Assessing the Rate of Replication in Economics

By James Berry, Lucas C. Coffman, Douglas Hanley, Rania Gihleb and Alistair J. Wilson*

Replications are a key component in the scientific process; helping the profession sift robust empirical findings from mistakes. However, while replications are desirable, there remains uncertainty over just how often they occur. Our paper is a coarse attempt to shed some light on this uncertainty. Focusing on all empirical papers in the *American Economic Review's* (AER) centenary volume, we find that a minority of the papers (29 percent) were replicated (to some degree), while a slight majority had either been replicated or extended (59 percent). In a complementary series of surveys, we find that about half of authors are confident of whether or not there own paper has been replicated, and average confidence of experts on the specific topic is even lower.

Our measurements here are complimentary to two other papers in this issue: Sukhtankar (2017), which examines replications of development papers; and Hamermesh (2017) which examines ten high-profile papers in labor. Where each of these papers examines a particular sub-field, our own work surveys all empirical works in a single volume for a top general-interest journal.

The definition of a "replication" is admittedly somewhat mercurial. In some fields a replication is an attempt to verify the original paper's results with the same data. For example, a graduate student might redo the analysis in order to better learn a technique and detect an error in the original code. Alternatively, a replication might reproduce the original paper's experiment in the field. For example, a treatment from the original paper could be used as a control in a follow-up paper that focuses on extending the original. To capture the variety of replication attempts, in our main coding exercise we take a top-level approach, defining a replication as any project that reports results that speak directly to the veracity of the original paper's main hypothesis.[1]

Our first data exercise manually codes published papers economics journals that cite recent, well-published papers. Cognizant that the focus on published papers may be too narrow, our second exercise surveys the authors of the original papers and a subsample of the citing papers to measure their beliefs and awareness of replications in the larger literature.

Manual Coding Sample

Our measurements examine the AER's 100th volume, published in 2010. The final *Volume Sample* for which we measured the rate of replication was given by 70 empirical papers.[2] We collated all published works citing a paper in our *Volume Sample* via *Web of Science* (WoS) in June of 2016. Every paper in the *Volume Sample* therefore had at least five years since publication to accrue citations. In total there were 2,945 citing papers. Restricting the citing papers to come from a top-200 economics journal (using WoS impact factors) lead to a final sample of 1,558 citations, which we refer to as our *Citing Sample*.[3]

* Berry: Cornell University. Coffman: Harvard University. Hanley: University of Pittsburgh. Gihleb: University of Pittsburgh. Wilson: University of Pittsburgh. We are thankful to Muriel Niederle for her advice and feedback on this project. In addition, we are grateful to the authors from our volume and citing sample that provided feedback and information on their works.

[1] As we were aware this approach leads to some subjectivity, we also measure replications using the narrower definitions in Clemens (forthcoming).

[2] In total the volume contains 223 papers. To focus on empirical replication of peer-reviewed original work we excluded: two Nobel addresses; 119 articles from *Papers & Proceedings*; nine comments/replies; and 23 papers that were purely theoretical in nature.

[3] Each *Citing Sample* entry can be thought of as a directed edge between a citing paper and a *Volume Paper*, as some papers cite multiple *Volume Sample* works.

Table 1—: Citation counts for our Volume sample ($N = 70$)

	Mean	**Min**	**Median**	**Max**
Google Scholar (GS)	227.6	7	139	1,246
Web of Science (WoS)	42.1	1	28.5	195
Top-200 Economics journal (WoS-200)	22.3	0	15	108

The published citation counts for *Volume Sample* papers had substantial variation. Table 1 provides summary statistics for the number of Google Scholar (GS) cites, the number of WoS citations, as well the top-200 Economics citations that we used to generate our *Citing Sample*. The median paper has 139 GS citations, 28.5 WoS citations, and 15 top-200 citations.

The *Citing Sample* papers were divided among the projects' coauthors for coding maximizing where the coders were specialists in the field. After accounting for a small number of citing papers which were not available (or in one case, not in English) a total of 1,546 papers were coded by the five coauthors.[4]

Every paper was coded according to: (i) the coder's subjective opinion on whether or not the paper was a replication of the relevant volume paper; (ii) the coder's opinion on whether the paper was an extension of the relevant volume paper; and (iii) three variables reflecting whether the paper used the same statistical model/specification, used the same data sample, and/or used data drawn from the same population as the relevant volume paper. These final three variables were recorded so that we could encode the more-concrete definitions of a replication/robustness tests in Clemens (forthcoming).

Manual Coding Results

Of the 1,546 citing papers, 52 were coded as replications. Hence roughly three and a half out of every one hundred citations contain content that speak to the veracity of the original result. Across the 70 *Volume Sample* papers, 29 percent have at least one citing paper coded as a replication attempt. Conditional on being replicated, the average number of replications per paper is 2.6. Though most papers with replication attempts have very few—eleven have just one, and three have two—five papers (7 percent) have five or more replications.[5]

In addition to our replication coding, we also report results in Table 2 for three alternative measurements (both for the *Volume* and *Citing Samples*): (i) extensions, coded as testing a closely related hypothesis to the original paper; (ii) *robustness* tests à la Clemens (forthcoming), which include altered econometric specifications on the same data or population and the same specification on a different population;[6] and (iii) *Any* of replication/extension/robustness.

In total 42 of the 70 volume papers have one or more citation coded as a replication/robustness/extension. Though this represents 60 percent of the empirical papers in the AER volume, the majority of this follow-up work is coming through robustness tests or extensions. Moreover, of the papers coded as being replications, just eight are papers with a focus on replicating the volume paper, where the remaining 44 are embedded within wider-scope papers.[7]

[4]One paper in the *Volume Sample* and one paper in the *Citing Sample* had authors in common with the present paper; neither were coded by the overlapping coauthor.

[5]Replication work is typically independent, not being produced by the original authors. Forty-eight of the 52 replicating papers (92 percent) have no authors in common with the original volume paper.

[6]Across the 1,546 coded citing papers we find: no verifications (using the same data and econometric specification) and two reproductions (the same econometric specification but a with a new dataset drawn from the same population, though both require a broad interpretation of same population).

[7]For each coded replication, we also examined whether the citing authors explicitly present the component of their work we code as a replication using the

Table 2—: Coding Rates

	Replications	Extension	Robustness	Any	N
Volume Sample	28.6% (20)	48.6% (34)	40.0% (28)	60.0% (42)	70
Citing Sample	3.4% (52)	7.8% (121)	4.7% (73)	11.0% (170)	1,546

(a) Citing Sample (b) Volume Sample

Figure 1. : Replications/Extensions by Field

Figure 1(a) breaks out the rate of any replication or extension by volume paper's field. Figure 1(b) provides a parallel illustration for the *Citing Sample*, indicating the fraction of citing papers (by field) that are replications or extensions. Though sample sizes are small, some patterns emerge. Just over half of the behavioral/experimental papers in the volume had at least one replication attempt. All other fields saw a published replication attempt in a minority of cases (between 12 and 33 percent of *Volume Sample* papers). In contrast to replications, citing papers extend the original work across all fields at much higher rates, with between 38 and 67 percent of the volume papers extended.

One consistent predictive variable for whether a particular volume paper is replicated is the number of times that paper is cited. In Figure 2(a) we illustrate the effect by graphing WoS citation CDFs for volume papers with no replications and for those with one or more. Here the figure clearly illustrates a stochastic ordering.[8] In particular, the figure shows that all papers in our sample with more than 100 published citations have replications. One interpretation is that the profession does a good job of replicating important findings.[9] It could also be that fields with more papers and citations also replicate work at a higher rate.

Volume paper replications accrue uniformly across the measured period. This is illustrated in Figure 2(b), which shows the cumulative fraction of volume papers with one or more replication from 2010 (the volume's publication year) to 2016 (the year we collected data on citations).

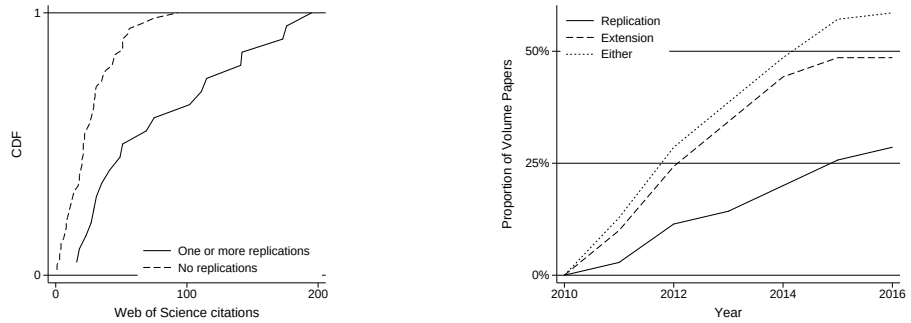
Survey Sample

The previous analysis measures replication attempts through the subjective judg-

word "replication" or similar. Eighteen of the 52 coded replications are explicitly presented as such. At the volume level, the proportion of papers with at least one explicitly presented replication was 13 percent, much closer to the replication rate found in Sukhtankar (2017).

[8]The estimated marginal effects from a probit suggests that ten additional published citations increases the chance that a volume paper has a replication by 5.2 percent (significant at the 1 percent level).

[9]Hamermesh (forthcoming) shows that the number of citations after the first five years is predictive of long-run impact, though citations in the second five years is a better predictor.



(a) Citation distributions

(b) Cumulative time trends

Figure 2. : Volume paper replications: Time and Citations

ment of this paper's five coauthors, researchers in the field reading published work in Economics. Though a starting point, the estimates may err for a couple of reasons. First, there may be replications not in the sample—unpublished work, papers not in Economics, undistributed graduate-student projects, etc. Second, the judgment of what constitutes a replication attempt may vary between the coders, as well as with those with more specific expertise on the topic.

In the second part of our data collection we attempt to corroborate our first measure, assess economists' awareness of replications and get a larger sense for how many replications might be out there. To these ends, we surveyed two sets of authors. First, we sent personal emails to one author from each *Volume Sample* paper. Second, we sent a link to an online survey to authors of papers in the *Citing Sample*.[10,11]

We elicited the authors' beliefs in both samples over the number of replications for the relevant volume paper that were: (i) publications, (ii) working papers, and (iii) projects never meant to be published.[12] Here our language in the elicitation was purposefully broad asking for the number of papers that "report a result that speaks directly to whether or not your paper's main hypothesis is true." For each response, we also asked about their confidence: "Do you think this number is pretty close, or is it more of a complete guess?" In total, one author from 37 of the 70 *Volume Sample* papers responded, and 58 of the 226 *Citing Sample* authors surveyed completed it.

Survey Results

Overall volume authors were not sure on how many replications of their work had occurred. For the 26 *Volume Sample* authors that reported their confidence, just over half (14) were sure on their responses. Examining responses from the 14 confident authors we find substantial concurrence with our manual coding. All eight papers where a volume author pointed to one or more published works have replications or extensions coded in our data (5 as replications, 3 as extensions). For the remaining six papers where a volume author was sure there had been no replication, five were coded as

[10] Data collection for authors in the *Volume Sample* was more informal, with clarifying questions answered over email, and some qualitative responses reported. Data collection from the *Citing Sample* did not have as much two-way communication, and elicited numerical responses.

[11] In particular, we sent surveys to one author on each citing paper (with independent authors from the volume paper) that: (i) was coded as a replication; (ii) was coded as an extension; and (iii) cited the volume paper most often (and at least twice) but was coded neither as a replication nor an extension (95, counting ties). The goal of these criteria were to narrow the sample to authors who knew the volume paper well but were independent of the original.

[12] Authors in the *Volume Sample* were also asked how many of the working papers they believed would be published at some point.

having no replications.[13] In addition, in all cases where a volume author pointed to *specific* published work that was contained in our *Citing Sample* the referenced work was always coded as either a replication or an extension.

Authors who were guessing on the number of published replications were generally more optimistic. For the 11 volume authors who report a guess, only one guesses that their paper has no published replications; in contrast for the 14 authors who were sure in their numbers, six report no published replications.

Citing authors reported even less confidence in their knowledge of replication attempts. On a 0 (complete guess) to 100 (very sure) percentage scale, *Citing-Sample* respondents report average confidence of 35 percent for published replications, 27 percent for working papers, and 15 percent for informal projects.[14] Though this sample was hand-selected to be experts on these specific literatures, there seems to be little confidence over the extent of replication work.

The authors of citing papers estimate high rates of replication. The median belief on published replication attempts is two replications per volume paper, with almost three-quarters reporting that the paper they cited had at least one published replication. That rate increases to 83 and 78 percent for working papers and informal projects, respectively.

The surveyed rates of replication are substantially higher than our coded sample. Where our manual coding suggests replications for just under three-in-ten papers, the survey responses suggest closer to seven-in-ten.[15] This difference could reflect narrowness in our coding—only using top-200 Economics papers—or differences in our subjective judgment of what constitutes a replication. However, considering the authors' beliefs to be the authoritative measure has to come with the qualification that the authors admit to being uncertain, and their estimates will be mechanically biased towards more replications than zero.

[13]The only stand-out was explained within the volume-author's response using an explicitly stricter definition of replication, where the number was positive on a broader definition.

[14]The large majority are on the unsure side of the scale: 74 percent report a confidence of 50 or below for published replication work. That proportion increases to 84 percent for confidence on working paper replications, and 97 percent for informal projects.

[15]The proportion of volume papers with author-reported replications does not increase substantially if we widen the net to include unpublished work.

Conclusion

Examining a set of well-published papers, and surveying experts on the specific topics, we find substantial uncertainty over how many replication attempts exist. As an attempt to shed some light on this uncertainty, the estimates from our coding exercise suggest that a majority of very well-published papers in Economics are not being replicated at all—though well-published *and* well-cited works are being replicated at much higher rates.

There are reasons to suspect that the true rate of replication might be higher or lower than the proportion we estimate. However, the measurements in our paper reflect very practical numbers: What do economists inside the literatures believe? What can be found through a search of the literature by those outside the literature?

REFERENCES

Clemens, Michael A. forthcoming. "The meaning of failed replications: A review and proposal." *Journal of Economic Surveys.*

Hamermesh, Daniel S. 2017. "Replication in Labor Economics: Evidence from Data and What It Suggests." *American Economic Review*, 107.

Hamermesh, Daniel S. forthcoming. "Citations in Economics: Measurement, Uses and Impacts." *Journal of Economic Literature.*

Sukhtankar, Sandip. 2017. "Replications in Development." *American Economic Review*, 107.