A Community Effort Towards Reproducible Management Science

INFORMS ⁝ 16-20 minutes
DOI: 10.1287/mnsc.2023.03556, Show Details

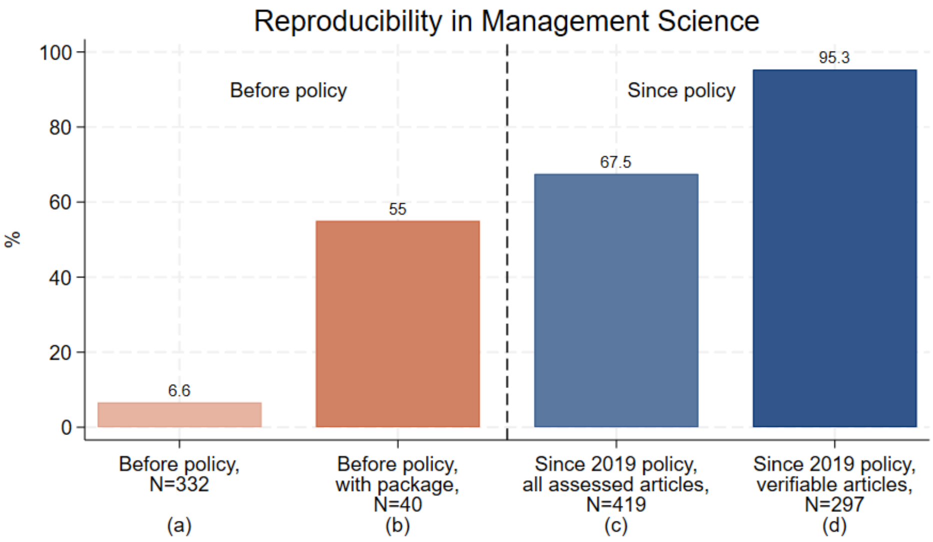
Posted by Milos Fisar, Ben Greiner, Christoph Huber, Elena Katok, and Ali I. Ozkes on 12/31/2023 at 09:59 am

Research in business and management aims to provide sound and credible evidence upon which business and policy leaders can base their decisions. But to what extent can we trust the scientific results? The answer depends on whether the results are transparently documented (reproducible) and whether they are robust and broadly applicable (replicable). While replicability is ultimately an empirical question, to be explored in further studies, reproducibility is a matter of scientific rigor, and provides the groundwork for replicability. In a recent article, *Reproducibility in Management Science*, Miloš Fišar, Ben Greiner, Christoph Huber, Elena Katok, and Ali I. Ozkes (Fišar et al. 2023) take a significant leap forward and take a closer look. They estimate, for the first time, the reproducibility of a broad range of almost 500 studies in *Management Science*, a leading academic journal in business and management.

To enable verification of scientific results, in 2019 *Management Science* introduced a policy that made it a requirement for authors to provide their study materials (that is, their data, code, and everything else needed for the empirical or computational analyses), with some exceptions applying. In the *Management Science Reproducibility Project*, the authors directed the collaborative effort of a community of more than 700 experts from relevant fields of research to (attempt to) reproduce a large and representative sample of articles published before and after this policy change. The findings of this project, reported in Fišar et al. article, provide a description of the current state of affairs, highlight the critical role of disclosure policies in scholarly research, and allow us to put forward suggestions for improving the reliability of research results.

Figure 1 shows the main results of Fišar et al. endeavor: the percentage of studies that can be fully or largely reproduced, both before and since the introduction of the disclosure policy.

Figure 1. Reproducibility rate before and since the introduction of the 2019 code and data disclosure policy (percentage)

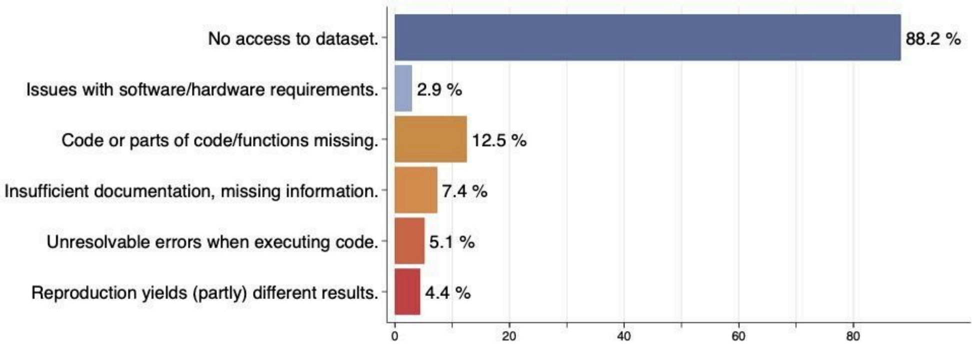


Consider the initial situation before the introduction of the policy, when providing code and data was voluntary. Because for 88% of articles, materials needed for reproducing a study’s results were not made available, only 7% out of 332 studies could be reproduced (see panel (a) in Figure 1). Among the 40 studies for which the authors did voluntarily provide materials, the reproduction rate is at 55% (see panel (b) in Figure 1).

In a sample of 419 studies published since the introduction of the policy and until January 2023, the authors observe a remarkable improvement: reproducibility climbed to almost 68% (see panel (c) in Figure 1). When, in addition, all data were available to the assessors and they could meet the soft- and hardware requirements, 95% of articles could be reproduced (see panel (d) in Figure 1).

These results reveal that the largest challenge to reproducibility since introduction of the disclosure policy is data accessibility. For a significant number of studies in Fišar et al. sample, data were not available to the assessors: the datasets may have been under NDA, not available for privacy reasons, or come from subscription databases or other commercial sources to which the assessor did not have access. Figure 2 displays the main reasons for limited reproducibility. Besides data accessibility, obstacles hindering reproduction include issues such as missing or incorrect code, insufficient documentation, and the complexity of technical requirements.

Figure 2. Reasons for non-reproducibility



Fišar et al. findings emphasize the critical importance of data and code disclosure policies in academic journals. Such policies seem not only to be associated with a considerably higher rate of reproducibility, but also encourage a culture of openness and integrity in academic publishing. They are essential for producing reliable and trustworthy research, which in turn informs sound decision-making in practice.

Several concrete steps can be taken to elevate reproducibility rates further. First, enhancing the data availability through various means such as including de-identified data in replication packages, forming agreements with subscription databases for data access, or providing data through specialized infrastructures that restrict use to specific purposes. Second, refining the review processes for code and data. This might involve making the acceptance of papers conditional upon the approval of replication packages, and integrating the code and data review as an essential step in the manuscript review process at academic journals. Third, professionalizing the code and data review either in-house at the journals or publishers, or by delegating reproducibility certification to specialized third-party agencies.

Such institutional reforms, along with a collaborative effort and awareness across the academic community, are key in enhancing the robustness and reliability of results published in academic journals in business and economics. *Management Science* has already gone a long way from when hardly any study materials used to be available to an enforced disclosure policy that requires that each article provides study materials (even if it allows for exceptions). However, this journey needs to be continued further, with sufficient resources made available by the publisher(s), to bring reproducibility to 100%.

Reproducibility is an essential feature of reliable research results, but it cannot guarantee replicability. It does not imply that redoing a study – in a different context, with different data, analyses, or research designs – will yield the same outcomes and conclusions. However, reproducibility lays the foundations, ensuring validity of reported results and provision of materials that enable replication attempts and robustness checks, thus supporting our aspiration of reliable and credible scientific evidence.

Reference:

Miloš Fišar, Ben Greiner, Christoph Huber, Elena Katok, Ali I. Ozkes, (2023) Reproducibility in *Management Science*. https://doi.org/10.1287/mnsc.2023.03556.

The Editor-in-Chief, David Simchi-Levi, asked two economists familiar with recent reproducibility studies, **Lars Vilhuber,** Data Editor, American Economic Association & Cornell University and **Sofia Encarnación**, Cornell University, to reflect on the uniqueness of the approach and importance and of the paper to the management science community and beyond. Below are their comments.

Comment on “Reproducibility in Management Science”

Lars Vilhuber and Sofia Encarnación

Fišar et al. describe the outcomes of a very exciting project: more than 700 scientists (the “Management Science Reproducibility Collaboration”) attempted the reproduction of nearly 500 articles that had provided replication packages in the course of the publication of their article. Reviewers were part of a “mega” project that used existing published replication packages, and attempted to reproduce the main results. Fišar et al. find that if accessing data, software, and resources are not a constraint, the vast majority of articles can be reproduced by faculty and advanced graduate students. Comparing to a small sample of articles that predate the journal’s current data and code availability policy, this is an enormous improvement. Failure to reproduce is due to the usual list of suspects: inaccessible data (as reported by reviewers), missing or faulty code, and in some cases, insufficient computational resources. These results are broadly consistent with both other published and unpublished attempts, such as our own (Herbert et al. 2023a; 2023b), our internal work at the American Economic Association (where we manage the work of the AEA Data Editor team), and observations from numerous replication games (see i4replication.org). Fišar et al. contribute to these findings by having a much more sophisticated mechanism to match skills and interests with to-be-reproduced articles, which should in principle improve the reproducibility, and should also, again in principle, alleviate the problems of access to data, software, and compute resources. They also provide very useful self-reported estimates of burden in terms of hours spent working on the reproducibility check.

Putting aside the numerical estimates provided, we want to emphasize two key points.

First, let’s consider the process itself, together with the burden on reviewers. Currently, there are a few different approaches of what one might call “enhanced reproducibility checking” across journals, ignoring the old school method of simply requesting unverified replication packages to be dumped onto a journal’s website. First, and requiring the least amount of effort, are “table-top” verifications such as those performed by Management Science data editors and a few other journals (in economics, we’ll highlight the Canadian Journal of Economics). The data editor simply verifies the contents of the provided replication package, using their experience to identify shortcomings or missing files. A more “intense” method then actually acquires any necessary data and runs code, prior to the publication of *every* manuscript otherwise acceptable to the journal. This method is in use at our journals, as well as a number of top economics and political science journals, but is also applied in some physical sciences. It can be outsourced (for instance, to institutions such as the Odum Institute or cascad, as mentioned in the article). In practice, this is equivalent to what the Management Science Reproducibility Collaboration did here, with one key difference: at our journals, this process, and any adjustments by the authors to the replication package (and the manuscript), happens **before publication**, whereas Fišar et al and the Management Science Reproducibility Collaboration did this **after publication of the manuscript**. So, a very good question to ask here is: **does the timing of the verification matter**?

There are a few indications that it should not matter. Consider two small facts reported in Fišar et al: Footnote 10 mentions that “the journal allows authors to submit an improved replication package”, and “115 authors [about a third]… submitted comments” (pg. 8). Our journals have a similar “revision policy” (American Economic Association 2019), and we receive requests to update or correct existing replication packages about a dozen times each year. Thus, when provided with evidence - before or after publication - that there are issues with their replication package, authors are quite willing to update it. Currently, such updates are occasional and ad-hoc events, but making them formal and regular - possibly through the aforementioned replication games - could induce authors of replication packages to internalize the risk of being verified much more strongly.

There are two important policy levers that are required to make this work: every article must have a (very high) probability of being verified, and there must be a mechanism to update replication packages that is transparent (i.e., the fact that the replication package was replaced should be visible). We posit that a policy that encourages or even regularly organizes post-publication verifications might enhance the effectiveness of the current policy enforcement at numerous journals that currently opt for the “table top” method of verification. Call this the “table top plus” method. We do note that timeliness might suffer - some of the articles in the Fišar et al sample were published four years ago, and were only verified in 2023.

A second issue that is worth highlighting is the **amount of effort and skill required** to properly conduct a reproducible analysis. In our own outreach efforts, we use the term "computational empathy" to emphasize that replication packages should be cognizant of other future replicators’ differences, such as a gap in computing capabilities or knowledge. Fišar et al indirectly address this in their survey of the volunteer replicators, including a question about the replicator’s familiarity/expertise with the methods and software used in the replication package. They then use this information to match replicators with replication packages that align with those skills. And yet: the average level of expertise on method is only about 85% and on topic is only 60% - even though all 900+ replicators were drawn from the 9000+ reviewers that the journal had used in the past 5 years. It would be interesting to compare the congruence of reviewer method and subject matter expertise during the regular refereeing process - in both cases, perfect congruence, but any systematic differences might reveal the difficulty (ex ante) of finding appropriate reviewers.

Somewhat more striking in our view is the fact that, despite the very high skill distribution in the Management Science Reproducibility Collaboration (the modal reviewer has a completed Ph.D. and some professional experience, Table 1), the reproducibility score distribution conditional on data availability (about 43% fully reproduced and 52% reproduced with “minor issues”, Figure 1), is eerily similar to the one we obtained in (Herbert et al. 2023a; 2023b) with a much less skilled reviewer pool (the modal reviewer had not completed the 3rd year of a bachelor's program!). There, we found about 45% fully reproduced and 43% what we called "partially reproduced". This suggests two things:

First, it is not the skill distribution of the *replicators* that is at issue, but the *skill distribution of the original authors* when crafting their replication package. Heuristically, that is also what we observe in the AEA verification checks we conduct, and is what suggested by the findings in (Pérignon et al. 2023), who had a single replicator reproduce 168 papers that tackled the same empirical problem.

But second, it also raises the following question: Is it really acceptable that a highly skilled group of researchers, all equipped with Ph.D., subject matter and technical expertise aligned with the paper to be reproduced, equipped with the same data and code, that this group should still need somewhere **between 8 and 13 hours** to simply reproduce the (small number of) tables and figures in the manuscript? Note that the median (non-reproduction) reviewer time in the social sciences is less than 5 hours (Mulligan, Hall, and Raphael 2013; Publons 2018)! Shouldn't this number be much closer to, say, one hour? After all, the reviewers are no longer trying to painstakingly clean the data, figuring out the best way to implement a particular econometric method, and lining up all the programs to produce the tables and figures in the paper. All that hard work has been done at least once. Shouldn't it be dramatically easier to simply reproduce that analysis?

Both of these observations are important, because they suggest that one path to improving the quality of replication packages can go through improved guidance, better training of new (and old) scholars, and increased outreach. We postulate that such improvement in the (technical) skills of authors across entire fields is not just something desirable from a reproducibility perspective (reducing the cost of future reproductions, re-use, and extensions, a public good benefit), but would also make the entire endeavor of a research project more efficient, providing a private benefit to each and every researcher.

We much appreciated reading this report on an important experiment, from which we learned a lot. In fact, we still have many questions that the authors may have some answers to, and which we hope to find answers to in the (naturally) fully reproducible replication package once it is posted!

References:

American Economic Association. 2019. "Policy on Revisions of Data and Code Deposits in the AEA Data and Code Repository." 2019. https://www.aeaweb.org/journals/data/revisions-policy.

Herbert, Sylverie, Hautahi Kingi, Flavio Stanchi, and Lars Vilhuber. 2023a. "The Reproducibility of Economics Research: A Case Study." Banque de France Working Paper. https://doi.org/10.2139/ssrn.4325149.

———. 2023b. "The Reproducibility of Economics Research: A Case Study." *Submitted*.

Mulligan, Adrian, Louise Hall, and Ellen Raphael. 2013. "Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers." *Journal of the American Society for Information Science and Technology* 64 (1): 132–61. https://doi.org/10.1002/asi.22798.

Pérignon, Christophe, Olivier Akmansoy, Christophe Hurlin, Anna Dreber, Felix Holzmeister, Juergen Huber, Magnus Johannesson, et al. 2023. "Computational Reproducibility in Finance: Evidence from 1,000 Tests." HEC Paris Research Paper. https://doi.org/10.2139/ssrn.4064172.

Publons. 2018. "Publons' Global State Of Peer Review 2018." 0 ed. London, UK: Publons. https://doi.org/10.14322/publons.GSPR2018.