The role of third-party verification in research reproducibility

Christophe Pérignon

July 4, 2023

Research reproducibility is defined as obtaining similar results using the same data and code as the original study. In practice, to check research reproducibility, third-party verification constitutes a useful complement to the work done by journals’ internal teams. Third-party verification services can also be used by individual researchers seeking a presubmission reproducibility certification to signal the reproducible nature of their research. Using the example of the cascad certification agency, which I co-founded in 2019 with Christophe Hurlin, I discuss the functioning, utility, comparative advantages, and challenges of third-party verification services. I thank Olivier Akmansoy, Jean-Edouard Colliard, Christophe Hurlin, Jacques Olivier, and Lars Vilhuber for their comments and support.

## Background

The quest for a reproducible science requires *three preconditions* to be met, and I believe all three are met today in the field of economics.

The first precondition is to have a good understanding of *what research reproducibility is*. Collectively, the survey of Christensen and Miguel (2018), the report of the National Academies of Sciences, Engineering, and Medicine (2019), and the work of the American Economic Association (Vilhuber, 2021) brought some much-needed clarity to the different concepts used to describe reanalyses in economics. Currently, the consensus is increasingly favoring the notion that an empirical result is deemed reproducible if it can be recreated by running the original code of the authors on the original data. This type of tests contrasts from other forms of reanalyses such as replications, robustness analyses or extensions (Vilhuber, 2020).

The second precondition is *to recognize that the current level of reproducibility is low*. Indeed, there is significant evidence that the success rate of reproducibility studies in economics and finance remains surprisingly low, mainly due to missing code/data/information and numerous bugs (Chang and Li, 2017; Gertler et al., 2018; Herbert et al., 2021; Pérignon et al., 2023). Depending on the studies, the success rate ranges between 14 and 52%.

The third precondition is *to acknowledge that this lack of reproducibility is problematic and that we need to act to improve the situation*. Following early decisions by the American Economic Association (Duflo and Hoynes, 2018) and the Royal Economic Society, most of the other leading scientific associations and academic journals are now considering strengthening their code and data availability policies.

Now that these three preconditions have been met, the most important challenge is *implementation*. As of today, reproducibility verification is conducted by dedicated verification teams working for some academic journals or associations or by third parties (e.g., *cascad* and the *Odum Institute*). Either internal or external to a journal, the verifier checks whether the submitted material complies with a set of guidelines and attempts to regenerate all the results from the code and data provided by the authors. While verifications at journals are typically conducted once the manuscript has been conditionally accepted, third-party verifications can be made at any time in the life of a paper. After successful reproductions, third-party verification services typically award reproducibility badges or certificates, which can be added to the manuscript when submitting to an academic journal.

## The advantages of an early third-party reproducibility verification

From the viewpoint of the researcher, I see three main reasons to conduct an early third-party reproducibility verification.

The first one is to *detect as early as possible mistakes or inconsistencies* in the analysis. Indeed, when preparing the materials required to request a verification, the authors regularly identify typos and mistakes and they have the opportunity to correct them at no cost. Differently, when such mistakes are discovered later in the process, and especially after publication, the research community, and in particular, journal editors, will have to decide whether this is an honest mistake or plain misconduct. In the latter case, the stigma in terms of reputation can be very large.

The second reason concerns the *cost* of conducting verification, mainly in terms of time for the researchers themselves. Today, when they target top journals in economics, researchers know that (1) the cost will be faced with probability one as most journals have systematic verification in place and that (2) the cost increases significantly with time, as more datasets, code versions, and forking paths are added to the analysis. While optimal timing will also reflect time preferences, waiting until the paper acceptance to start thinking about reproducibility is unlikely to be an optimal strategy.

The third reason to conduct an early reproducibility verification is to *build trust*, and in particular among coauthors. Indeed, most academic papers have multiple authors, and the latter tend to specialize in exploiting their comparative advantages. Furthermore, some specialized coauthors may not have the time, nor the skills, to monitor and review tasks that are not under their operational control. In this case, a third-party verification provides some reassurance for all the parties involved. However, it is important to acknowledge that a pre-publication reproducibility verification is not an “all-risks insurance”. Indeed, there are many aspects of the research that a third-party verifier does not check, such as the correspondence between the claims and equations in the paper and the content of the code, the presence of typos in the code, or whether the authors engaged in data manipulation or fabrication. These problems can subsequently be identified by other researchers by reviewing the original code and datasets (see the [www.datacolada.org](file:///C:\Users\perignon\AppData\Local\Microsoft\Windows\INetCache\Content.Outlook\FAT02JQU\www.datacolada.org) website for such forensic investigations).

## The cascad certification agency

Christophe Hurlin and I founded *cascad* ([www.cascad.tech](http://www.cascad.tech)) in 2019 with a double objective: (i) to help individual researchers signal the reproducible nature of their research by granting reproducibility certificates and (ii) to help other scientific actors (e.g., academic journals, universities, funding agencies, scientific consortia, data providers) verify the reproducibility of the research they publish, fund, or contribute to the production of.

In terms of organization, *cascad* is a nonprofit research laboratory funded by the French National Center for Scientific Research (CNRS) along with several universities and research institutions. While it operates within France, *cascad* collaborates with researchers, academic journals, and other users from all around the world. Its workforce comprises full-time reproducibility engineers, part-time graduate students, and a group of faculty oversees the operations and promotes the services offered.

The establishment of *cascad* was driven by two firm beliefs. First, we believe that for science to be taken seriously, there needs to be a serious commitment to reproducibility. Put it simply, if you want the chain of science to be strong and useful to society, you do not want reproducibility to be its weakest link. Second, we hold the conviction that merely making code and data publicly accessible does not fully address the reproducibility challenge. We have come to this resolute belief after engaging in several years of management at RunMyCode ([www.runmycode.org](file:///C:\Users\churl\Downloads\www.runmycode.org)), a repository for code and data used by various economics and management journals. In this capacity, we often saw researchers failing to share all the essential components (code, data, explanations) required to regenerate their results. This was frequently due to hurdles such as copyright issues, non-disclosure agreements (NDAs), or concerns related to data privacy. Moreover, even when all components were available, other researchers frequently struggled to execute them, and occasionally failed entirely (for consistent evidence, see Chang and Li, 2017, Gertler et al. 2018, Trisovic, 2022, and Pérignon et al. 2023).

We realized that a third party could be useful in this context. First, when all the required resources can be shared, a third party can run and regenerate all the results before uploading the code and data on an online repository. Second, when some data cannot be shared, the third party can ask permission to access such data, to be able to run the code and reproduce the results (Pérignon et al. 2019). Finally, third-party verifiers can also be useful to academic journals (i) when the third party has permanent access to some restricted data, (ii) when it owns a license of, or expertise in, a software that the journals do not have, or (iii) when the journals do not have enough staff or computing power to verify all the newly accepted papers.

## Examples of collaborations

*Collaborations with economics journals:* Since 2019, *cascad* has provided verification reports to the data editors of the American Economic Association and the Royal Economic Society. Such verifications concern conditionally accepted articles in one of the eleven journals managed by these two associations (e.g. *American Economic Review*, *American Economic Journal: Macroeconomics*, *Economics Journal*). To date, around 60 verifications have been conducted by *cascad* for these journals.

*Collaboration with a restricted data access center:* Since 2020, the *cascad* agency has partnered with the Centre d'Accès Sécurisé aux Données (CASD), a French public research infrastructure that enables researchers to access granular, individual data from the French Institute of Statistics and Economic Studies (INSEE) and from various French public administrations and ministries. In total, CASD hosts data from 378 sources and offers a data provider service to 742 user institutions. The CASD also gives access to restricted access data from the Banque de France, as well as individual health data, and environmental data. This example allows us to illustrate the economy of scale argument introduced earlier. Indeed, Colliard et al. (2023) found 134 articles on Google Scholar using CASD data, published in 91 different academic journals. To verify the reproducibility of all these articles, each journal would have had to go through a lengthy accreditation process to access the same original data. Instead, *cascad* offers a single point of entry to all academic journals seeking a reproducibility check for articles using restricted data accessed through CASD.

*Collaboration with a scientific consortium:* In 2021, *cascad* was tasked with assessing the reproducibility of the empirical results of 168 international research teams, gathered from more than 200 universities, who were participating in the Fincap project (Menkveld et al., 2023). Each team had to answer the same six research questions using the same dataset consisting of 720 million financial transactions. Pérignon et al. (2023) showed that running the original researchers’ code on the same raw data regenerated exactly the same results only 52% of the time. Reproducibility was higher for researchers with better coding skills and for those who exerted more effort. It was lower for more technical research questions, those with more complex code, and for outlier results. Neither researcher seniority, nor peer-review ratings appeared to be related to the level of reproducibility.

**The business model of third-party verifiers**

Launching and operating a third-party reproducibility verification service is costly. Colliard et al. (2023) decomposed the total costs between the fixed costs corresponding to the IT infrastructure (including software) and the variables costs corresponding to labor costs, computing costs, and the costs of accessing data. They showed that exploiting economies of scale could lower the average cost per paper from $763 to $330.

Our experience at *cascad* suggests that in addition to accessing restricted data, the most challenging and time-consuming task is to reconstruct the computing environment used by the original authors (recall that dockers are not yet widespread in economics). Another challenge in practice is to be able to locate the results in the regenerated logfile because a surprisingly large fraction of code still does not automatically generate tables and figures (see Pérignon et al., 2023). These challenges suggest that one way to reduce verification costs is to increase automation in the verification process, raise awareness among researchers, and increase their coding skills.

The question of who should pay for the extra cost associated with reproducibility checks is also key. Should the readers pay, including nonacademic ones? Should the authors pay? Should only those who get their papers accepted or all authors who submit their manuscripts pay? Differently, should the costs be covered by research funding agencies or universities? Obviously, designing a sustainable business model is a prerequisite for third-party verifiers to scale up and operate efficiently.

## Conclusion

We have shown in this paper that third-party verification services are useful actors in the reproducibility ecosystem. They complement the journals’ verification efforts, especially when the research is based on restricted data or requires special skills or computing environments. We have shown that to prosper in the long term, third-party verifiers need to automate their labor-intensive process and clarify their business models.

Third-party verifiers could also be useful to systematically verify empirical findings based on *online experiments*, such as those conducted on Amazon Mechanical Turk ([MTurk](https://www.mturk.com/)) and [Qualtrics](https://www.qualtrics.com/). While this kind of study often relies on pre-analysis plans and shares final datasets on public repositories (e.g. Open Science Foundations), the existence of several forensic investigations (<http://datacolada.org/109>) and subsequent paper retractions suggest that it is more important than ever to allow third-party verifiers to access the raw data collected on the online platforms.

## References

Chang, A. C., & Li, P. (2017) A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review*, 107(5), 60–64.

Christensen, G., Miguel, E. (2018) Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3), 920–980.

Colliard, J.-E., Hurlin, C., and Pérignon, C. (2023) The economics of computational reproducibility, Working Paper, HEC Paris.

Duflo, E., and Hoynes, H. (2018) "Report of the search committee to appoint a data editor for the AEA." *AEA Papers and Proceedings*, 108: 745.DOI: 10.1257/pandp.108.745

Gertler, P., Galiani, S., and Romero, M. (2018) How to make replication the norm. *Nature*, 554 (7693), 417–419.

Herbert, S., Kingi, H., Stanchi, F., & Vilhuber, L. (2021). The reproducibility of economics research: A case study. Banque de France Working Paper Series, WP #853.

Menkveld, A., Dreber, A., Holzmeister, F. Huber, J., Johannesson, M., Kirchler, M., Razen, M., Weitzel U., et al. (forthcoming) Non-standard errors, *Journal of Finance*.

National Academies of Sciences, Engineering, and Medicine. (2019) Reproducibility and replicability in science. The National Academies Press.

Pérignon, C., O. Akmansoy, C. Hurlin, A. Menkveld, Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M. Razen, M., Weitzel U. (2023) Computational reproducibility in finance: Evidence from 1,000 tests. Working Paper, HEC Paris.

Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R. and Debonnel E. (2019) Certify reproducibility with confidential data, *Science*, 365 (2019), 127–128.

Trisovic, A., Lau, M. K., Pasquier T., and Crosas, M. (2022) A large-scale study on research code quality and execution. *Scientific Data*, volume 9, 60.

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

Vilhuber, L. (2021). Report by the AEA data editor. *AEA Papers and Proceedings*, 111, 808–817.