

Review

Replication Papers

Peter Harremoës [†] 

Niels Brock Copenhagen Business College, GSK department, 1358 Copenhagen, Denmark; harremoës@iee.org
[†] Current address: Rønne Alle 1, st, 2860 Søborg, Denmark.

Received: 31 December 2018; Accepted: 17 July 2019; Published: 22 July 2019



Abstract: Reproductions and replications of experiments and surveys are important for ensuring the healthy development of modern science. The so-called replication crisis is a problem that needs to be addressed in various ways. In this paper, we propose to make a special category for replication papers, where the focus should be to verify or falsify the results of previously-published experiments or surveys. We also propose some guidelines for the types and content of replication papers.

Keywords: repetition; reproduction; replication; reproducibility; replicability

1. Introduction

Reproducibility was first discussed by Boyle and Huygens in the 17th Century [1] and has been a corner stone in experimental sciences ever since. A recent report from the Royal Dutch Academy of Sciences [2] also emphasized this point, and in the report, one can find various advice regarding good practice to enhance replicability. In traditional experimental sciences like physics and chemistry, reproduction of important experiments is an integrated part of education at all levels, while this is not the case in social sciences, where the use of systematic experiments and surveys was introduced much later. A replication crisis has been declared in certain social sciences, but even in more traditional experimental sciences, replicability sometimes turn out to be a problem.

Academic publishing is an activity that is growing fast with more than 28,100 peer reviewed journals publishing more than 2.5 million papers each year [3]. Surprisingly, the publication rate per researcher has not gone up for the last 100 years if we take into account that it has become more common with papers with multiple authors [4]. To a large extent, this huge number of papers is needed in order to run the academic merit system that is important for the career of a steadily increasing number of people in academic positions. Each year, more than 225,000 PhD students are educated worldwide [5], and as part of their education, they are expected to produce several papers. A recent survey found that a modern PhD could be given for anything between one and 12 publications, and the average number of publications was 4.5 per PhD thesis [6]. To become a professor, one would have to have more publications, but the numbers depend greatly on the specific field. Official regulation will only give minimum numbers. In India [7], assistant professors need at least two papers, associated professors need at least seven papers, and full professors need to have produced at least 10 papers. In very competitive fields, the number of publications may be much higher. Although there are great variations between different academic fields, these figures illustrate why the expression “publish or perish” has become important advice to young ambitious researchers.

The pressure on researchers to publish has led to much low-quality research. The current situation was summarized by B.Nosek, who stated that “There is no cost to getting things wrong. The cost is not getting them published” [8]. Currently, any piece of research can be published somewhere. If a manuscript is rejected from one journal, the author may submit it to another journal. If the manuscript is improved and then submitted to a journal that better fits to the topic of the manuscript, then resubmitting a rejected manuscript is recommended [9,10]. The problem is that even fake articles

can sometimes go through a peer review and get published in journals with high impact factors. This was first demonstrated by the Sokal affair [11], and later, similar hoaxes followed [12].

During the last decade, there has been an increasing awareness that in some research areas, most experiments and surveys were never replicated. When people try to replicate them, the success rate was kind of depressing. For this reason, a replication crisis has been declared, and there have been several attempts to do something about it. The problem was addressed in a recent opinion paper in this journal [13]. Here, we shall focus on what the publishers can do.

2. Terminology

Before we go any further, we have to be precise about our terminology. Goodman et al. had a longer discussion [14]. Repeatability means the extent to which one can specify the details of an experimental setup to diminish the variations in the observations. The more details one gives on when, where, how, and who did the experiment, the more repeatable the experiment will be.

In statistical modeling, one typically aims at independent identically distributed variables. This is equivalent to having exchangeable repetitions of the experiment. The more one aims at repeatability, the more restricted are the conclusions that one can draw from the experiments. A lack of repeatability may be due to bad control over the parameters of the experiment, *p*-value hacking, cherry picking, random effects (in 5% of the cases), or fraud.

In our terminology, repetition is a theoretical model about how a certain dataset is collected and how the statistical analysis is performed. Repetitions are opposed to other models of data collection like time series, where the order in which the data are collected matters. When we say repetition, we claim that the order does not matter or perhaps there is no order. Different parts of the dataset may be collected simultaneously at different locations. In practice, some parameters may change during an experiment, and this will often be the source for statistical variations in the data. When we use repetition as a model, we claim that these changes are of an unsystematic nature that can neither be predicted nor controlled. It is quite common that some patterns in the data lead us to reject our repetition model.

An experiment is reproducible if a similar experiment will support the same conclusions. The more variation that is allowed in an experimental setup that still supports the same conclusion, the more valid is the conclusion. Checking that minor variations in the parameters have no influence on the result is sometimes called a robustness check. If an experiment or survey is described in sufficient detail, it should be possible to reproduce it. If it is not possible to reproduce an experiment, it is an indication that the authors of the original study overlooked or omitted details that influenced the result.

In principle, one may merge a data sequence from a reproduction together with the data sequence from the original experiment in order to obtain a longer data sequence. In practice, however, one will often keep data from the original study and the reproducing study apart and use the new data sequence as an independent confirmation of the conclusions of the original study. The point in keeping the data sequences apart is that subsequent knowledge or analysis may show that certain details in the different studies turn out to be important.

In a statistical analysis, one will often do statistical testing in order to provide evidence for a certain interesting hypothesis *A*. This is typically done by rejecting a null hypothesis H_0 , where the null hypothesis states that there is no effect. It is important to note that the null hypothesis H_0 typically represents a conservative view or skeptical view, while the alternative hypothesis *A* represent some kind of knowledge that is claimed to be valid. A statistical test is then designed so that it gives a certain significance level (often 5%). Such a statistical analysis only makes sense if one aims at reproducibility. If the researchers in a certain scientific field work with a significance level of 5%, then in 95% of all cases where a true null hypothesis has been rejected, a later reproduction will confirm the original study, while in 5% of the cases, later studies will demonstrate that the original conclusion was false.

Often, it is neither possible nor desirable to make a reproduction of an experiment. Nevertheless, it might still be relevant to make similar studies or experiments in order to support the overall conclusion of a study. Reproducibility is in contrast to replicability, which means the ability to achieve similar conclusions independently (but not identical conclusions), when there are differences in sampling, research procedures, and data analysis methods. In some sciences, the state of the world may change so that repetition or reproduction is not possible or not desirable. In cases where the goal of a study is to obtain new conclusions, one should not call it a replication.

3. The Replication Crisis

B. Nosek illustrated how publication practices favored novelty and positive results over consolidation of existing knowledge [15]. The problem of the lack of replication studies in psychology was emphasized already in 1970 [16].

During the last decade, there has been an increased awareness that many published research studies can neither be reproduced nor replicated. The problem has been addressed in various meta-studies, and here, only a brief overview will be given with some pointers to the literature. It is difficult to give precise numbers for a problem of this nature, and it is also difficult to compare different fields, because different fields have different standards for peer review, operate with different significance levels, etc. Therefore, the numbers given below should be taken with great caution.

In a sample of 1576 scientists, more than 70% had failed to reproduce another scientist's experiments, and more than 50% had failed to reproduce their own experiments [17]. It was also reported that 24% succeeded in publishing reproduced experiments and 13% published unsuccessful reproduction attempts. Not all succeeded in publishing reproduction attempts: 12% got successful reproductions rejected, and 10% got unsuccessful reproduction attempts rejected. In a survey, 270 scientists ranked the main problems that modern science is facing [18]. The statement "Replicating results is crucial and rare" was ranked as the third most important problem after "Academia has a huge money problem" and "Too many studies are poorly designed".

Below are some examples of academic disciplines where the replication crisis is well documented. The problem also exist in other disciplines, but here, we will just document that there is a problem that has to be solved. It is much harder to say anything precise about the full extent of the problem and its impact on society. An extensive bibliography on the replication crisis was compiled by Rathemacher [19].

3.1. Medicine

Out of 49 medical studies from 1990–2003, with more than 1000 citations, 45 studies claimed that the studied therapy was effective [20]. Subsequent studies contradicted 16% of these studies; 16% found stronger effects than did subsequent studies; 44% were replicated; and 24% remained largely unchallenged.

In a meta-analysis from 2012 [21], only 11% of preclinical results in cancer research could be reproduced, and these results were published in high impact journals, while the papers all had a high number of citations. Surprisingly, the papers that reported results that could be reproduced had fewer citations than the papers for which it was not possible to reproduce the results. This indicated that papers with unreproducible results may initiate much research that is compromised by the negative reproduction attempts. The Centre for Open Science and the Science Exchange tried to reproduce key experiments in cancer research, and their success rate was 40% [22].

3.2. Psychiatry

In 83 highly-cited papers in psychiatry, it was claimed that certain treatments were effective [23]. Only 16 of these studies were verified by replications. In 27 of these studies, replication attempts were unsuccessful. In 11 cases, the effect was smaller than originally reported, and in 16 cases, the replication even contradicted the original paper. The rest of the studies were not replicated.

3.3. Psychology

An analysis of the publication history in the top 100 psychology journals between 1900 and 2012 indicated that approximately 1.6% of all psychology publications were replication attempts [24]. Of these studies, 500 were randomly selected for further examination and yielded a replication rate of 1.07% (342 of the 500 studies were actually replications). In the 500 studies, analysis indicated that 78.9% of published replication attempts were successful. The rate of successful replication was significantly higher when at least one author of the original study was part of the replication attempt (91.7% relative to 64.6%).

A team lead by B.Nosak tried in 2015 to replicate 100 psychological studies from journals with high impact factors [25]. They were only able to replicate 39 of these studies.

3.4. Educational Sciences

In the top 100 journals in educational science, only 221 out of 164 589 articles tried to replicate previous studies [26]. Of these, 28.5% were what we call replications, and the other ones were what the authors for the meta study called “conceptual repetitions”. Of the replications, 48.2% were performed by the same researchers that had produced the original study. When the same authors published a replication in the same journal, 88.7% of the replications succeeded. If the same authors published in a different journal, only 70.6% were successful replications. If a different author tried to replicate, the success rate was only 54%.

3.5. Social Sciences

In a recent paper, 21 experimental studies in the social sciences published in Nature and Science were replicated [27]. The original authors reviewed these replications. The sample sizes were on average about five-times higher than in the original studies. For 13 (62%) of the studies, there was a significant effect in the same direction as the original study. The relative effect size of true positives was estimated to be 71%, suggesting that both false positives and inflated effect sizes of true positives contributed to imperfect reproducibility.

4. Replication Proposal

To address the replication crisis, one should allow a type of paper called a replication paper. For instance, the Multi Disciplinary Publishing Institute, which publishes this journal, has 42 paper categories, but none of these are intended for replication papers. In a replication paper, the author should try to provide independent evidence for or against the conclusions of a published paper. The purpose of a replication paper should not be to introduce new ideas or explanations, but should only focus on providing more evidence for or against the previous conclusions. The evidence should also be of a nature that is indisputably along the same lines as the original paper. The conclusion of a replication should be that either the conclusion is verified or not, where the last case happens when the replication attempt either gives evidence that does not support the conclusion or the replication attempt turns out not to be possible due to flaws in the description in the original paper.

Some attempts have already been made in this direction. Early in the replication crisis in 2010, the journal Public Finance Review called for replication papers and provided guidelines for these papers [28]. The journal ReScience Journal publishes replication studies, but this has only lead to few replication papers. Similarly, the journal Economics allows replication papers in the area of economics and has developed detailed guidelines for this type of paper [29].

A replication paper should go through the same type of strict review process as other papers in the particular journal. If possible, one or more authors of the original paper should act as reviewers of a replication paper, but there should also be one or more review reports from independent researchers. In order to ensure that the replication papers are of the same quality as the original papers, it would be optimal if replication papers were published in the same journals as the original papers. If a replication

paper is published in the same journal as the original paper, it would also be natural to add a hyperlink for the web page of the original to the replication paper. If a replication paper is published in a different journal, the journal of the original paper may abstain from linking to the replication paper because the replication paper is published according to different editorial standards.

Some journals that focus on publishing replications have been established, but they are few and tend to have a lower impact factor than the journals where the original papers were published. If a journal publishes good papers and good replication papers, this may lead to a higher impact factor if proper cross referencing is facilitated, because any citation of the good paper may also lead to a citation of the replication paper. One should note that a replication paper that falsifies the conclusions of a previous paper would lead to an extra citation of the original paper. This problem already exists without a special category for replication papers. Often, highly problematic papers get many citations, and many bibliometric measures are not able to disguise between citations indicating the high quality of the cited paper and citations indicating the low quality of the cited paper. These problems in bibliometrics should not be an argument against replication papers.

An author of a paper may feel a strong incentive to get other researchers to replicate his/her results, so it would be extra important that the authors of a replication paper do not have a conflict of interests. Any respectable journal has rules about conflicting interests for their reviewers, and the same rules should be applied to authors of a replication paper. The authors of a replication paper should be allowed to ask the authors of the original study about the details of how the original experiment or survey was performed, and such an interaction should be reported in a replication paper.

5. Replication Types

When speaking about replications, one may think of classical experiments, but there are other studies that may also qualify as replications in that they attempt to consolidate previously-published results or perhaps falsify published results. If one allows replication papers in this more broad sense, one will need slightly different quality criteria for different types of replication papers. Here, we will comment on three quite different types of replication papers in order to illustrate the diversity of papers that may fall into the category of replication papers. A more comprehensive list with detailed guidelines may be needed for journals that allow submission of replication papers.

5.1. Experiment and Surveys

A paper that refers to observations in experiments should describe these experiments in such detail that it is possible to replicate the experiment. If an experiment is not described in sufficient detail, a replication may give different results. If this is the case, it means that the original paper did not provide sufficient evidence for the conclusions. If an experiment is reproduced with a slightly different setup and still gives the same result, it would actually give stronger evidence for the conclusions than a repetition of the experiment.

If an experiment or a survey is replicated rather than reproduced, the replication provides new knowledge, but new knowledge should not be the prime intention for a replication paper. Obviously, a replication can be part of a research article, but contrary to a replication study, the focus in a research article should be to provide new knowledge and insight.

5.2. New Proof

A paper in mathematics contains proofs of the theorems, and in principle, the reviewers should check all technical details in the proofs. In modern mathematics, the proofs are often long and rely on highly-specialized knowledge, and it may be difficult to find reviewers that have the time and expertise to check all details. Often, reviewers in mathematics check that the statements sound reasonable and only check the detailed argument if they find some statement that they find dubious. If there are errors in proofs, these are often found by the author himself or by coauthors, although they have been through a review. Since the authors of a manuscript often have the most specialized expertise

in the particular topic, the reviewers often have great trust in the author's ability to write down the proofs correctly.

A replication of a proof would be a different proof that proves the same theorem. One cannot define how different a proof should be in order to count as a different proof. That would be up to the reviewers and the editor to judge. It is already an established activity to make new proofs of published theorems. Such new proofs would normally be published together with other mathematical theory, and the novelty of the whole manuscript would be important for the decision on whether the manuscript should be published as a research article. Replication with a new proof should only be judged on whether the new proof provides confidence in the correctness of the theorem even for mathematicians that do not have the specialized expertise to follow the original proof.

5.3. Simulation and Numerical Calculations

Many technical papers contain sections with simulations or numerical calculations that are intended to illustrate certain points. The results of simulations are often given in terms of figures or tables. Often, the authors have done more simulations or numerical calculations than the ones reported in their paper, and this leaves room for cherry picking. Often, simulations and numerical calculations have given the authors some insight that they want to convey to the reader, but the insight of the authors is often more due to the process of doing the simulations or numerical calculations than due to the published figures and tables.

My experience as an editor is that reviewers rarely have really critical comments on such sections. Sometimes, they comment that it is unclear what the simulations or numerical calculations illustrate, and sometimes, they recommend that such sections be shortened or removed. Only in very few cases have the reviewers checked the calculations.

Replication of simulations and numerical calculations could be of two types. One type would be to run the calculations using different software or different hardware. Such a reproduction would check for software errors and errors in programming the software. If the computations are demanding for the hardware resources, faster hardware or specialized hardware may allow for better simulations or numerical calculations. Another type of replication would be to run the simulation or numerical calculation with slightly different parameters. This would give a robustness check of the results.

Obviously, simulations and numerical calculations can be published in other ways than in journals, but peer-review in journals is the best way of ensuring the quality of the work, and it gives the researcher the credit for the results, which is both important for the career of the individual researcher and for the funding agencies.

6. Conclusions

We proposed to introduce a paper type called a replication paper. The purpose of such a paper is to reproduce or to replicate part of the results of published papers. One should allow for more variation than defined by repeatability.

A 2016 study of replications in social and behavioral sciences concluded that few replications took place, but that these replications were often conducted in the context of teaching [30]. The authors of this study proposed that replications become an integrated part of curriculum, and for this to work, the journals should be more open to publishing replication papers. Obviously, a person that has only published replication papers should not be granted a position as a researcher. One may hope that some day, it will not be possible to get a research position in a topic with experiments and surveys without having any experience with performing replications.

Claim: Whenever authors publish a paper containing surveys or experiments that are analyzed using Neumann–Pearson-type hypothesis testing, then replication is relevant. That means that if any null hypothesis, any p -value, any significance level, or any confidence interval appears in a paper, it is relevant to try to reproduce or replicate the study.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Shapin, S.; Schaffer, S. *Leviathan and the Air-Pump*; Princeton University Press: Princeton, NJ, USA, 1985.
2. KNAW. *Replication Studies. Improving Reproducibility in the Empirical Sciences*; Advisory Report; Royal Netherlands Academy of Arts and Sciences (KNAW): Amsterdam, The Netherlands, 2018.
3. Ware, M. *An Overview of Scientific and Scholarly Journal Publishing*; STM: Montreal, QC, Canada, 2015.
4. Fanelli, D.; Larivière, V. Researchers' Individual Publication Rate Has Not Increased in a Century. *PLoS ONE* **2016**, *11*. [CrossRef] [PubMed]
5. OECD. *Science, Technology and Innovation Outlook 2016*; OECD: Paris, France, 2016.
6. Mason, S.; Merga, M. A Current View of the Thesis by Publication in the Humanities and Social Sciences. *Int. J. Dr. Stud.* **2018**, *13*, 139–154. [CrossRef]
7. University Grants Commission. *UGC Regulations on Minimum Qualifications for Appointment of Teachers and Other Academic Staff in Universities and Colleges and Measures for the Maintenance of Standards in Higher Education*; University Grant Commission: New Delhi, India, 20 July 2018.
8. Unreliable Research: Trouble at the Lab. *Economist*, 18 October 2013, pp. 26–30.
9. Durso, T. Editors' Advice To Rejected Authors: Just Try, Try Again. *TheScientist*, 15 September 1997.
10. Mudrak, B. Your Paper Was Rejected—What Next? Available online: <https://www.aje.com/en/arc/your-paper-was-rejected-what-next/> (accessed on 9 February 2019).
11. Sokal, A. Transgressing the boundaries: Toward a transformative hermeneutics of quantum gravity. *Soc. Text* **1996**, *46/47*, 217–252. [CrossRef]
12. Couronne, I. 'Real' fake research hoodwinks US journals. *Phys.org*, 5 October 2018.
13. Kun, A. Publish and Who Should Perish: You or Science? *Publications* **2018**, *6*, 18. [CrossRef]
14. Goodman, S.N.; Fanelli, D.; Ioannidis, J. What does research reproducibility mean? *Sci. Transl. Med.* **2016**, *8*, 341–353. [CrossRef] [PubMed]
15. Nosek, B.A.; Spies, J.R.; Motyl, M. Scientific utopia: II. Reconstructing incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* **2012**, *7*, 615–631. [CrossRef] [PubMed]
16. Smith, N. Replication studies: A neglected aspect of psychological research. *Am. Psychol.* **1970**, *25*, 970–975. [CrossRef]
17. Baker, M. 1500 scientists lift the lid on reproducibility. *Nature* **2016**, *533*, 452–454. [CrossRef] [PubMed]
18. Belluz, J.; Plumer, B.; Resnick, B. The 7 Biggest Problems Facing Science, According to 270 Scientists. Available online: <https://www.vox.com/2016/7/14/12016710/science-challenges-research-funding-peer-review-process> (accessed on 5 February 2019).
19. Rathemacher, A. Reproducibility Crisis Bibliography. 17 March 2017. Available online: http://digitalcommons.uri.edu/cgi/viewcontent.cgi?filename=1&article=1047&context=lib_ts_presentations&type=additional (accessed on 31 December 2018).
20. Ioannides, J. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **2005**, *294*, 218–228. [CrossRef] [PubMed]
21. Begley, C.; Ellis, L. Raise standards for preclinical cancer research. *Nature* **2012**, *483*, 531–533.
22. McRae, M. More Cancer Studies Have Just Passed an Important Reproducibility Test. Available online: <https://www.sciencealert.com/two-more-cancer-studies-have-just-passed-an-important-reproducibility-test> (accessed on 31 December 2018).
23. Tajika, A.; Ogawa, Y.; Takeshima, N.; Hayasaka, Y.; Furukawa, T. Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *Br. J. Psychiatry* **2015**, *207*, 357–362. [CrossRef] [PubMed]
24. Makel, M.C.; Plucker, J.A.; Hegarty, B. Replications in Psychology Research: How Often Do They Really Occur? *Perspect. Psychol. Sci.* **2012**, *7*, 537–542. [CrossRef] [PubMed]
25. Nosek, B. Estimating the reproducibility of psychological science. *Science* **2015**, *349*. [CrossRef]
26. Tyson, C. Failure to Replicate. *Inside Higher ED*, 14 August 2014.

27. Camerer, C.F.; Dreber, A.; Holzmeister, F.; Ho, T.H.; Huber, J.; Johannesson, M.; Kirchler, M.; Nave, G.; Nosek, B.A.; Pfeiffer, T.; et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2018**, *2*, 637–644. [CrossRef]
28. Burman, L.; Reed, W.; Alm, J. A Call for Replication Studies. *Public Finance Rev.* **2010**, *38*, 787–793. [CrossRef]
29. Snower. Replication Guidelines. Available online: <http://www.economics-ejournal.org/special-areas/replications-1> (accessed on 31 December 2018).
30. Fecher, B.; Frässdorf, M.; Wagner, G. *Perceptions and Practices of Replication by Social and Behavioral Scientists: Making Replications a Mandatory Element of Curricula Would Be Useful*; IZA—Institute of Labor Economics: Bonn, Germany, 2016.



© 2019 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).