Fiala, Nathan; Neubauer, Florian; Peters, Jörg

Working Paper

Do economists replicate?

Ruhr Economic Papers, No. 939

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Fiala, Nathan; Neubauer, Florian; Peters, Jörg (2022) : Do economists replicate?, Ruhr Economic Papers, No. 939, ISBN 978-3-96973-100-0, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen, https://doi.org/10.4419/96973100

This Version is available at:
https://hdl.handle.net/10419/250076

RUHR
ECONOMIC PAPERS

Nathan Fiala
Florian Neubauer
Jörg Peters

Do Economists Replicate?

RWI #939

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, e-mail: W.Leininger@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Manuel Frondel, Prof. Dr. Torsten Schmidt, Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #939

Responsible Editor: Manuel Frondel

ISSN 1864-4872 (online) – ISBN 978-3-96973-100-0

Ruhr Economic Papers #939

Nathan Fiala, Florian Neubauer, and Jörg Peters

Do Economists Replicate?

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/96973100
ISSN 1864-4872 (online)
ISBN 978-3-96973-100-0

Nathan Fiala, Florian Neubauer, and Jörg Peters[1]

Do Economists Replicate?

Abstract

Reanalyses of empirical studies and replications in new contexts are important for scientific progress. Journals in economics increasingly require authors to provide data and code alongside published papers, but how much does the economics profession indeed replicate? This paper summarizes existing replication definitions and reviews how much economists replicate other scholars' work. We argue that in order to counter incentive problems potentially leading to a replication crisis, replications in the spirit of Merton's 'organized skepticism' are needed – what we call 'policing replications'. We review leading economics journals to show that policing replications are rare and conclude that more incentives to replicate are needed to reap the fruits of rising transparency standards.

JEL-Codes: A11, C18

Keywords: Replication; research transparency; generalizability

February 2022

1 Nathan Fiala, RWI and University of Connecticut; Florian Neubauer, RWI and University of Connecticut; Jörg Peters, RWI and University of Passau.- *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 thank Gunther Bensch, Abel Brodeur, Daniel Hamermesh, David McKenzie, Frank Müller-Langer, Maximiliane Sievert, Sandip Sukhtankar, Séverine Toussaert, Tim Salmon, and Alistair Wilson for valuable comments and suggestions. We are particularly grateful to Niklas Benner of RWI for supporting us with textual analysis techniques to identify citations in the abstracts of papers published in the Top 50 journals. Caitlin Chan provided valuable research assistance. –* *All correspondence to: Jörg Peters, RWI, Hohenzollernstr. 1/3, 45128 Essen, Germany, e-mail: peters@rwi-essen.de*

1. Introduction

Replications are an important tool in all empirical disciplines to verify results, uncover errors and fraud, and test the generalizability of previous findings to new contexts. This is especially true for empirical economics, with its profound implications for policy decisions. At the same time, a growing recent literature raises concerns about the replicability of economics research (Brodeur et al., 2016, 2020; Camerer et al., 2016; Christensen and Miguel, 2018; Ferraro and Shukla, 2020; Huntington-Klein et al., 2021; Ioannidis et al., 2017; Peters et al., 2018; Vivalt, 2020), suggesting that the economics profession might need some introspection as to the extent that Robert Merton's norm of 'organized skepticism' is being maintained (Merton, 1973). An essential component of organized skepticism, on top of peer review, we argue, are replications that critically reflect on published studies. In the present paper, we examine how widespread replication is in economics.

First, we review the three existing systematic reviews that estimate replication rates in economics (that is, how often replications are being conducted): Berry et al. (2017), Mueller-Langer et al. (2019), Sukhtankar (2017). They provide a wide range of replication rates and, as we will show, the most important reason for this are different definitions of what constitutes a replication[1]. Despite attempts such as Hamermesh (2007) and most notably Clemens (2017), there is still no convention among economists about what constitutes a replication and which different subtypes exist. We contend that the different definitions used in the three reviews are reasonable but differ in what they capture as the purpose of replications. Broader definitions of replications include studies that build on an existing empirical finding by slightly modifying the research question and applying it in a new context. Such *implicit* replication work is daily fare in economics and hence they find that replication rates are high.

[1] Note that throughout the paper we use the terms 'replication' and 'replicate' to refer to the process of replicating a previously published study (e.g. through a reanalysis or an extension), not to a succesful attempt of obtaining the same result as the previously published study.

While we acknowledge the scientific importance of this type of research, we argue that a narrower type of replication is needed that stress-tests published results to uncover purposeful or unintentional questionable research practices, as they have been diagnosed, for example, in Ferraro and Shukla (2020). We refer to this as *policing replication*.[2] Questionable research practices are not necessarily "*blatantly improper*" but "*offer considerable latitude for rationalization and self-deception*" (John et al., 2012). They comprise *p*-hacking (Brodeur et al., 2020; Ferraro and Shukla, 2020; Huntington-Klein et al., 2021), ex-post theorizing (Kerr, 1998), reporting underpowered results (Dahal and Fiala, 2020; Ioannidis et al., 2017), uncorrected multiple hypothesis testing (Anderson, 2008; Fink et al., 2014), and coding errors (Foote and Goetz, 2008). The 'policing replication' category complements the replication nomenclature in the seminal Clemens (2017) work (which we refer to as 'Clemens nomenclature' going forward; see Table 1 for an overview). The constitutive feature of a replication to qualify as policing, we argue, is the direct engagement with the original work, for example in the abstract. In this sense, all of Clemens' categories a priori qualify, although some are likelier to be policing replications than others.

In a next step, we examine how frequent policing replications are in economics. We check for policing replications in three ways: we first review the Top 50 economics journals for papers that directly challenge a previously published paper. We find that 176 (or 0.6%) of all 29,643 published papers in the Top 50 economics journals between 2010 and 2020 fit our definition of policing replication. Second, we corroborate this finding by isolating those replications from Berry et al. (2017), Sukhtankar (2017), and Mueller-Langer et al. (2019) that would qualify as policing replications. We, third, look at how many comments have been published in the *American Economic Review (AER)* over time, one of the profession's leading journals. Comments typically discuss papers that were published in the same journal and hence arguably have strong policing components. We find that there is a continuous downward sloping trend over the past

[2] The term 'policing' has been coined by Ofosu and Posner (2019) in their review of pre-analysis plans in registries, see below.

decades and in recent years less than 3 % of papers published in the AER have been comments.

Our paper contributes to a growing meta-scientific literature in economics and its research transparency debate (Christensen and Miguel, 2018). More specifically, we add to the discussion of rising transparency standards. Important reviews demonstrate that the economics profession has made tremendous progress in recent years on the availability of data and code for published work (Christensen et al., 2020; Christensen and Miguel, 2018; Miguel, 2021; Vilhuber, 2020). Yet, this only leads to more incentives for credible research if in parallel a replication culture is established that takes the data to the test (Höffler, 2017). Ofosu and Posner (2019) and Laitin (2013) make a similar case for the effectiveness of pre-analysis plans (PAP) in combating *p*-hacking and publication bias. As Ofosu and Posner (2019) point out: "*whatever the benefits of pre-registration may be in theory, PAPs are unlikely to enhance research credibility without vigorous policing*". We, therefore, conclude by a plea for more incentives to replicate: all types of replications are important, but the profession's reward system should particularly facilitate more policing work.

2. Replication definitions and replication rates in economics

There is no universally accepted definition of replication in economics. Several papers structure what types of replications exist. The most influential ones in economics are Hamermesh (2007) and Clemens (2017), followed by an overview paper on research transparency by Christensen and Miguel (2018). Another compelling categorization is Freese and Peterson (2017) from quantitative sociology, a sister discipline. Table 1 summarizes the replication categories according to whether the new paper uses the same specification, the same population, and the same sample.[3]

[3] Another dimension that could be added is whether a replication looks into the raw data of a published paper or only into the cleaned data set uploaded on the journal's webite. See, for example, Huntington-Klein et al. (2021) and Ozier (2021).

Table 1: Most significant replication definitions in the social sciences

Author(s), Year	Category	New paper uses the same…		
		Specification[1]	Population	Sample
Clemens, M.A., 2017	Verification	✓	✓	✓
	Reproduction	✓	✓	✗
	Reanalysis	✗	✓	✓/✗[2]
	Extension	✓	✗	✗
Freese, J., Peterson, D., 2017	Verifiability	✓	✓	✓
	Robustness	✗	✓	✓
	Repeatability	✓	?[3]	✗
	Generalization	✗	?[3]	✗
Hamermesh, D.S., 2007	Pure replication	✓	✓	✓
	Statistical replication	✓	✓	✗
	Scientific replication	✓/?[4]	✗	✗

[1]Hamermesh (2007) calls it 'model', and Freese and Peterson (2017) vary in their wording between 'analysis', 'specification', 'procedure', and 'method'. For consistency across the three papers, we named it 'specification'. [2]According to Clemens, a reanalysis can use exactly the same data as the original study or a new sample from the same population. [3]Freese and Peterson (2017) do not specify whether the new sample shall come from the same or a new population. [4]Similar but not identical specification.

The question of how much economists replicate has been systematically addressed in three reviews, summarized in Table 2. The diagnosed replication rates range between less than 1% and 60%, which is the upper bound of Berry et al. (2017). The broad range can be ascribed to different definitions of 'replication rate'. To start with, the replication rate can measure two things: first, how many published papers are *replicated*. Or, second, how many published papers are *replications*. We refer to the first as the *selective replication* approach, pursued by Berry et al. (2017) and Sukhtankar (2017) and also similar to the logic of Hamermesh (2017), who argues that only influential studies need to be replicated. We refer to the second approach as the *total replication* approach, used by Mueller-Langer et al. (2019), who review all papers in the Top 50 journals published between 1974 and 2014 to check how many of these are replications. The logic here is that any empirical finding that is seriously published contributes to the knowledge base and should be put to scrutiny. Because more influential studies attract more replications, the selective replication approach delivers much higher replication rates than the total replication approach.

The three reviews also use different definitions for what constitutes a replication in the first place. Berry et al. (2017) use a very pragmatic approach and define three categories: 'replication', 'extension', and 'robustness tests'.[4] A 'replication', according to Berry et al. (2017), is "*any project that reports results that speak directly to the veracity of the original paper's main hypothesis*" (p. 27). An 'extension' is a paper that is "*testing a closely related hypothesis to the original paper*" (p. 28). 'Robustness tests' are papers that either use the same specification in a new sample and population or different specifications on the same data. Berry et al. (2017) find that 28.6%, 48.6%, and 40% of papers in the *AER* volume under scrutiny are 'replications', 'extensions', and 'robustness tests', respectively. The authors emphasize in their abstract that 60% of papers in this *AER* volume have either a 'replication', 'robustness test', or an 'extension'. All three categories are very inclusive and broad.[5] Berry et al. (2017) also document narrower categories like 'verifications' and 'reproductions' (using the Clemens nomenclature, see Table 1) for which they find only zero and two cases, respectively.

Sukhtankar (2017), the other review that uses the selective replication approach, applies a much narrower definition, strictly following the Clemens nomenclature. Correspondingly, the overall replication rate in Sukhtankar (2017) is much lower at 6.2% when replications in working papers are included and at 3.3% when the focus is on peer-reviewed replications only. Likewise, Mueller-Langer et al. (2019) apply a narrower definition, guided by the Hamermesh (2007) categories, to elicit the total replication rate of how many papers published in the Top 50 journals are replications. Very few are, only 130 in all 126,505 published papers between 1974 and 2014 (0.1%).

[4] Note that Berry et al. (2017) deviate from the Clemens definition of these terms.

[5] For example, Berry et al. (2017) coded Magnan et al. (2015) as a 'replication' of the seminal paper by Conley and Udry (2010), which looks at social learning in driving the adoption of fertilizer for pineapple production in Ghana. Five years later, Magnan et al. (2015) investigate how social learning affects the demand for a water-saving agricultural technology in India.

Table 2: Overview of papers investigating replication rates

Paper	Replication rates	Definition of replication	Search engine	Inclusion criteria	Search strategy for replications, coding
A. Selective Replications - "How many published papers are replicated?"					
Sukhtankar 2017	Overall (incl. working papers): 6.2% Published: 3.3 % RCTs: 12.5% 71 replication studies were found, and they include: Replication verification: 32.4% Replication extension: 0% Robustness reanalysis: 77.5% Robustness extension: 36.6% (they don't add up to 100% because some studies included different replication types)	<u>Clemens nomenclature[1]</u> I. Replication: a) Verification b) Reproduction II. Robustness: c) Reanalysis d) Extension	GS[3]	<u>Original papers:</u> - Top five journals (AER, Econometrica, JPE, QJE, ReStud) and next five general-interest journals: AEJ:AE, AEJ:EP, EJ, JEEA, ReStat - JEL code: O - 2000-2015 <u>Replicating papers:</u> Published, working papers	- First step: Formalized word search in GS search among citing papers for "replicate OR replicates OR replicated OR replication OR replicating" - Second step: Subjective coding of replications, i.e., no formalized criteria or protocol in decision whether a paper is a replication or not - Supplemented GS search with search on other websites (see column "additional sources")
Berry et al. 2017	Replication: 28.6%[2] Extension: 48.6% Robustness: 40% Any of the three: 60%	Definitions of rates in previous column: A. 'Replication': "Any project that reports results that speak directly to the veracity of the original paper's main hypothesis" B. 'Extension': "Testing a closely related hypothesis to the original paper" C. 'Robustness': Clemens' Robustness categories: Robustness reanalysis, Robustness extension	WoS	<u>Original papers:</u> - AER centenary volume (2010) <u>Replicating papers:</u> - Top 200 economics journals, Published papers only - 2010-2016	- Checked every citing paper of the 70 papers in the AER centenary volume whether it is a replication or not - Subjective coding of replications, i.e., no formalized criteria or protocol in decision whether a paper is a replication or not
B. Total Replications - "How many published papers are replications?"					
Mueller-Langer et al. 2019	0.1%	A. Narrow: Same data and code B. Wide replication: a) new data, same methods, same models b) same data, new methods, new models c) new data, new methods, new models	WoS	<u>Original and replicating papers:</u> - Top 50 Econ journals - Published papers - 1974-2014	- First step: Formalized word search in title and abstract for keywords such as “repli*,” “reexamin*,” “comment,” “revisit,” “retesting,” or “reappraisal” (among others), as well as references to other articles - Second step: Used frequency and location of keywords to determine likelihood of being a replication, then ranked them for each journal and looked at the 100 highest ranked papers in each journal in detail - Also included all eligible replications from ReplicationWiki in their dataset

Notes: Mueller-Langer et al. (2019) and Sukhtankar (2017) both used the replication database of the University of Göttingen as an additional source to find replications; Sukhtankar (2017) further used replicationnetwork.com and the 3ie Replication Paper Series. [1]See also Table 1. [2]These categories do not reflect the Clemens nomenclature, except for the "Robustness" category, which comprises Clemens' reanalysis and extension categories. See Section 2 for the definition of the categories used in Berry et al. (2017). [3]GS is Google Scholar. WoS is Web of Science.

3. Policing replication

3.1 A plea for more clarity: Assuming the burden-of-proof

We fully acknowledge the scientific value of replications in the spirit of broader definitions as they are used, for example, in Berry et al. (2017). Yet, such *implicit* replication does not organize the skepticism that Merton (1973) called for. To this end, also *explicit* replications are needed that directly scrutinize whether a paper's claim is valid. Building on Ofosu and Posner (2019), we propose the term 'policing replications' for this type of replication.

The deficiency we lament here is that for most replications in broader senses, it is effectively left to the reader to perceive a study as a replication or not (or to the coder, as in the case of the three summarized reviews). We believe there should be more clarity about whether a new paper "*speaks to the veracity*" (Berry et al., 2017) of an influential previous study and we would therefore re-emphasize another important proposition of Clemens (2017): "*the burden of proof [for] a study to demonstrate that it should have obtained identical results to the original*" is with the authors of the (potential) replication.

3.2 Definition

We propose a straightforward definition: to qualify as a policing replication, the replication should directly challenge a previously published empirical paper and address this original paper prominently, that is, in the title or abstract. The rationale of this is that an act of policing must be directly attributable to a case. Just like previous papers conceptualizing sub-types of replications, we acknowledge that this is no clear-cut definition. How does our proposal relate to the Clemens nomenclature? A 'verification' is a policing replication in all cases, 'reproductions' and 'reanalyses' have strong policing features in most cases, but even 'extensions' can be policing replications.[6]

[6] Hamermesh's 'pure replication' and Freese and Peterson's 'verifiability' and 'robustness' categories are those types that most clearly overlap with our policing replication definition.

The term policing is meant to convey that empirical scientific discovery needs to be controlled systematically to institutionalize incentives that prevent questionable research practices and fraud. We acknowledge that the term might evoke some negative connotations. We use policing in its very positive sense, that is, a regulatory act preventing intentional or unintentional bad behavior. The police do not sentence. The police only investigate and compile evidence for a case. This evidence is then used by prosecutors and, potentially, a verdict is pronounced by a court of law. In this sense, a policing replication investigates a previously published paper – the role of the prosecutor and the court of law is with the scientific community as the readership. An excellent piece of scholarship in this regard is Ozier (2021).

3.3. Policing replication rates

We now push further by asking how many policing replications are being published. First, we screened all papers published in the Top 50 economics journals[7] between 2010 and 2020 for whether they are policing replications (thereby looking for the *total replication rate*) and scraped all papers that

- directly cite another paper in their abstract or title, or
- that include the word "comment" in the title, or
- that include the word "replic*", "reanal*", or "revisit*" in title, abstract, or keywords.

Out of 29,643 papers published in total in the Top 50 journals between 2010 and 2020, 967 papers meet these formalized search criteria (see Table A1 in the appendix for a comprehensive list)[8]. We read all abstracts of the resulting papers – which is where the policing ambition must become apparent according to our definition – and coded them as policing replications or not. We acknowledge that this can be – at least to some degree – a subjective exercise and there may be a grey area where it is sometimes not

[7] We used the Top 50 journals as listed on https://ideas.repec.org/top/top.journals.simple.html, accessed last on July 28, 2021.

[8] The search results as well as the coding of the 969 papers that meet the formalized search criteria are in the online appendix and can be obtained from the authors upon request.

clear if a paper is a policing replication or not. To be conservative, we code papers for which we were on the fence as policing.[9] Our estimate of the number of policing replications is thus likely an upper bound.

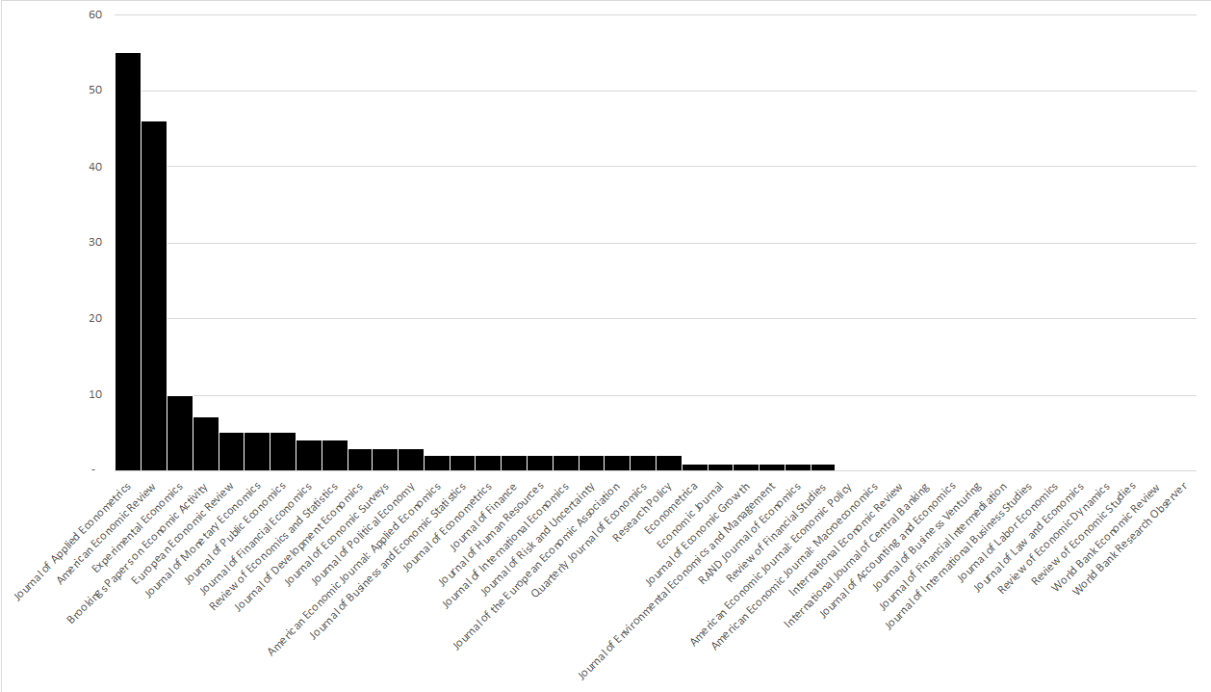
We identified 176 policing replications (i.e., 0.6% of all published papers, see Table A1 in the appendix), of which 156 papers cite the replicated study in the title or abstract. As can be seen in Figure 1, 14 of the 42 journals that were eventually included[10] have not published any policing replication since 2010 and two journals (the *AER* and the *Journal of Applied Econometrics*) account for almost 60% of all published policing replications.

Second, to corroborate our review, we isolated those replications identified in Berry et al. (2017), Sukhtankar (2017), and Mueller-Langer et al. (2019) that qualify as policing replications by going through their abstracts and coding them as policing or not. For Berry et al. (2017), except for one paper, none of the 52 replications qualify as policing. For Sukhtankar (2017), 50 out of 71 papers meet our policing criterion. Policing replication rates in these two reviews are hence at 1.2% and 4.4%, respectively (see Table A2 in the Appendix). Recall that both reviews look at selective replication rates, which are higher by design since influential papers are more likely to be replicated. In Mueller-Langer et al. (2019), a review of selective replication rates like ours, all except nine papers coded by the authors as replications are policing replications in our sense, leading to a policing replication rate of 0.1%.

[9] A special case are papers that review multiple papers. A priori, we would consider a replication of a limited number of papers as policing, since it would still uncover paper-specific problems. However, we would not consider systematic reviews and meta-analyses as replications. The demaraction when a replication of a limited number of papers turns into a meta-analysis is not always clear. We tried to be conservative by coding a paper as a policing replication in case of an ambiguous description in the abstract. Also note that this only affects <1% of the papers that met the formalized search criteria.

[10] From the Repec Top 50 list we excluded journals of federal reserve banks because they are not listed on Scopus (*Proceedings, Federal Reserve Bank of Cleveland*, *Proceedings, Federal Reserve Bank of San Francisco*, and *Quarterly Review, Federal Reserve Bank of Minneapolis*). We further excluded the *Journal of Economic Perspectives*, the *Journal of Economic Literature*, the *Annual Review of Economics, and Foundations and Trends in Econometrics* because they are review journals and unlikely to publish replication studies. The *Journal of Economic Theory* is excluded because we concentrate our work on empirical studies. The *Journal of Business* was discontinued in 2006 and is therefore not part of this table, either. Two journals were jointly listed on rank 21, which is the reason for arriving at 42 journals in total.

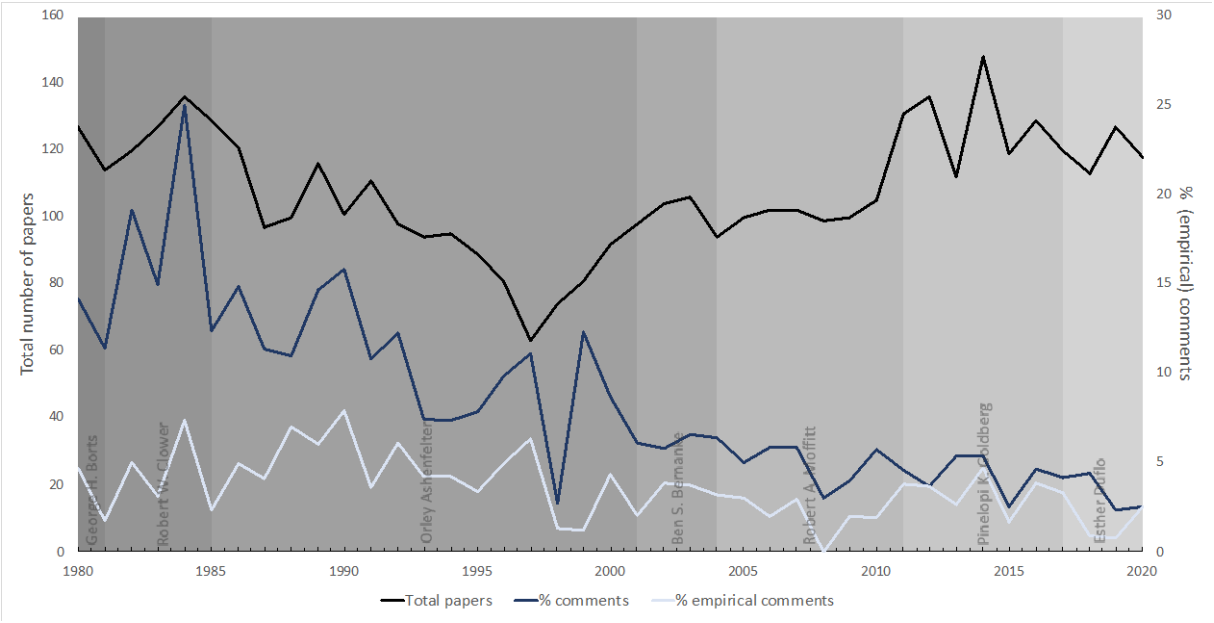
Figure 1: Policing replications in the Top 50 economics journals between 2010 and 2020



Third, we now zoom into one of the two journals that publish most policing replications, the *AER*, and investigate how many comments it has published since 1980. Comments typically discuss papers that were published in the *AER* before, and hence – if empirical – arguably have strong policing components. We find that there is a continuous downward sloping trend of published comments, from a high level of between 10 and 20 % of all papers in the 80s and 90s to below 5 % in the early 2010s and 2-3% in the most recent years (see Figure 2).[11] Yet, most of the early comments were on theoretical papers. When we only look at comments on *empirical* papers – the requirement to qualify as a policing replication – we see that their share of total papers has still decreased considerably, just now starting from an already very low level: The average number of empirical comments in the 80s was at 4.5% and has been at around 2.7% in the past 10 years. This drop is a noteworthy development, especially considering the sharp increase in empirical work, including the AER (Angrist et al., 2017).

[11] This trend was diagnosed for an earlier period already in Coelho et al. (2005).

Figure 2: Total number of papers and share of (empirical) comments in the AER since 1980



Note: Gray shaded areas indicate the periods of AER editors-in-chief. Source: Own data.

4. Conclusion

Complementing the nomenclature by Clemens (2017) but also Hamermesh (2007) and Freese and Peterson (2017), we propose a dedicated type of replication that polices previously published work in the spirit of Merton's organized skepticism. We have also taken stock of how much policing replications are being published in economics journals. Our finding suggests that below 1% of published papers indeed police previous work and between 1 and 4% of very influential papers are subject to a policing replication. Whether this is a reason for concern or not depends on one's prior about whether there is a replicability problem in the profession. We believe some recent meta-scientific work suggests there is.

As a matter of course, the policing replication rates diagnosed in our paper are based on replications that make it into working paper status or peer-reviewed journals. A lot of replication work is happening in economics classes on both graduate and undergraduate levels where influential papers are re-analysed or subject to robustness checks. Vilhuber (2020) argues that only a fraction of this work will be published, and

this fraction might be biased towards non-confirmative findings. This is certainly true. A replication registry documenting both successful and unsuccessful replication attempts, also by students, would help to obtain a deeper understanding of classroom replications. We would contend, though, that the sensitive work of policing replications in Merton's spirit should not be left to students who are then supposed to confront powerful original authors with potential problems. Experienced scholars should engage in policing replications as well.

To reach this, mainstreaming the term 'policing replication' will not be enough. More important are incentives for researchers. Our claim (based on Clemens, 2017) that authors of replications should assume the burden-of-proof is at odds with the current novelty norm in economics. This novelty norm makes it difficult to publish replications in journals that pay off for academic careers. Moreover, policing replications are often perceived as hostile in the profession. Both of these make it a risky career strategy, especially for young scholars (Hamermesh, 2017; Janz, 2015).

Clearly, reforms and incentives are needed to catalyse a cultural change. Coffman et al. (2017) argue that change must come from the top down, and they call on journals to offer a regular section for replications in each issue. Not least, "*citations to the original paper [should] include citations to its replication*" (Coffman et al., 2017). This might also lead to a positive feedback loop by countering the argument that publishing comments and replications are costly for editors as they dedicate scarce journal space to papers that are then hardly cited (Whaples, 2006).[12] Alternative approaches encompass reforming promotion incentives within academic institutions. Tenure decisions, for example, could be based not only on publications but also on the replicability of the candidate's work and whether she has conducted replications herself.

We also reiterate a very simple and straightforward proposal made by Clemens (2017): The *American Economic Association* and the *Journal of Economic Literature (JEL)* could add

[12] On this note, to the extent editors are incentivezed by impact factors, these metrics could be modified so that journals are not punished for publishing replications. For example, impact factors could exclude replications from their calculation or even credit them positively. A precondition for this, obviously, is a universally accepted defintion of what consitutes a replication – something that is currently not in sight.

explicit codes to the JEL code structure on the different types of replications. This would help to clarify the terminology and at the same time signal that replications are endorsed by the profession's flagship association. While we acknowledge that there is no logically superior definition of the different sub-types of replications, the Clemens nomenclature lends itself to this new JEL code structure, potentially complemented by the policing category proposed in the present paper. Such JEL codes would also facilitate finding replications and hence including them in systematic reviews and overview articles (Coffman et al. 2017).

Furthermore, leading economics journals should make explicit whether they generally accept or even encourage comments and replications.[13] At the very least, we believe, if a paper published in a journal is replicated or commented on, the original paper's website should include links to these comments – something that is standard in other professions but currently not the case for many economics journals, including those of the *AEA*. As of December 2021, the *AER*, for example, does not even provide links on the replicated paper's website if the comment was published in the *AER* itself.

[13] See the *Institute for Replication's* website for a recent survey among editors of leading journals: https://i4replication.org/publishing.html

References

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2017). Economic Research Evolves: Fields and Styles. *American Economic Review*, *107*(5), 293–297.

Berry, J., Coffman, L. C., Hanley, D., Gihleb, R., and Wilson, A. J. (2017). Assessing the Rate of Replication in Economics. *American Economic Review*, *107*(5), 27–31.

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

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, *8*(1), 1–32.

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., and Wu, H. (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science*, *351*(6280), 1433–1436.

Christensen, G., and Miguel, E. (2018). Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature*, *56*(3), 920–980.

Christensen, G., Wang, Z., Paluck, E. L., Swanson, N., Birke, D. J., Miguel, E., and Littman, R. (2020). Open Science Practices are on the Rise: The State of Social Science (3S) Survey. *Working Paper*.

Clemens, M. A. (2017). The Meaning of Failed Replications: A Review and Proposal. *Journal of Economic Surveys*, *31*(1), 326–342.

Coelho, P. R. P., De Worken-Eley III, F., and McClure, J. E. (2005). Decline in Critical Commentary, 1963-2004. *Econ Journal Watch*, *2*(2), 355–361.

Coffman, L. C., Niederle, M., and Wilson, A. J. (2017). A Proposal to Organize and Promote Replications. *American Economic Review*, *107*(5), 41–45.

Conley, T. G., and Udry, C. R. (2010). Learning about a New Technology: Pineapple in Ghana. *American Economic Review*, *100*(1), 35–69.

Dahal, M., and Fiala, N. (2020). What Do We Know about the Impact of Microfinance? The Problems of Power and Precision. *World Development*, *128*, 104773.

Ferraro, P. J., and 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.

Fink, G., McConnell, M., and Vollmer, S. (2014). Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures. *Journal of Development Effectiveness*, *6*(1), 44–57.

Foote, C. L., and Goetz, C. F. (2008). The Impact of Legalized Abortion on Crime: Comment. *Quarterly Journal of Economics*, *123*(1), 407–423.

Freese, J., and 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.

Hamermesh, D. S. (2017). Replication in Labor Economics: Evidence from Data and what it Suggests. *American Economic Review*, *107*(5), 37–40.

Höffler, J. H. (2017). Replication and Economics Journal Policies. *American Economic Review*, *107*(5), 52–55.

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

Ioannidis, J. P. A., Stanley, T. D., and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, *127*(605), F236–F265.

Janz, N. (2015). Bringing the Gold Standard into the Classroom: Replication in University Teaching. *International Studies Perspectives*, 392–407.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Laitin, D. D. (2013). Fisheries Management. *Political Analysis*, *21*(1), 42–47.

Magnan, N., Spielman, D. J., Lybbert, T. J., and Gulati, K. (2015). Leveling with Friends: Social Networks and Indian Farmers' Demand for a Technology with Heterogeneous Benefits. *Journal of Development Economics*, *116*, 223–251.

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

Miguel, E. (2021). Evidence on Research Transparency in Economics. *Journal of Economic Perspectives*, *35*(3), 193–214.

Mueller-Langer, F., Fecher, B., Harhoff, D., and 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.

Ofosu, G., and Posner, D. N. (2019). Pre-analysis Plans: A Stocktaking. *Working Paper*.

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

Peters, J., Langbein, J., and Roberts, G. (2018). Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity. *The World Bank Research Observer*, *33*(1), 34–64.

Sukhtankar, S. (2017). Replications in Development Economics. *American Economic Review*, *107*(5), 32–36.

Vilhuber, L. (2020). Reproducibility and Replicability in Economics. *Harvard Data Science Review*, *2*(4).

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, *18*(6), 3045–3089.

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

Appendix

Table A1: Total number of published papers and policing replications in the Top 50 economics journals between 2010 and 2020

Journal	Total number of published papers	Papers meeting formalized search criteria	Number (and share*) of policing replications	Number (and share*) of policing replications that cite the original study in the title or abstract
Quarterly Journal of Economics	446	8	2 (0.4%)	2 (0.4%)
Econometrica	727	28	1 (0.1%)	1 (0.1%)
Journal of Economic Growth	135	5	1 (0.7%)	1 (0.7%)
Journal of Financial Economics	1387	14	4 (0.3%)	4 (0.3%)
Review of Financial Studies	1197	22	1 (0.1%)	1 (0.1%)
American Economic Review	2387	90	46 (1.9%)	43 (1.8%)
Journal of Political Economy	565	18	3 (0.5%)	3 (0.5%)
Journal of Finance	774	11	2 (0.3%)	2 (0.3%)
Review of Economic Studies	607	15	0 (0%)	0 (0%)
Journal of Monetary Economics	845	135	5 (0.6%)	4 (0.5%)
Journal of Labor Economics	380	2	0 (0%)	0 (0%)
Brookings Papers on Economic Activity	260	130	7 (2.7%)	7 (2.7%)
Journal of Accounting and Economics	444	3	0 (0%)	0 (0%)
American Economic Journal: Macroeconomics	361	13	0 (0%)	0 (0%)
Journal of the European Economic Association	579	17	2 (0.3%)	1 (0.2%)
Journal of Econometrics	1682	23	2 (0.1%)	2 (0.1%)
American Economic Journal: Applied Economics	436	3	2 (0.5%)	2 (0.5%)
RAND Journal of Economics	393	5	1 (0.3%)	0 (0%)
Review of Economics and Statistics	852	15	4 (0.5%)	4 (0.5%)
Journal of Applied Econometrics	668	65	55 (8.2%)	50 (7.5%)
Economic Journal	979	15	1 (0.1%)	0 (0%)
Journal of International Economics	947	35	2 (0.2%)	1 (0.1%)
Journal of Financial Intermediation	310	1	0 (0%)	0 (0%)
Journal of Business Venturing	503	4	0 (0%)	0 (0%)
Journal of Business & Economic Statistics	598	60	2 (0.3%)	1 (0.2%)
Journal of Public Economics	1272	17	5 (0.4%)	2 (0.2%)
World Bank Economic Review	363	3	0 (0%)	0 (0%)
Journal of International Business Studies	712	24	0 (0%)	0 (0%)
Journal of Development Economics	1058	17	3 (0.3%)	2 (0.2%)
Experimental Economics	415	22	10 (2.4%)	10 (2.4%)
American Economic Journal: Economic Policy	456	3	0 (0%)	0 (0%)
Journal of Environmental Economics and Management	741	11	1 (0.1%)	1 (0.1%)
Journal of Law and Economics	335	1	0 (0%)	0 (0%)
Journal of Human Resources	388	5	2 (0.5%)	2 (0.5%)
World Bank Research Observer	97	4	0 (0%)	0 (0%)
Journal of Risk and Uncertainty	260	7	2 (0.8%)	1 (0.4%)
Research Policy	1605	33	2 (0.1%)	1 (0.1%)
Journal of Economic Surveys	506	12	3 (0.6%)	3 (0.6%)

International Economic Review	610	13	0 (0%)	0 (0%)
International Journal of Central Banking	484	6	0 (0%)	0 (0%)
European Economic Review	1346	33	5 (0.4%)	5 (0.4%)
Review of Economic Dynamics	533	19	0 (0%)	0 (0%)
Total	29,643	967	176 (0.6%)	156 (0.5%)

Note: We used the Top 50 journals as listed on https://ideas.repec.org/top/top.journals.simple.html, accessed last on July 28, 2021. We excluded journals of federal reserve banks because they are not listed on Scopus ("*Proceedings, Federal Reserve Bank of Cleveland*", "*Proceedings, Federal Reserve Bank of San Francisco*", and "*Quarterly Review, Federal Reserve Bank of Minneapolis*"). We further excluded the *Journal of Economic Perspectives*, the *Journal of Economic Literature*, the *Annual Review of Economics*, and *Foundations and Trends in Econometrics* because they are review journals and unlikely to publish replication studies. The *Journal of Economic Theory* is excluded because we concentrate our work on empirical studies. The *Journal of Business* was discontinued in 2006 and is therefore not part of this table, either. Two journals were jointly listed on rank 21, which is the reason for arriving at 42 journals in total. *Share of policing replications in the total number of published papers.

Table A2: Policing replications in Berry et al. (2017), Sukhtankar (2017), and Mueller-Langer et al. (2019)

Paper	Number of replications as coded in paper (1)	Number of policing replications (our coding) (2)	Share of policing replications (3)	Replication rate according to paper (4)	Policing replication rate [(3)*(4)] (5)
Berry et al. (2017)	52	1	2%	60.0%*	1.2%
Sukhtankar (2017)	71	50	70%	6.2%[+]	4.4%
Mueller-Langer et al. (2019)	130	121	93%	0.1%	0.1%

Note. *The reported replication rate from Berry et al. (2017) presented here is the one in which the authors include any of the three categories branded by them as "Replication", "Extension", or "Robustness". [+]The reported replication rate from Sukhtankar (2017) is based on a review of both peer-reviewed replications and working papers.