per year since their publication – compared to an average of 74 citations per year for the OP since publication of the comment. Comments are, hence, not cited much in absolute terms, and a lot less than the OP. The latter implies that most OP citations ignore the comment. This issue has been discussed by Coffman et al. (2017) who call for a normative change towards citing the replication next to its OP, which would also ensure that "well-executed replications receive credit". We furthermore find that the publication of a comment does not affect the OP's citation trend. These findings confirm Coupé and Reed (2022), who, likewise, do not find what they call a "penalty" of replications on post-replication citations.

Not all citations on GS are of equal quality and perhaps self-correction mechanisms are more effective in academically more prestigious circles. We therefore also focus on influential high-quality citations by zooming into how OPs and comments are cited in the *Journal of Economic Literature* (*JEL*), the *Journal of Economic Perspectives* (*JEP*), and the *AER* itself. We found 205 references in 192 *AER*/*JEL*/*JEP* papers to 43 OPs from our sample. We corroborate our previous findings in that most OPs are cited in the *AER*, *JEL,* and *JEP* papers without mentioning the comment alongside it (only 41 out of 205 cite the comment). This finding also holds for the other Top 5 journals.

In a next step, we address the fact that not all comments raise equally substantive concerns. Thus, the need for scientific self-correction differs across comments. We therefore read and rated all comments as to whether the respective comment, in our view, must be cited (against the alternative options 'sometimes' and 'never'). In fact, some comments are confirmative or criticize only parts of the OP, without questioning the OP's key contribution. Others fundamentally challenge the OP's main claims. Especially this latter group should not be ignored by the literature if the self-correction paradigm is to hold. Moreover, we conducted a short survey among authors of OPs and comments to elicit whether they believe the respective comment must be cited alongside the OP. Using their assessment of the debates as well as ours, we test for the robustness of our results when focusing on those comments that are rated as must-cites or sometimes-cites. We find that our interpretation holds even for these cases.

Our paper contributes to several literatures. We build on the growing meta-scientific literature in economics, which diagnoses increasing transparency standards (Christensen and Miguel, 2018; Miguel, 2021). Furthermore, we contribute to attempts in economics and other disciplines to shed more light on how replications are received in the scientific community (Coupé and Reed, 2022; Hardwicke et al., 2021; Schafmeister, 2021; Serra-Garcia and Gneezy, 2021; von Hippel, 2022), whether self-correcting mechanisms in science work (Ioannidis, 2012; Peterson and Panofsky, 2021; Vazire and Holcombe, 2021) and how the scientific community deals with

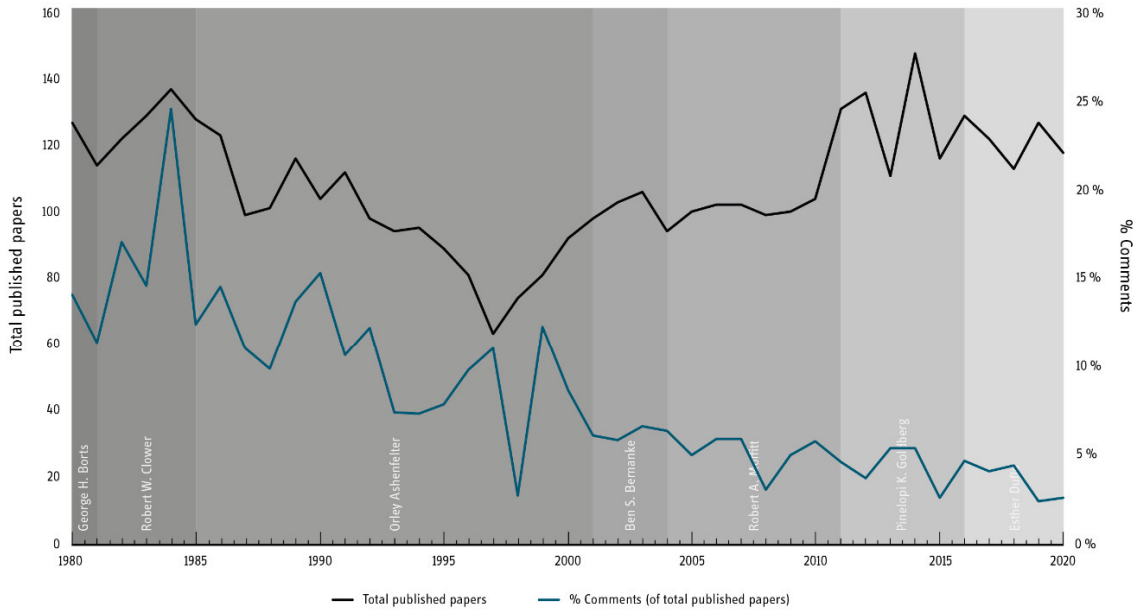
demonstrably erroneous papers (Budd et al., 1998; Fernández et al., 2019; Malkov et al., 2023).

2. Replication and comments: The *AER* policy

The *AER* has a long history of publishing comments and has a clear guideline for what comments are supposed to do and how they are handled (AEA, 2023):

> *"Comments submitted to the Review are refereed both by the author of the article being commented on and by other referees. Replies to Comments are sent to the author of the Comment and to other referees. There is no automatic right to Reply; the author of a Reply must provide substantive and material discussion of the issues in question. Comments and Replies which appear only on the AER web page are also sometimes considered.*[2] *These papers go through the same refereeing process as all Comments and Replies, but may be judged to be more appropriate for internet posting instead of publication in the printed AER."*

Figure 1: Papers and comments published in the *AER* between 1980 and 2020



Notes: Gray-shaded areas indicate the periods of AER editors-in-chief. Comments are elicited for each volume. Source: Own data.

Throughout the 1980s and into the 1990s, more than 10% of the papers published in the *AER* were comments (Figure 1). This share has declined considerably over time, down to 2-3% in most recent years. Ankel-Peters et al. (2023a) investigate this in more

[2] We checked the *AER*'s "archived internet comments" but only found one, published online in 2006 during Robert Moffitt's term (accessed on July 11, 2022).

detail and find that in the 1980s most comments were theoretical. Since then, the share of empirical papers in the *AER* has increased considerably, which has not been accompanied by a simultaneous increase in empirical comments.

We conducted a short survey among the *AER*'s editors since 1985 to obtain a better understanding of the *AER* comment policy over time. Thankfully, all editors responded to our three-question survey, and all agree that there has not been an explicit change in the policy regarding the publication of comments.[3] The following quotes are interesting attempts to explain the development over time:

Orley Ashenfelter:

> *The comment/reply format is a tedious one to referee. In addition, I believe that in tenure decisions comments do not receive the same 'points' as other publications. The change to judging publications on a point system probably started in the 1980s. This clearly reduces author incentives to write comments.*

Ben Bernanke:

> *It's a little surprising that there are fewer comments now because 12 AER issues plus four field journals mean a lot more available space. Maybe what used to be comments are now more likely to be expanded and accepted as regular papers. Economic Insights, another new journal, also publishes shorter papers.*[4]

Robert Moffitt:

> *I know that many people feel, today, that submitting a comment has the problem that the author will almost surely reply if the comment is negative, and that will generate many months of back-and-forth debate with the original authors. […] I would not be surprised if many people also don't think a comment is particularly strong on a c.v. I suspect that many people feel it is just better to write a new, original paper which explicitly or implicitly criticizes the original paper, than submit a comment.*

Pinelopi Goldberg:

> *One hypothesis is that the published research has become both more complicated and more rigorous. There are many complaints about the*

[3] The complete answers of all editors are in Section F of the Online Appendix.

[4] Indeed, in 2019, the *AEA* launched a new journal, the *American Economic Review: Insights (AER:I)*. However, the *AER:I*, has not published any comment since its inception. What is more, in a companion paper, we have searched for replications in the *AER* and other top journals and only found a small number that qualify as replications but are not published as comments (Ankel-Peters et al., 2023a).

> *increasing length of published papers […]. Perhaps the flip side of longer papers is that they cover more ground, provide more robustness checks, and leave fewer open questions.*

Esther Duflo:

> *We publish comments when the point made is of significant interest for the general readership, so either when results in very influential papers are overturned or there is a methodological contribution in the comment (and the comment is correct as far as we can see).*

The *AER* is also a forerunner in terms of code and data-sharing policies. It is important to demarcate the replication work that *AER* comments are based on (to the extent they are empirical, see Section 3.1) from the newly established replications conducted by the *American Economic Association* (*AEA*) data editor (Vilhuber, 2019). For this demarcation, we refer to the nomenclature of different replication sub-types defined by Dreber and Johanneson (2023) and the Institute for Replication (2022, see Table 1).[5]

Table 1: Replication definitions

Author(s), Year	Category	New paper uses the same…		
		Specification	Population	Sample
Institute for Replication;	Computational Reproduction	✓	✓	✓
Dreber and Johannesson (2023)	Recreate Reproduction[1]	✓/✗[2]	✓	✓
	Robustness Replication	✗	✓	✓
	Direct Replication	✓	✓/✗[3]	✗
	Conceptual Replication	✗	✓/✗[3]	✗

Notes: [1]Dreber and Johanneson (2023) introduce this additional category which differs from "computational reproduction" only in that it emphasizes the usage of raw data and not having the analysis code of the OP. This category is not included in the *I4R* definition. [2]The specification in the reproduction is not always identical to the OP as the replicator does not have access to the original code but tries to recreate the analysis based on the given information in the OP. [3]*I4R*'s definitions of direct and conceptual replication only require new data but it does not matter if it is from the same population or not. Dreber and Johanneson (2023) further subdivide between the same, similar, and different populations.

Already in 2005, during Robert Moffitt's tenure, the *AEA* launched and implemented a new policy that made data sharing mandatory. Moreover, in 2018, they appointed a data editor to rigorously enforce the policy by conducting what Dreber and Johanneson (2023) call a *computational reproduction* on every accepted paper. The main purpose of this policy is to check whether the data and code are accessible and

[5] Similar definitions with different nomenclatures exist, see for example Clemens (2017), Freese and Peterson (2017), and Hamermesh (2007). They all share very similar dimensions to distinguish different sub-types, that is, according to whether the replication uses the same specification, population, and sample. See also Ankel-Peters et al. (2023a) for a more detailed review.

complete and to ensure that the code reproduces the results (Vilhuber, 2019). Virtually all comments in our sample, if empirical, are based on *robustness replications*, *direct replications*, or *conceptual replications*. Perhaps closest to a *computational reproduction* are a few comments that find coding errors in the original study, but it is unlikely that the reproductions conducted as part of the *AEA*-checks would have uncovered the deeper coding issues that underlie these comments. Nevertheless, this rigorous data and code-sharing policy might have signaling effects altering the incentive structure towards transparency and replicability (see e.g., Askarov et al., 2022).

3. Citation patterns for comments and original papers

We conducted a systematic review covering eleven volumes of the *AER* from 2010 until 2020 and screened the *AER* website for all papers that included the word "comment" in the title. In total, we found 56 comments, written on 53 OPs[6], 37 of which also received a reply from the original authors. For every OP, comment, and, if applicable, reply – henceforth a *debate* – we elicited the number of citations in GS.[7] We use the average annual citations since the publication of the comment in the *AER* as the main citation indicator throughout the paper to ensure the comparability of citation counts between OP and comment.

3.1 Descriptive statistics on authors and original papers

Table 2 shows some author characteristics to examine whether comment authors differ from original authors in terms of career status and influence. Authors of OPs are more senior and more influential than comment authors (as measured by top five[8]

[6] Note that two OPs received more than one comment (Andreoni and Sprenger, 2012; Long and Ferrie, 2013). We also want to mention that the comment by Rothstein (2017) replicates two papers: Chetty et al. (2014a) and Chetty et al. (2014b). After successfully replicating both papers, though, the comment focuses on the former. Thus, we only include Chetty et al. (2014a) in this list of OPs underlying our analyses.

[7] This task was done between February 15 and 19, 2022. We also considered citations in Web of Science (WoS) but recent studies report a very high correlation between GS and WoS metrics – above 90% for economics (Hamermesh, 2018; Martín-Martín et al., 2018).

[8] The top five journals in economics are: *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *The Quarterly Journal of Economics*, and *The Review of Economic Studies*.

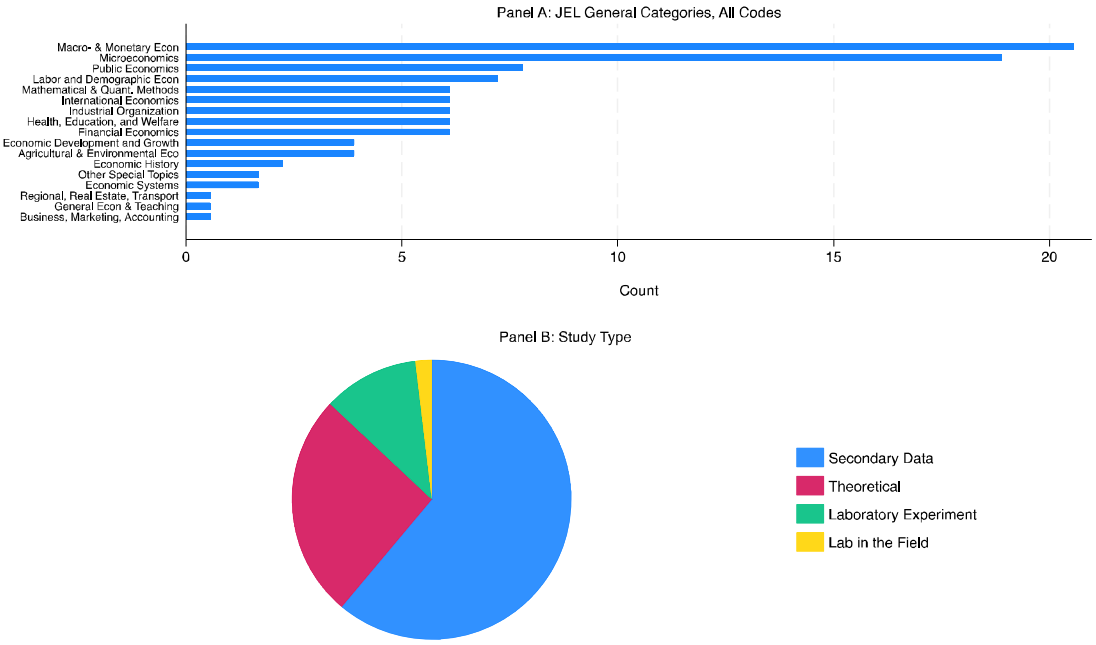
publications and GS citations). Some comment authors are likewise established researchers, but at the time of writing the comment, many were at the beginning of their careers. (See Tables C3-C5 in the Online Appendix for more author features; location, field of work, and highest obtained degree).

Table 2: Author characteristics

	OP authors		Comment authors	
N (Number of papers)	53		56	
N (Number of authors)	117		111	
Size of author team (average)	2.21		1.98	
	Team average	Maximum per team	Team average	Maximum per team
Number of years since PhD in t-1[1,2]	15.3	20.0	9.0	13.6
# of top five publications up to t-1	6.0	10.3	1.4	2.3
Share of authors with <100 GS citations in t-1	0.22	0.47	0.63	0.84
GS citations in total	16,126	29,930	5,471	10,739
GS citations in 2021	1,426	2,675	463	855
GS citations in t-1	855	1,492	277	542

Notes. [1] t-1 is the year before comment publication. [2] 97% of authors have a PhD. The number of observations is different for the indicator 'Average number of years since PhD' due to missing data (OP: 110, Comment: 108). GS = Google Scholar.

Figure 2: *JEL*-codes and methods of original papers (OPs)



Notes. Panel A shows all *JEL* codes of all 53 OPs in our sample (multiple JEL categories per paper possible; N=180). Panel B is on a per-paper basis, i.e., N=53. Panel A is based on *JEL* codes, while panel B is based on our own judgment. Whittington et al. (1990) does not have any *JEL* codes and is, therefore, not included in panel A. Panel B includes one paper that collected primary but non-experimental data (Bonjour et al., 2003). We include it in the "Secondary Data" category. We coded all non-empirical papers as "Conceptual".

The OPs in our sample cover a broad range of topics, but the *JEL*-code '*macro- and monetary economics*' dominates (Figure 2, Panel A).[9] Most OPs use secondary data (33 OPs), some also laboratory experiments, and a few are also non-empirical, that is theoretical or methodological. There is no Randomized Controlled Trial.

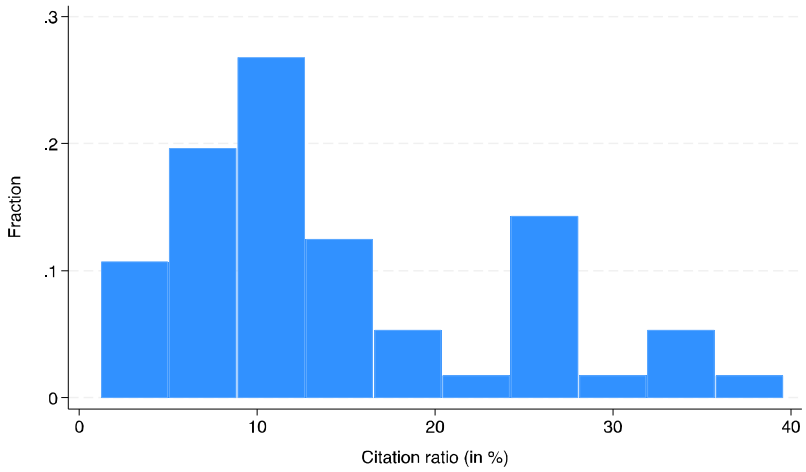
3.2 How much are comments cited vis-à-vis the original paper?

In this section, we examine the average annual citations since the comment's publication for comments and OPs and the citation ratio of comment to OP.[10] Figure 3 shows the distribution of citation ratios. This is a key indicator, because to maintain the claim that economics is self-correcting, one would expect a high ratio. The average citation ratio across all debates is 14%. Looking at the distribution in Figure 3, 50% of the debates have citation ratios of less than 11%. No citation ratio is higher than 40% and only four of the 56 comments in our sample have an average annual citation ratio above 30% - of which all associated OPs are from the bottom half of Figure 4, i.e., they are among the least cited: Brunner et al. (2011) at 39.6% (OP: Selten and Chmura, 2008), Mattauch et al. (2020) at 35.7% (OP: Lemoine and Rudik, 2017), Caselli and Ciccone (2019) at 34.0% (OP: Jones, 2014), and Crump et al. (2011) at 32.3% (OP: Whittington et al., 1990). Hence, for most debates, the comment is largely ignored by the future literature.

Taking the difference in author prestige into account that we discussed in Section 3.1, we explore debates for which the gap flips in favor of the comment authors.[11] We find that for these 15 debates the average citation ratio increases to 22% but is still on a low level.[12]

[9] The *Journal of Economic Literature* (*JEL*) codes are used to classify literature in economics.
[10] Table A1 in the Appendix (at the end of this paper) provides comprehensive descriptive statistics for all OPs and comments.
[11] We specifically investigate the group of debates for which the comment author team exceeds the OP author team in at least one of the following four variables: 1) Average number of top five publications of the author team, 2) Maximum number of top five publications of a member of the author team, 3) Average citations of the author team, and 4) Maximum citations of a member of the author team.
[12] See Figure D12 for a scatter plot of all citation ratios, separated by author prestige, and Table D7 for citation statistics.

Figure 3: Citation ratio distribution (in %)



Notes. Citations are counted since comment publication for each debate.

Figure 4 shows the average annual citations of the OPs (since the publication of the comment) and their respective comments.[13] The difference between citations of OPs and comments is large: Total average citations are 15 times higher for the OPs than for comments, but also the average annual citation count of OPs since comment publication is more than ten times higher (see Table A1 at the end of this paper). Most OPs are influential papers with total citation counts way above the *AER* average, but some OPs also have low citation counts. The debate between Acemoglu et al. (2001) and Albouy (2012) is a striking outlier: While the comment is the second most cited comment, the OP dwarfs its citation count. Replies are even less cited with an average of 30 total citations (compared to 56 total citations of comments and 567 of OPs). Here again, the reply by Acemoglu et al. is an outlier with more than 200 citations.

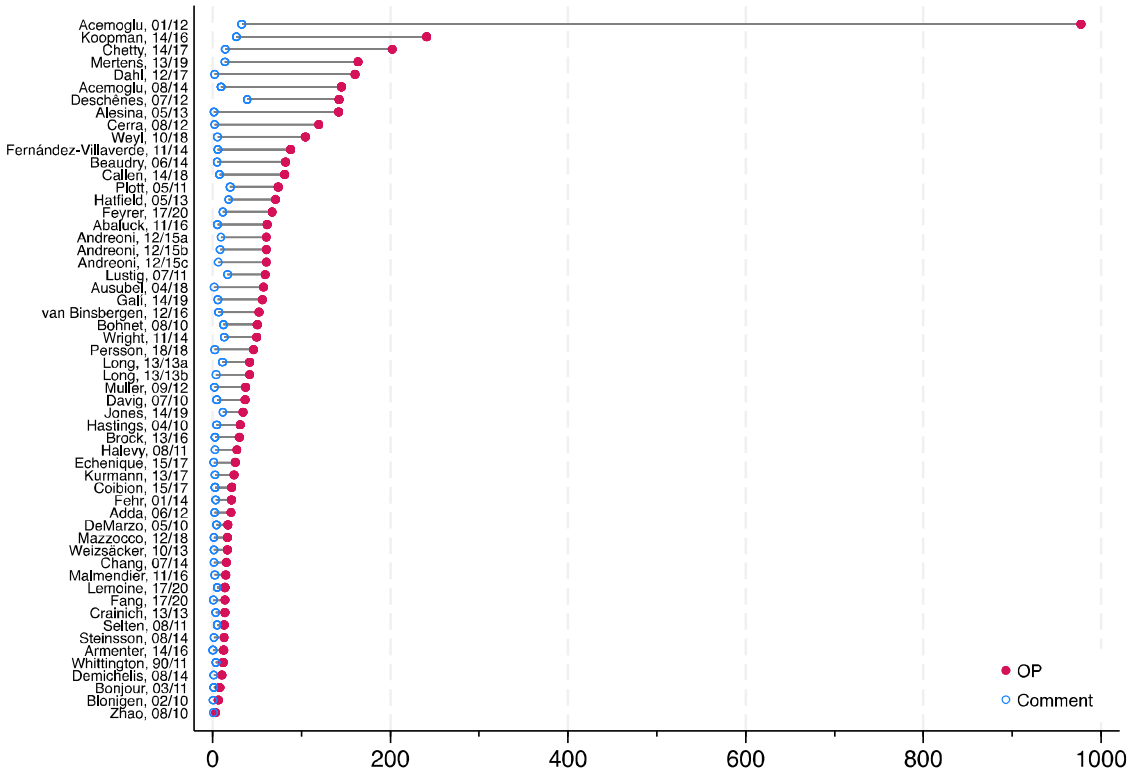
While it seems that comments are largely ignored by the general literature, it may be that at least original authors consider the comment when citing their own OP in their subsequent work. We check this for all 117 original authors[14] in our sample and find

[13] Only four of the 56 comments have at least ten citations in total prior to their *AER* publication, probably because discussion paper versions had circulated before: Albouy (2012), Burnside (2011), Cheung (2015), and Fisher et al. (2012).

[14] While there are 113 unique authors in our sample, four of them authored two OPs: Acemoglu, Johnson, and Robinson co-authored both Acemoglu et al. (2001) and Acemoglu et al. (2008); Sprenger co-authored Andreoni and Sprenger (2012) and Callen et al. (2014). We therefore report results for 117 debate-author pairings.

that 68 of them cite their OP after the comment publication in a total of 190 papers. Only 22 of those papers (or 12%) also cite the comment.

Figure 4: Average annual citations – difference between original paper and comment



Notes. We included all comments published in the *AER* between 2010 and 2020 and their respective OP. Citations are counted since comment publication for each debate. OPs have red markers; comment markers are blue. The labels on the y-axis show the first author of the OP, its publication year, and the publication year of the comment. The OP by Andreoni and Sprenger (2012) received three comments that we mark by the letters "a" (Miao and Zhong, 2015), "b" (Cheung, 2015), and "c" (Epper and Fehr-Duda, 2015). Similarly, the paper by Long and Ferrie (2013) received two comments that we mark with "a" (Xie and Killewald, 2013) and "b" (Hout and Guest, 2013).

In principle, citations could also dismiss the referenced paper's content. To check this, we briefly investigate *how* the OPs are cited by using the scite.ai tool *Reference Check*, which classifies citation statements in referencing papers into 'supporting', 'mentioning' or 'contrasting' the referenced paper (with 'unclassified' as a fourth option in case the tool cannot assign the statement to one of the three categories).[15] We find that almost all citation statements about OPs in our sample (93.8%) are

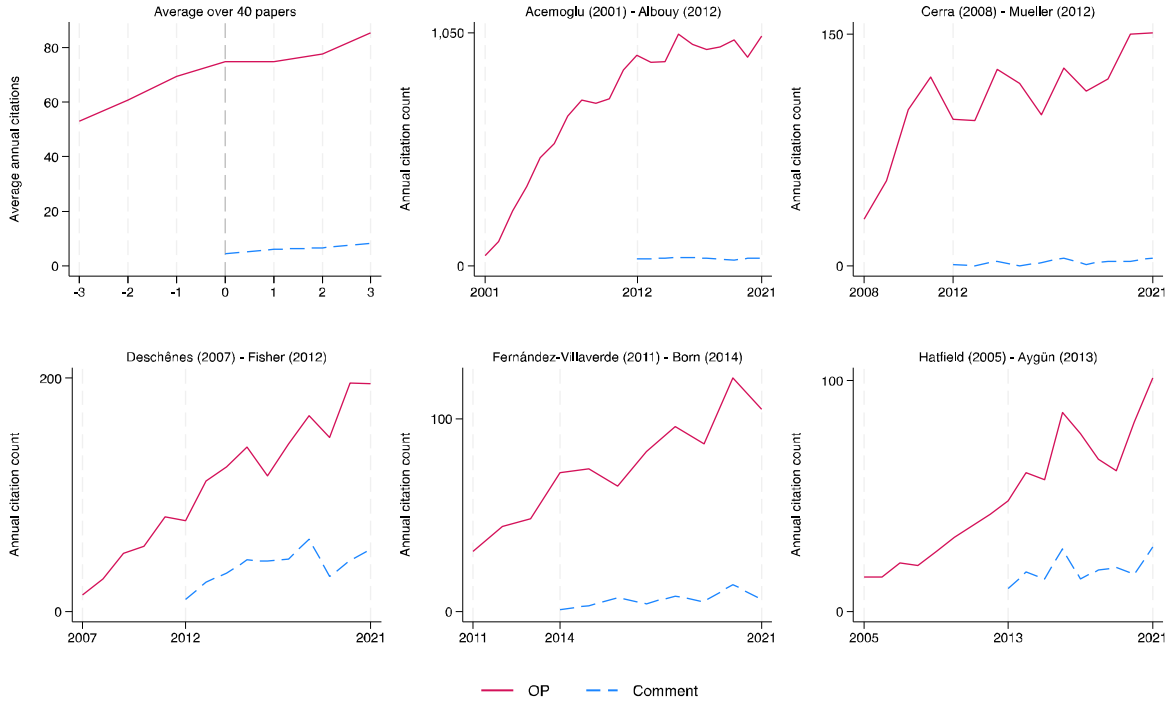
[15] One OP, Whittington et al. (1990), could not be included in the scite.ai analysis because a) it is not in their databank and b) scite.ai was not able to analyze the paper based on the uploaded PDF.

categorized as 'mentioning'. Further, 4.4% are 'supporting', only 0.7% are 'contrasting', and 1% are 'unclassified'.

3.3 Do comments affect citation trends of original papers?

We now investigate whether the publication of a comment leads to a reappraisal of the OP's influence in the literature, measured by the citation count of the OP. A transparent way to do this, given the limited number of observations in our sample, is via visual inspection of citation trends.[16] This is a suitable approach since we are interested in a sharp decrease of citations after publication of the comment, not in subtle differences, because, as we argue, such a sharp decrease is needed for the self-correction assumption to hold.

Figure 5: Original Papers' and comments' average annual citation counts before and after comment publication



Notes. For each panel, the blue line depicts the annual citation count of the OP, and the red line depicts that of the comment. In the first panel, 0 is the year of comment publication in the *AER*. The number of observations is lower than the full sample since the figure only includes debates for which a) the comment and OP were published at least three years before 2021, and b) the comment was published at least three years after the OP. For the two OPs that received multiple comments, we added the citations of the comments and included them in the calculation of the average together. Treating three comments as one is conservative in so far as a citing paper "corrects" the OP if at least one comment is cited. In panels 2-6, the first paper in the title is the OP, and the paper mentioned second is the comment. In both cases, we only depict the first author for space reasons.

[16] See Coupé and Reed (2022) for an econometric analysis of a larger sample of replications and OPs.

The publication of comments does not lead to a decrease in OP citations. What is more, most OPs (46 or 82%) have higher average annual citation counts after the comment than before. The first panel of Figure 5 shows the average annual citations of OPs and comments in the three years before and after the comment publication (t0): The positive trend of rising citations for the OP continues after t0. We test whether this observation holds for different time horizons before and after t0, and it does (see Figure D1 in the Online Appendix). The caveat of this robustness test is that the sample size of eligible comments decreases quite considerably the wider the sample period is, i.e., the more years before and after t0 we include.

It is likely that some comments had circulated as discussion papers before they were published in the *AER*. Yet, while 30 of 56 comments have been cited prior to their publication in the *AER* (probably as discussion paper versions), only four received more than 10 citations prior to publication. Moreover, most citation counts of OPs increase over time; thus, it is unlikely that the discussion paper versions had a strong effect on OP citations.

We also scrutinize individual debates and their citation trends. Figure 5 shows a selection of these, and Figures D3-D6 in the Online Appendix comprehensively depict all debates. Most debates exhibit an increasing citation trend. That is, OPs are cited more each year. For a few cases, the trend seems to stagnate post-comment publication – and we cannot rule out that this is due to the comment. However, even in these stagnating cases, the OPs keep on collecting the same number of citations every year.[17]

As we have emphasized, it is not our intention to make a precise causal statement of how much a comment affects the OP's citation trend. Yet, we might miss relative decreases of OP citation counts vis-á-vis similar non-replicated papers that are induced by the comment's publication. To examine this, we select the 20 most cited

[17] To explore whether an increase in source documents in GS is responsible for the non-declining citation trends, we collected data on citations from WoS, and find no systematic difference between the two search engines (see also Hamermesh, 2018, and Martín-Martín et al., 2018). Moreover, the fact that our results are consistent across the different years of comment publication underpins this.

OPs in our sample and approximate each OP's counterfactual trend by plotting the citation counts for all *AER* papers published in the same issues (see Figures D7 and D8 in the Online Appendix). We find that the citation trends of the OPs follow a similar pattern compared to their respective same-issue papers, which reassures us that the counterfactual trend of replicated OPs would not have looked much different in the absence of a comment.[18]

Furthermore, we quantitatively probe into post-comment citation trends of those OPs in our sample for which there are three years of pre- and post-comment citation data (40 OPs). For each OP, we plot the citation trends before the publication of the comment against the trend post-comment publication and test for a) a flattening of the slope post-comment publication (H0: $\beta 2 \geq \beta 1$), and b) a downward-sloping post-comment citation trend (H0: $\beta 2 \geq 0$). Figures D9-D11 in the Online Appendix show that there is a statistically significant break in the citation trend for 23 OPs, albeit the post-comment slope is significantly negative for only five. One might interpret this as a weak signal of correction, yet these quantitative robustness checks confirm that for most OPs, citations continue to rise or remain constant.

3.4. Citations in the *American Economic Review*, *Journal of Economic Literature* and *Journal of Economic Perspectives*

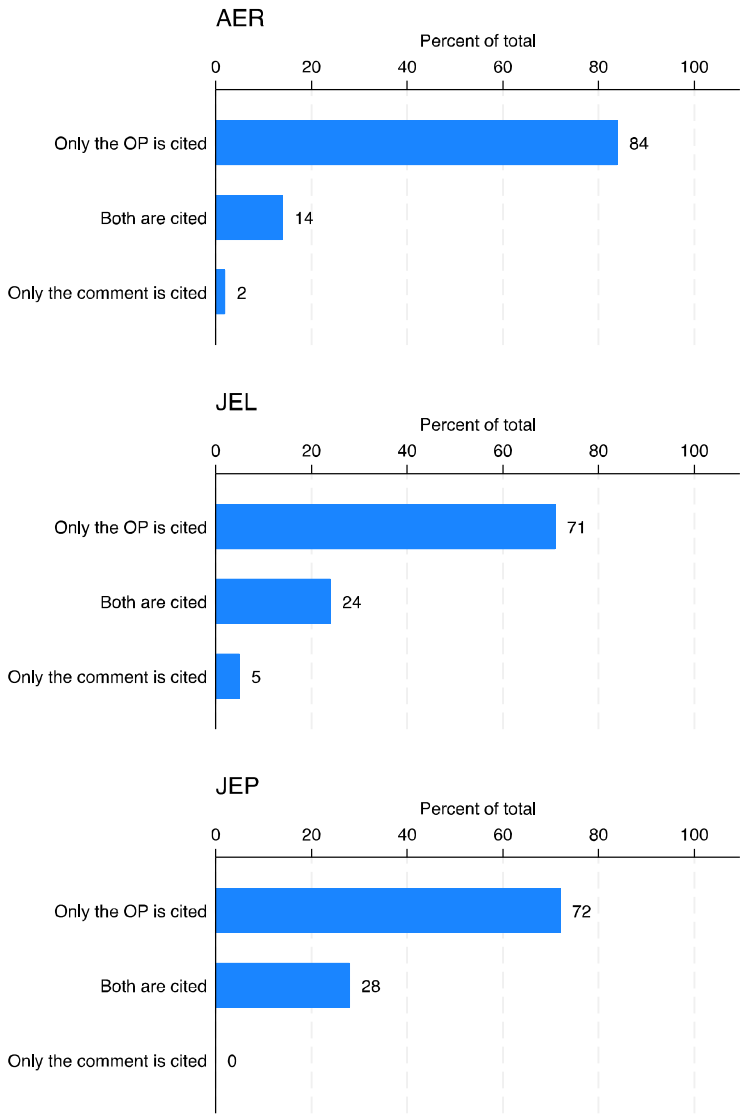
We now zoom into citations of particularly high influence. We focus on three journals for which one can argue that they are written with particular diligence and citations are made carefully: The *AER*, *JEL*, and *JEP*.[19] The *JEL* and the *JEP* are important review journals that summarize the state of affairs in the literature and especially the *JEL* arguably represents the highest standard in economics. We will show that while the

[18] See also Coupé and Reed (2022) for matching-based comparison of confirmatory and contradictory replications and their effect on citations of replicated papers.

[19] We did the analysis for these three journals for our full sample. In the Online Appendix, we show the results of the same analysis for a random sample of 15 debates for the other top five journals: the *Quarterly Journal of Economics*, *Econometrica*, the *Journal of Political Economy*, and *The Review of Economic Studies* (see Figure E13 and Table E10).

citation ratio in *AER* papers is similar to the one shown in Section 3.3, the citation ratio in *JEL* and *JEP* papers is indeed higher, but the majority of such high-quality citations still ignore the comment.

Figure 6: Share of *AER*, *JEL*, and *JEP* papers citing either the OP or the comment, or both



We searched WoS for all *AER*, *JEL*, and *JEP* papers (excluding book reviews and eulogies) that cite our sample of OPs and comments and found 192 individual papers, of which 140 are published in the *AER*, 35 in the *JEL*, and 17 in the *JEP*.[20] Combined,

[20] A few OPs are cited in multiple *AER*/*JEL*/*JEP* papers; Acemoglu et al. (2001), for example, is cited in 12 *AER*/*JEL*/*JEP* papers.

they cite 43 of our debates. Some of these 192 *AER*/*JEL*/*JEP* papers cite more than one OP, which is why we have 205 references in our dataset.[21]

Figure 6 shows that between 71% and 84% of the references in each journal cite the OP without mentioning the comment, and between 14% and 28% cite both the OP and the comment. The journal specific citation ratios are at 17% for the *AER*, 28% for the *JEP* and 31% for the *JEL*, versus 14% in Section 3.3 (not shown in Figure 6).

We also dug deeper into the quality of citations in the *AER*/*JEL*/*JEP* papers as well as the remaining top five journals and find that most citations to the OP are very prominent, including those that ignore the comment. The results can be found in Tables E9 and E10 in the Online Appendix. Even if the comment is cited next to the OP, it is rarely mentioned explicitly that there is a debate. This underpins that the OPs continue to have a profound impact on the thinking in the profession.

4. Subjective ratings and author survey

4.1 Subjective ratings

Not all comments put forward equally fundamental criticism of the OP. Comments diagnosing (or claiming to diagnose) deeper problems are arguably more important to be cited. We, thus, assess the substance of each comment in this section. We first rated each debate ourselves and, second, we surveyed the authors of OPs and comments to obtain their assessment. For our own subjective rating, all three co-authors read the entire debate and answered the question "Should the comment be cited whenever the OP is cited?", with three possible answers: a) Yes, in virtually all cases; b) Yes, but only in some cases; c) No, the comment does not have to be cited. In case our ratings deviated, disagreements were resolved through discussion. We asked the authors the same question in our survey.

[21] An overview of the number of citing papers, references, and whether they cite the OP, the comment, or both, can be found in Table E8 in the Online Appendix.

The subjective leeway in rating the debates is obvious. For example, OPs might include several results of equal importance, but the comment only criticizes one. The reply by the original authors might acknowledge problems with this one result but claims that the OP provided several results, and the others still hold. Other, similar, scenarios are possible. Thus, any replication needs to be qualified and a debate is inevitable. Some comments stake out the implication of their criticism for the contribution of the OP very clearly, others do not. Most replies, in turn, do not acknowledge the diagnosed problem, or they acknowledge parts of it, but question the extent and the implications.

We also used the author survey to find out more about the debate.[22] We programmed the survey in Qualtrics and sent out the link to all authors via e-mail.[23] That is, if a paper is written by three authors, all three received a survey link and were informed that their co-authors were also contacted. We assured respondents anonymity in our invitation e-mail and, therefore, will not report paper-specific responses. Response rates are shown in Table 3.

Table 3: Author survey response rates

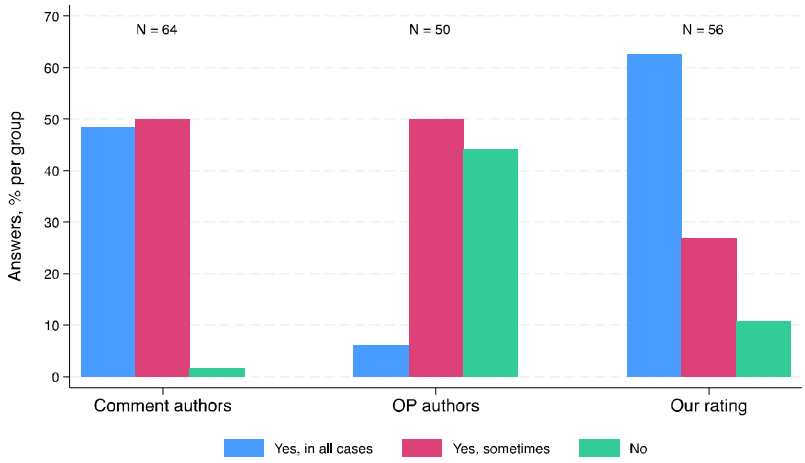
	By authors			By papers		
Survey	**Contacted**	**Replied**	**%**	**Contacted**	**Replied**	**%**
Comment	98	66	67%	51	43	84%
OP	111	53	48%	51	38	75%
Total	209	119	57%	102	81	79%

Notes. The response rates by papers counts a paper where at least one author responded to our survey as "replied". We depict the response rates separate for the four groups of authors: 1) Authors of comments that did not receive a reply from the original authors, 2) Authors of comments that did receive a reply from the original authors, 3) Authors of OPs that did not write a reply, and 4) Authors of OPs that did write a reply. We did not contact authors of debates where the OP received more than one comment. This applies to two OPs: Andreoni and Sprenger (2012) and Long and Ferrie (2013). In addition, 3 authors deceased, and we could not find working e-mail addresses of 5 authors.

Figure 7 shows the results for the author group ratings and our own. The difference between comment authors and original authors is striking, while our rating is closer to the comment authors.

[22] All details on the implementation of the survey, the comprehensive list of questions, the scripts of the survey, invitation e-mails, and reminders, as well as the details on the feedback rates can be found in Online Appendix B.
[23] We exclude the debates where the OP received multiple comments: Andreoni and Sprenger (2012) and Long and Ferrie (2013).

Figure 7: Ratings: "If a researcher cites the original paper, do you think the comment should be cited, as well?"



Notes. The number of observations reflects the responses we received from the OP and comment authors to this question, and in our case our rating of each of the 56 comments.

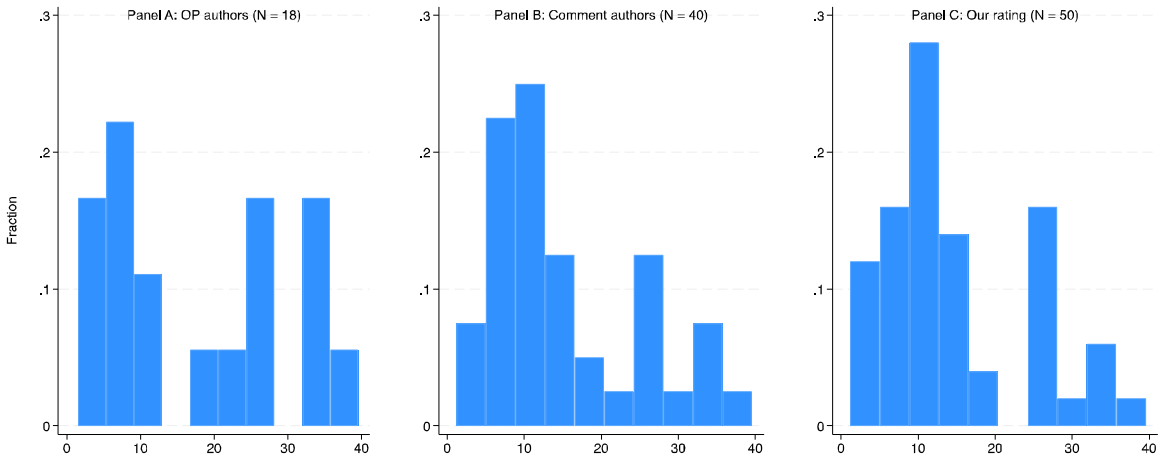
4.2. Citation patterns for comments rated as 'must-cite' and 'sometimes-cite'

We use the ratings by the three author groups to check whether our previous findings change if we exclude those comments that are rated as 'never-cite'. We first corroborate the distribution analysis of the citation ratio, now only for those comments that were rated either as 'must-cites' or 'sometimes-cites'. Figure 8 confirms our previous analysis in Figure 3 for comment authors' and our own rating. Even for those comments rated as 'must-' or 'sometimes-cites' by the OP authors, most reveal citation ratios around or below 10%.

Figure 9 shows annual citations for comments and OPs, now distinguished by our rating. Comments come only close to OPs for debates in which the OP itself is hardly cited. Several heavily cited OPs received comments that we rated as 'must-cites' – and yet the comments are barely cited. It is noteworthy that some OPs at the bottom of the figure are hardly cited, and, at the same time, we rated the comments as 'must-cites'.[24] It is arguably possible that the comment in these cases has worked as a self-correcting mechanism.

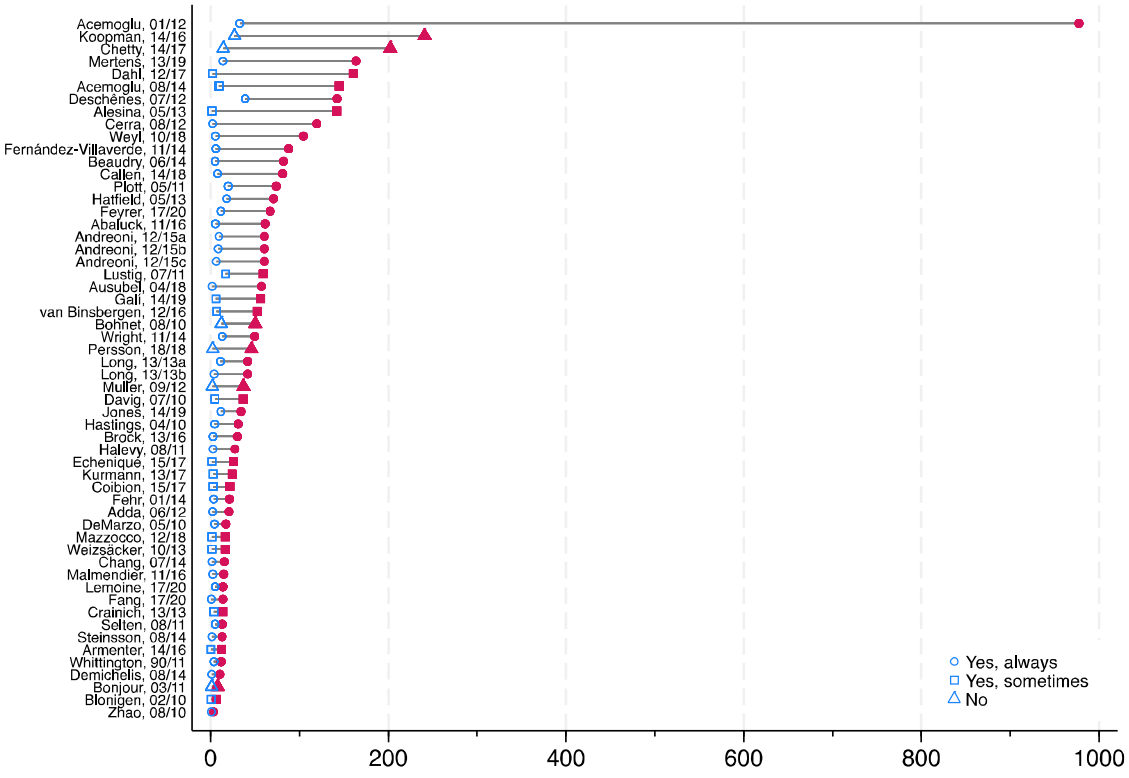
[24] In fact, in ten cases, the comment was published in the same year as the OP (four times) or only two years after (six times). Four out of the ten OPs have average annual citation counts below 25 since publication of the comment and are either rated as "must-cites" or "sometimes-cites".

Figure 8: Citation ratio distribution (in %) for comments rated as 'must-cite' and 'sometimes-cite'



Notes. The number of observations reflects the number of debates included in each panel, i.e., all debates where the comment was scored as either "yes, [should be cited in all cases]" or "sometimes". For the few debates where we received multiple responses with different scores from the same author group (original authors or comment authors), we calculated the average score. We only included debates with a score equal to or less than 2. Our scoring system was coded as 1 - "yes, [should be cited in all cases]", 2 – "sometimes", and 3 - "never".

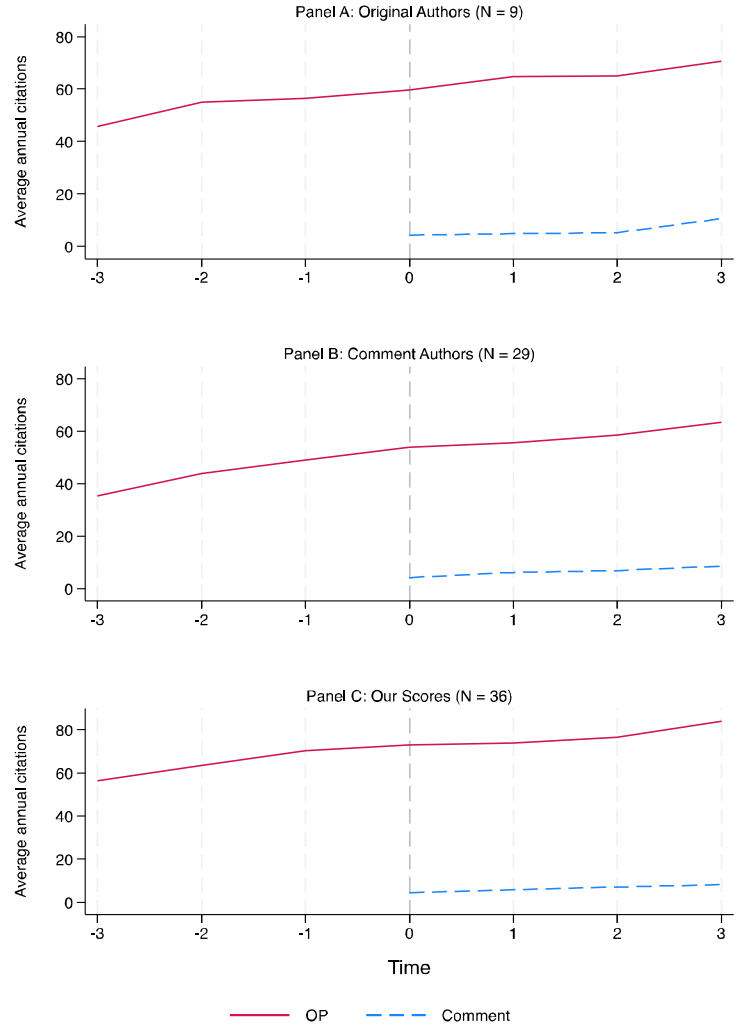
Figure 9: Average annual citations - difference between original paper and comment (by must-cite/sometimes-cite/never-cite)



Notes. We included all comments published in the *AER* between 2010 and 2020 and their respective OP. Citations are counted since comment publication for each debate. OPs have red markers; comment markers are blue. The coding of the debates, i.e., the marker shape, is based on our subjective ratings of the debates as discussed in Section 4.1. The labels on the y-axis show the first author of the OP, its publication year, and the publication year of the comment. The OP by Andreoni and Sprenger (2012) received three comments that we mark by the letters "a" (Miao and Zhong, 2015), "b" (Cheung, 2015), and "c" (Epper and Fehr-Duda, 2015). Similarly, the paper by Long and Ferrie (2013) received two comments that we mark with "a" (Xie and Killewald, 2013) and "b" (Hout and Guest, 2013).

Figure 10 tests for the robustness of our citation trend analysis pre- and post-publication of the comment in Panel A of Figure 5.[25] Again, our verdict holds: comments do not seem to lead to a reappraisal of the OPs' influence in the literature, even for 'must-cite' comments.

Figure 10: Average annual citation counts before and after comment publication for debates with comments rated as 'must-cite' or 'sometimes-cite'



4.3. Author survey results

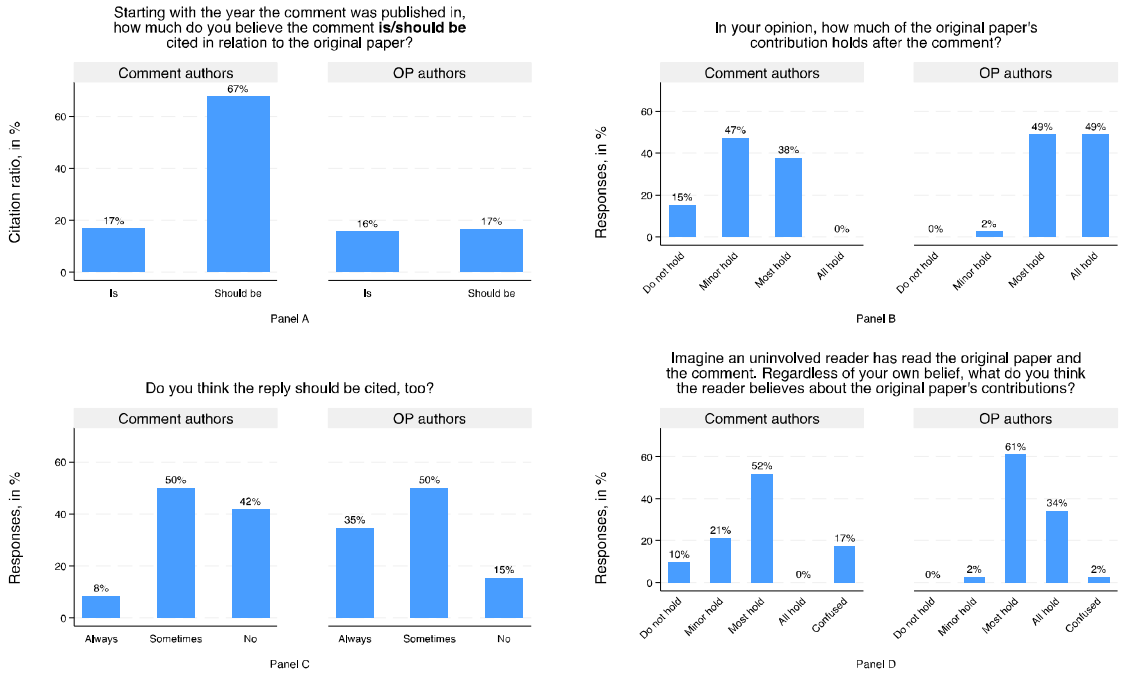
Some of the debates are very controversial and, especially those with a reply by the original authors, are not easy to rate. Even for comments that in our opinion clearly change fundamental parts of the OP's results, replies push back harshly or at least try

[25] See Figure D2 in the Online Appendix for additional robustness tests on different pre- and post-comment horizons, and for paper-specific citation trends see Figures D3-D6 in the Online Appendix.

to maintain the OP's contribution. We, therefore, added a few questions to the author survey on how the authors perceive the debate – and show some of the results in this section (see Online Appendix B for the comprehensive list of questions).[26]

Figure 11 provides different perspectives on how much both groups diverge, sometimes drastically, also for their opinion on the contribution of the OP after the comment. What we derive from this is that many uninvolved readers of the debates will have difficulties reassessing the OP's influence in the literature.

Figure 11: Authors' responses on the OP's contribution



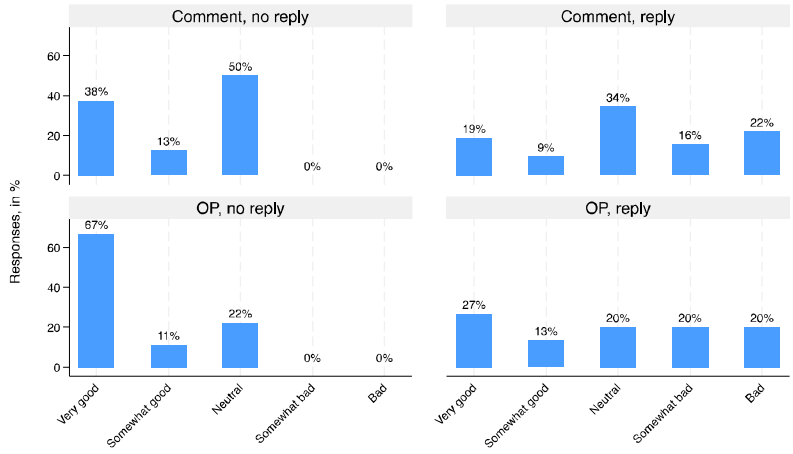
Notes. Values above the bars are rounded to the nearest integer. The exact answer options for Panels B and C were 1) "Do not hold anymore", 2) "Only minor contributions hold", 3) "Most important contributions hold", 4) "Hold in their entirety", 5) "The reader is likely confused". Respondents could also choose "Don't know" and "Refuse to answer". For Panel B, three comment authors chose "Refuse to answer", and for Panel D, two comment authors chose "Refuse to answer", and two comment authors as well as one original author chose "Don't know". Because we omit responses, the percentages in Panels B, C, and D add up to 100% for each group of authors, except for OP authors in Panel D due to rounding.

One of the most frequently mentioned impediments to replications is that they are perceived as hostile. We asked all authors about their sentiment towards the interaction with the other author team and, as Figure 12 shows, the picture is mixed. For those debates without a reply, both author teams overwhelmingly think positively

[26] The results to all questions can be obtained from the authors upon request.

or at least neutral of the interaction. For those with a reply, in both teams, a considerable percentage is unhappy with the interaction.

Figure 12: "How would you rate your interaction with the authors of the other paper?"



Note. Values above the bars are rounded to the nearest integer. Respondents could also choose "Don't know" and "Refuse to answer". One comment author chose "Refuse to answer" and one original author chose "Don't know". Because we omit these responses, the percentages add up to 100% for each group of authors. Differences are due to rounding. The numbers of observations are 16 for 'Comment, no reply', 32 for 'Comment, reply', 9 for 'OP, no reply', and 15 for 'OP, reply'.

The unpleasant experiences in their interaction with the other team are not only visible in the numbers but also in the text responses to open questions we asked. For example, one comment author said: "*We were aiming to clarify their views on some key issues, but we found it difficult to get them to engage substantively*". Another said that the relationship was good at first, but "*relations became more difficult when the nature of our criticism became clear*". Some comment authors were clearly frustrated by the exchange, as this testament shows: "*We could not pinpoint all of the issues, because they would not share their data. Once the data was published, we could begin the difficult task of unpacking their (many) errors. One of the authors was a bit more receptive, but the more powerful of the two […] simply wasn't willing to engage substantively. It was not pretty.*". While we received much negative feedback from comment authors, the original authors were less talkative but one original author said: "*The authors sent many different results some in support of our original paper, some in conflict, but the comment they wrote only contained the most negative*

one, without mention that it was not robust. They were essentially trying to score some cheap points to get a publication."

5. Conclusion

In this paper, we have shown that replications, published as comments in the *AER*, are not cited much and have no discernable impact on the citation trends of the replicated OP. We interpret this as evidence for the absence of self-correction mechanisms in economics. This verdict implies a narrow definition of scientific self-correction, what Peterson and Panofsky (2021) call 'formal self-correction'. Formal self-correction relies on 'diagnostic replication'[27] and "*its outcome is some change to the original study that either emends or retracts it*". To specify our interpretation, we hence provide strong evidence that economics is not subject to *formal* self-correction.

Our results are not directly at odds, though, with a broader definition, what Peterson and Panofsky (2021) call 'organic self-correction'. This happens "*largely through the unpublished backchannels of a field. […] Formal self-correction remembers wrongness; organic self-correction forgets that which is not useful.*" In fact, some of the barely cited OPs at the bottom of Figures 4 and 9 might have been subject to organic self-correction, catalyzed by the comment that got published shortly after (and hence, the literature has not even started to cite the OPs). Yet, several highly cited OPs in the upper part of those figures suggest anecdotally that economics is not very effective in organically self-correcting either.

Self-correction in economics is perhaps also difficult because the results of many replications are very controversial (see as well Ozier, 2021, and Roodman and Morduch, 2014). This holds for our sample: for cases with a reply by the original authors, the extent to which a comment changes the contribution of the OP is a matter of fierce debate. For a neutral reader – like us – it is often confusing. It also resonates

[27] In a companion paper we introduce a term, policing replication, that is similar in meaning to what Peterson and Panofsky (2021) call 'diagnostic replication' (see Ankel-Peters et al., 2023a).

with results from the author survey we conducted, as well as with statements we obtained from the former *AER* editors. It also is similar to what Harry Collins called the 'experimenter's regress': "*the problem is that, since experimentation is a matter of skillful practice, it can never be clear whether a second experiment has been done sufficiently well to count as a check on the results of the first*" (Collins, 1992, p. 2). We believe this problem is particularly hard to overcome in economics and other social sciences – disciplines that mostly work outside the laboratory – where the leeway for both researchers and replicators is very high (see Ankel-Peters et al., 2023; Breznau et al., 2022; Bryan et al., 2019; Huntington-Klein et al., 2021).

The absence of a clear-cut definition of robustness and replicability raises questions about the extent to which empirical economics can live up to the Popperian definition of 'science'. It does not have to, other reasonable epistemologies exist, which are not falsified by the absence of replicability, like Imre Lakatos's understanding of scientific progress through the progressiveness of research programs. This would imply, though, a humbler interpretation of research results and more modest communication to the outside world. Irrespective of this deeper epistemological question, economics could do more to reveal its appreciation for replication. The *AER*, to begin with, deserves to be applauded for systematically publishing comments and a rigorous data-sharing policy. In response to a previous version of our paper, the current *AER* editor has let us know that the journal has changed its policy and now, for new comments, will provide a link on the original paper's website. This is a small but perhaps important first step to giving replication work in economics the attention it needs and deserves.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *The American Economic Review*, *91*(5), 1369–1401.

Acemoglu, D., Johnson, S., Robinson, J. A., & Yared, P. (2008). Income and Democracy. *American Economic Review*, *98*(3), 808–842. https://doi.org/10.1257/aer.98.3.808

AEA. (2023). *AER: Editorial Policy*. https://www.aeaweb.org/journals/aer/about-aer/editorial-policy

Albouy, D. Y. (2012). The Colonial Origins of Comparative Development: An Empirical Investigation: Comment. *American Economic Review*, *102*(6), 3059–3076. https://doi.org/10.1257/aer.102.6.3059

Andreoni, J., & Sprenger, C. (2012). Risk Preferences Are Not Time Preferences. *American Economic Review*, *102*(7), 3357–3376. https://doi.org/10.1257/aer.102.7.3357

Ankel-Peters, J., Fiala, N., & Neubauer, F. (2023a). Do Economists Replicate? *Journal of Economic Behavior & Organization*, *212*, 219–232. https://doi.org/10.1016/j.jebo.2023.05.009

Ankel-Peters, J., Vance, C., & Bensch, G. (2023b). Spotlight on researcher decisions–Infrastructure evaluation, instrumental variables, and first-stage specification screening. *Ruhr Economic Papers, No. 991.*

Askarov, Z., Doucouliagos, A., Doucouliagos, H., & Stanley, T. D. (2022). The Significance of Data-Sharing Policy. *Journal of the European Economic Association*, jvac053. https://doi.org/10.1093/jeea/jvac053

Berry, J., Coffman, L. C., Hanley, D., Gihleb, R., & Wilson, A. J. (2017). Assessing the Rate of Replication in Economics. *American Economic Review*, *107*(5), 27–31. https://doi.org/10.1257/aer.p20171119

Bonjour, D., Cherkas, L. F., Haskel, J. E., Hawkes, D. D., & Spector, T. D. (2003). Returns to Education: Evidence from U.K. Twins. *American Economic Review*, *93*(5), 1799–1812. https://doi.org/10.1257/000282803322655554

Bornmann, L., & Daniel, H. (2008). What Do Citation Counts Measure? A Review of Studies on Citing Behavior. *Journal of Documentation*, *64*(1), 45–80. https://doi.org/10.1108/00220410810844150

Breznau, N., Rinke, E. M., Wuttke, A., & Nguyen, H. H. V. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *PNAS*, *119*(44). https://doi.org/10.1073/pnas.2203150119

Brodeur, A., Carrell, S., Figlio, D., & Lusher, L. (2023). Unpacking p-Hacking and Publication Bias. *American Economic Review (Forthcoming)*. https://doi.org/10.1257/aer.20210795

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review*, *110*(11), 3634–3660. https://doi.org/10.1257/aer.20190687

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, *8*(1), 1–32. https://doi.org/10.1257/app.20150044

Brunner, C., Camerer, C. F., & Goeree, J. K. (2011). Stationary Concepts for Experimental 2 × 2 Games: Comment. *American Economic Review*, *101*(2), 1029–1040. https://doi.org/10.1257/aer.101.2.1029

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate. *Proceedings of the National Academy of Sciences*, *116*(51), 25535–25545. https://doi.org/10.1073/pnas.1910951116

Budd, J. M., Sievert, M., & Schultz, T. R. (1998). Phenomena of Retraction: Reasons for Retraction and Citations to the Publications. *JAMA*, *280*(3), 296. https://doi.org/10.1001/jama.280.3.296

Burnside, C. (2011). The Cross Section of Foreign Currency Risk Premia and Consumption Growth Risk: Comment. *American Economic Review*, *101*(7), 3456–3476. https://doi.org/10.1257/aer.101.7.3456

Callen, M., Isaqzadeh, M., Long, J. D., & Sprenger, C. (2014). Violence and Risk Preference: Experimental Evidence from Afghanistan. *American Economic Review*, *104*(1), 123–148. https://doi.org/10.1257/aer.104.1.123

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. https://doi.org/10.1126/science.aaf0918

Card, D., & DellaVigna, S. (2020). What Do Editors Maximize? Evidence from Four Economics Journals. *The Review of Economics and Statistics*, *102*(1), 195–217. https://doi.org/10.1162/rest_a_00839

Caselli, F., & Ciccone, A. (2019). The Human Capital Stock: A Generalized Approach: Comment. *American Economic Review*, *109*(3), 1155–1174. https://doi.org/10.1257/aer.20171787

Chang, A. C., & Li, P. (2022). Is Economics Research Replicable? Sixty Published Papers FromThirteen Journals Say "Often Not." *Critical Finance Review*, *11*. https://doi.org/10.1561/104.00000053

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Cheung, S. L. (2015). Comment on "Risk Preferences Are Not Time Preferences": On the Elicitation of Time Preference under Conditions of Risk. *American Economic Review*, *105*(7), 2242–2260. https://doi.org/10.1257/aer.20120946

Christensen, G., & Miguel, E. (2018). Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature*, *56*(3), 920–980. https://doi.org/10.1257/jel.20171350

Clemens, M. A. (2017). The Meaning of Failed Replications: A Review and Proposal. *Journal of Economic Surveys*, *31*(1), 326–342. https://doi.org/10.1111/joes.12139

Coffman, L. C., Niederle, M., & Wilson, A. J. (2017). A Proposal to Organize and Promote Replications. *American Economic Review*, *107*(5), 41–45. https://doi.org/10.1257/aer.p20171122

Collins, H. (1992). *Changing order: Replication and induction in scientific practice.* University of Chicago Press.

Collins, H. M. (1991). The Meaning of Replication and the Science of Economics. *History of Political Economy*, *23*(1), 123–142. https://doi.org/10.1215/00182702-23-1-123

Coupé, T., & Reed, W. R. (2022). Do Negative Replications Affect Citations? *University of Canterbury, Department of Economics and Finance Working Papers in Economics (No. 22/16).*

Crump, R., Shah Goda, G., & Mumford, K. J. (2011). Fertility and the Personal Exemption: Comment. *American Economic Review*, *101*(4), 1616–1628. https://doi.org/10.1257/aer.101.4.1616

Dahal, M., & Fiala, N. (2020). What Do We Know about the Impact of Microfinance? The Problems of Power and Precision. *World Development*, *128*, 104773. https://doi.org/10.1016/j.worlddev.2019.104773

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, *76*(4), 587–603.

Dreber, A., & Johanneson, M. (2023). A Framework for Evaluating Reproducibility and Replicability in Economics. *I4R Discussion Paper Series No. 38*. https://doi.org/10.2139/ssrn.4458153

Epper, T., & Fehr-Duda, H. (2015). Comment on "Risk Preferences Are Not Time Preferences": Balancing on a Budget Line. *American Economic Review*, *105*(7), 2261–2271. https://doi.org/10.1257/aer.20130420

Fernández, L. M., Hardwicke, T. E., & Vadillo, M. A. (2019). *Retracted Papers Clinging on to Life: An Observational Study of Post-Retraction Citations in Psychology* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/cszpy

Ferraro, P. J., & Shukla, P. (2020). Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics? *Review of Environmental Economics and Policy*, *14*(2), 339–351. https://doi.org/10.1093/reep/reaa011

Fisher, A. C., Hanemann, W. M., Roberts, M. J., & Schlenker, W. (2012). The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather: Comment. *American Economic Review*, *102*(7), 3749–3760. https://doi.org/10.1257/aer.102.7.3749

Freese, J., & Peterson, D. (2017). Replication in Social Science. *Annual Review of Sociology*, *43*, 147–165.

Hamermesh, D. S. (2007). Viewpoint: Replication in Economics. *Canadian Journal of Economics/Revue Canadienne d'économique*, *40*(3), 715–733. https://doi.org/10.1111/j.1365-2966.2007.00428.x

Hamermesh, D. S. (2018). Citations In Economics: Measurement, Uses, and Impacts. *Journal of Economic Literature*, *56*(1), 115–156. https://doi.org/10.1257/jel.20161326

Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., van den Akker, O. R., Nuijten, M. B., & Ioannidis, J. P. A. (2021). Citation Patterns Following a Strongly Contradictory Replication Result: Four Case Studies From Psychology. *Advances in Methods and Practices in Psychological Science*, *4*(3), 251524592110408. https://doi.org/10.1177/25152459211040837

Hensel, P. G. (2021). Reproducibility and Replicability Crisis: How Management Compares to Psychology and Economics – A Systematic Review of Literature. *European Management Journal*, *39*(5), 577–594. https://doi.org/10.1016/j.emj.2021.01.002

Hout, M., & Guest, A. M. (2013). Intergenerational Occupational Mobility in Great Britain and the United States Since 1850: Comment. *American Economic Review*, *103*(5), 2021–2040. https://doi.org/10.1257/aer.103.5.2021

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The Influence of Hidden Researcher Decisions in Applied Microeconomics. *Economic Inquiry*, *59*(3), 944–960. https://doi.org/10.1111/ecin.12992

Institute for Replication. (2022). *Defintions*. https://i4replication.org/definitions.html

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, *127*(605), F236–F265. https://doi.org/10.1111/ecoj.12461

Jones, B. F. (2014). The Human Capital Stock: A Generalized Approach. *American Economic Review*, *104*(11), 3752–3777. https://doi.org/10.1257/aer.104.11.3752

Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *American Economic Review*, *73*(1), 31–43.

Lemoine, D., & Rudik, I. (2017). Steering the Climate System: Using Inertia to Lower the Cost of Policy. *American Economic Review*, *107*(10), 2947–2957. https://doi.org/10.1257/aer.20150986

Long, J., & Ferrie, J. (2013). Intergenerational Occupational Mobility in Great Britain and the United States Since 1850. *American Economic Review*, *103*(4), 1109–1137. https://doi.org/10.1257/aer.103.4.1109

Malkov, D., Yaqub, O., & Siepel, J. (2023). The Spread of Retracted Research into Policy Literature. *Quantitative Science Studies*, *4*(1), 68–90. https://doi.org/10.1162/qss_a_00243

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A Systematic Comparison of

Citations in 252 Subject Categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/j.joi.2018.09.002

Mattauch, L., Matthews, H. D., Millar, R., Rezai, A., Solomon, S., & Venmans, F. (2020). Steering the Climate System: Using Inertia to Lower the Cost of Policy: Comment. *American Economic Review*, *110*(4), 1231–1237. https://doi.org/10.1257/aer.20190089

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is Psychology Suffering from a Replication Crisis? What Does "Failure to Replicate" Really Mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Merton, R. K. (1973). The Normative Structure of Science. In N. W. Storer (Ed.), *The Sociology of Science*. The University of Chicago Press.

Miao, B., & Zhong, S. (2015). Comment on "Risk Preferences Are Not Time Preferences": Separating Risk and Time Preference. *American Economic Review*, *105*(7), 2272–2286. https://doi.org/10.1257/aer.20131183

Miguel, E. (2021). Evidence on Research Transparency in Economics. *Journal of Economic Perspectives*, *35*(3), 193–214. https://doi.org/10.1257/jep.35.3.193

Mirowski, P., & Sklivas, S. (1991). Why Econometricians Don't Replicate (Although They Do Reproduce). *Review of Political Economy*, *3*(2), 146–163. https://doi.org/10.1080/09538259100000040

Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019). Replication Studies in Economics—How Many and Which Papers Are Chosen for Replication, and Why? *Research Policy*, *48*(1), 62–83. https://doi.org/10.1016/j.respol.2018.07.019

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

Ozier, O. (2021). Replication Redux: The Reproducibility Crisis and the Case of Deworming. *The World Bank Research Observer*, *36*(1), 101–130. https://doi.org/10.1093/wbro/lkaa005

Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity. *The World Bank Research Observer*, *33*(1), 34–64. https://doi.org/10.1093/wbro/lkx005

Peterson, D., & Panofsky, A. (2021). Self-correction in science: The diagnostic and integrative motives for replication. *Social Studies of Science*, *51*(4), 583–605. https://doi.org/10.1177/03063127211005551

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Taylor & Francis. http://nukweb.nuk.uni-lj.si/login?url=http://search.ebscohost.com/login.aspx?authtype=ip&direct=true&db=nlebk&AN=143035&site=eds-live&scope=site&lang=sl

Roodman, D., & Morduch, J. (2014). The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence. *The Journal of Development Studies*, *50*(4), 583–604. https://doi.org/10.1080/00220388.2013.858122

Rothstein, J. (2017). Measuring the Impacts of Teachers: Comment. *American Economic Review*, *107*(6), 1656–1684. https://doi.org/10.1257/aer.20141440

Rubin, A., & Rubin, E. (2021). Systematic Bias in the Progress of Research. *Journal of Political Economy*, 000–000. https://doi.org/10.1086/715021

Schafmeister, F. (2021). The Effect of Replications on Citation Patterns: Evidence From a Large-Scale Reproducibility Project. *Psychological Science*, *32*(10), 1537–1548. https://doi.org/10.1177/09567976211005767

Selten, R., & Chmura, T. (2008). Stationary Concepts for Experimental 2x2-Games. *American Economic Review*, *98*(3), 938–966. https://doi.org/10.1257/aer.98.3.938

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable Publications are Cited More than Replicable Ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

Siler, K., Lee, K., & Bero, L. (2015). Measuring the Effectiveness of Scientific Gatekeeping. *Proceedings of the National Academy of Sciences*, *112*(2), 360–365. https://doi.org/10.1073/pnas.1418218112

Sukhtankar, S. (2017). Replications in Development Economics. *American Economic Review*, *107*(5), 32–36. https://doi.org/10.1257/aer.p20171120

Teplitskiy, M., Duede, E., Menietti, M., & Lakhani, K. R. (2022). How status of research papers affects the way they are read and cited. *Research Policy*, *51*(4), 104484. https://doi.org/10.1016/j.respol.2022.104484

The Replication Network. (2023). Replication Studies. *The Replication Network*. https://replicationnetwork.com/replication-studies/

Vazire, S., & Holcombe, A. O. (2021). Where are the Self-Correcting Mechanisms in Science? *Review of General Psychology*, 1–12. https://doi.org/10.1177/10892680211033912

Vilhuber, L. (2019). Report by the AEA Data Editor. *AEA Papers and Proceedings*, *109*, 718–729. https://doi.org/10.1257/pandp.109.718

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, *18*(6), 3045–3089. https://doi.org/10.1093/jeea/jvaa019

von Hippel, P. T. (2022). Is Psychological Science Self-Correcting? Citations Before and After Successful and Failed Replications. *Perspectives on Psychological Science*, 174569162110725. https://doi.org/10.1177/17456916211072525

Whaples, R. (2006). The Costs of Critical Commentary in Economics Journals. *Econ Journal Watch*, *3*(2), 275–282.

Whittington, L. A., Alm, J., & Peters, H. E. (1990). Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States. *American Economic Review*, *80*(3), 545–556. https://www.jstor.org/stable/2006683

Xie, Y., & Killewald, A. (2013). Intergenerational Occupational Mobility in Great Britain and the United States Since 1850: Comment. *American Economic Review*, *103*(5), 2003–2020. https://doi.org/10.1257/aer.103.5.2003

The Online Appendix can be accessed on the OSF website under this link: https://osf.io/kpywn/?view_only=2c4ecfdcc66f40459b6073ad9e7b42ce.

Appendix

Table A1: Citations on comments published in the AER between 2010 and 2020 and their respective original papers (OPs)

	OP				Comment		
	Since OP Publication		Since Comment Publication				
OP First Author / Comment First Author	Total Citations (1)	Average Annual Citations (2)	Total Citations (3)	Average Annual Citations (4)	Total Citations (5)	Average Annual Citations (6)	Citation Ratio (Comment/ OP) Since Comment Publication (in %) (7)
Median	**395**	**33**	**229**	**37**	**30**	**5**	**11.2**
Mean	**839**	**60**	**567**	**74**	**56**	**7**	**14.3**
Acemoglu (2001) / Albouy (2012)	15,347	731	9,772	977	328	33	3.4
Alesina (2005) / Tella (2013)	1,912	112	1,277	142	15	2	1.2
Acemoglu (2008) / Cervellati (2014)	1,827	131	1,161	145	77	10	6.6
Koopman (2014) / Los (2016)	1,677	210	1,446	241	162	27	11.2
Deschênes (2007) / Fisher (2012)	1,652	110	1,423	142	389	39	27.3
Cerra (2008) / Mueller (2012)	1,503	107	1,195	120	23	2	1.9
Dahl (2012) / Lundstrom (2017)	1,409	141	802	160	12	2	1.5
Chetty (2014) / Rothstein (2017)	1,339	167	1,011	202	71	14	7.0
Ausubel (2004) / Okamoto (2018)	1,264	70	229	57	7	2	3.1
Plott (2005) / Isoni (2011)	1,129	66	813	74	218	20	26.8
Beaudry (2006) / Kurmann (2014)	1,070	67	656	82	40	5	6.1
Weyl (2010) / Tan (2018)	935	78	418	105	22	6	5.3
Mertens (2013) / Jentsch (2019)	850	94	491	164	42	14	8.6
Hatfield (2005) / Aygün (2013)	846	50	638	71	163	18	25.5
Fernández-V. (2011) / Born (2014)	826	75	703	88	48	6	6.8
Lustig (2007) / Burnside (2011)	769	51	652	59	186	17	28.5
Bohnet (2008) / Bolton (2010)	629	45	605	50	146	12	24.1
Abaluck (2011) / Ketcham (2016)	577	52	369	62	33	6	8.9
Andreoni (2012) / Cheung (2015)	519	52	424	61	60	9	14.2
Andreoni (2012) / Epper (2015)	519	52	424	61	44	6	10.4
Andreoni (2012) / Miao (2015)	519	52	424	61	66	9	15.6
Callen (2014) / Vieider (2018)	512	64	325	81	31	8	9.5
Davig (2007) / Farmer (2010)	502	33	440	37	58	5	13.2
Hastings (2004) / Taylor (2010)	483	27	373	31	56	5	15.0
Fehr (2001) / Petersen (2014)	470	22	170	21	27	3	15.9
Wright (2011) / Bauer (2014)	466	42	397	50	106	13	26.7
Muller (2009) / Fraas (2012)	417	32	371	37	21	2	5.7
Binsbergen (2012) / Schulz (2016)	413	41	313	52	41	7	13.1
Galí (2014) / Miao (2019)	395	49	168	56	18	6	10.7
Long (2013) / Hout (2013)	375	42	375	42	36	4	9.6
Long (2013) / Xie (2013)	375	42	375	42	100	11	26.7
Halevy (2008) / Saito (2011)	343	25	301	27	29	3	9.6
Adda (2006) / Abrevaya (2012)	296	19	207	21	23	2	11.1
Whittington (1990) / Crump (2011)	289	9	130	12	42	4	32.3
Feyrer (2017) / James (2020)	269	54	134	67	23	12	17.2
DeMarzo (2005) / Che (2010)	231	14	207	17	54	5	26.1
Brock (2013) / Krawczyk (2016)	229	25	181	30	16	3	8.8
Chang (2007) / Takahashi (2014)	203	14	124	16	13	2	10.5
Malmendier (2011) / Schneider (2016)	190	17	89	15	15	3	16.9
Jones (2014) / Caselli (2019)	187	23	103	34	35	12	34.0
Persson (2018) / Matsumoto (2018)	184	46	184	46	10	3	5.4
Weizsäcker (2010) / Ziegelmeyer (2013)	177	15	150	17	14	2	9.3
Steinsson (2008) / Iversen (2014)	171	12	104	13	13	2	12.5

Selten (2008) / Brunner (2011)	169	12	144	13	57	5	39.6
Mazzocco (2012) / Shrinivas (2018)	169	17	67	17	6	2	9.0
Kurmann (2013) / Cascaldi-Garcia (2017)	165	18	121	24	14	3	11.6
Demichelis (2008) / Heller (2014)	160	11	84	11	10	1	11.9
Bonjour (2003) / Amin (2011)	148	8	89	8	15	1	16.9
Echenique (2015) / Doğan (2017)	147	21	128	26	7	1	5.5
Coibion (2015) / Gagnon (2017)	137	20	108	22	14	3	13.0
Crainich (2013) / Ebert (2013)	125	14	125	14	32	4	25.6
Blonigen (2002) / Kelly (2010)	116	6	79	7	8	1	10.1
Armenter (2014) / Blum (2016)	110	14	75	13	3	1	4.0
Lemoine (2017) / Mattauch (2020)	56	11	28	14	10	5	35.7
Zhao (2008) / Chen (2010)	41	3	40	3	10	1	25.0
Fang (2017) / Matsumoto (2020)	40	8	28	14	2	1	7.1

Notes. The two OPs that received multiple comments are only included once in the calculation of the mean and median. Thus, the number of observations are N=53 for OPs and N=56 for comments. Some comments are cited as discussion papers prior to their publication in the *AER*. In this table, we only include citations after their publication. However, only four out of the 56 comments have at least 10 citations prior to the *AER* publication, probably because discussion paper versions had circulated before: Albouy (2012). Burnside (2011), Cheung (2015), and Fisher et al. (2012). The same table can be found in Table D6 in the Online Appendix where we include the total citations of the reply as an additional column.