3-2024

Reproducibility in Management Science

Milo ˇs FIŠAR

Ben GREINER

Christoph HUBER

Elena KATOK

Ali I. OZKES

See next page for additional authors

Citation

FIŠAR, Milo ˇs; GREINER, Ben; HUBER, Christoph; KATOK, Elena; OZKES, Ali I.; and CHANG, Hannah H.. Reproducibility in Management Science. (2024). *Management Science*. 70, (3), 1343-1356.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/7441

Author

Miloˇs FIŠAR, Ben GREINER, Christoph HUBER, Elena KATOK, Ali I. OZKES, and Hannah H. CHANG

Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Reproducibility in Management Science

Miloš Fišar, Ben Greiner, Christoph Huber, Elena Katok, Ali I. Ozkes, and the Management Science Reproducibility Collaboration

To cite this article:
Miloš Fišar, Ben Greiner, Christoph Huber, Elena Katok, Ali I. Ozkes, and the Management Science Reproducibility Collaboration (2023) Reproducibility in Management Science. Management Science

Published online in Articles in Advance 22 Dec 2023

. https://doi.org/10.1287/mnsc.2023.03556

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–14
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Reproducibility in *Management Science*

Miloš Fišar,[a] Ben Greiner,[b,c,*] Christoph Huber,[b] Elena Katok,[d] Ali I. Ozkes,[e] and the *Management Science* Reproducibility Collaboration[‡]

[a] Masaryk University Experimental Economics Laboratory, Department of Public Economics, Masaryk University, 60200 Brno, Czech Republic; [b] Wirtschaftsuniversität Wien, Institute for Markets and Strategy, 1020 Vienna, Austria; [c] University of New South Wales, School of Economics, Sydney NSW 2052, Australia; [d] Department of Operations Management, Naveen Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080; [e] SKEMA Business School, GREDEG, Université Côte d'Azur, Campus Grand Paris, 92156 Suresnes, France

*Corresponding author

‡A complete list of the members of the *Management Science* Reproducibility Collaboration is included in Online Appendix A.

Contact: milos.fisar@econ.muni.cz, https://orcid.org/0000-0003-4153-3500 (MF); bgreiner@wu.ac.at, https://orcid.org/0000-0002-4493-8100 (BG); christoph.huber@wu.ac.at, https://orcid.org/0000-0001-5820-571X (CH); ekatok@utdallas.edu, https://orcid.org/0000-0002-7037-7896 (EK); ali.ozkes@skema.edu, https://orcid.org/0000-0002-8720-2494 (AIO)

Received: November 2, 2023
Revised: November 27, 2023
Accepted: November 29, 2023
Published Online in Articles in Advance: December 22, 2023

https://doi.org/10.1287/mnsc.2023.03556

Abstract. With the help of more than 700 reviewers, we assess the reproducibility of nearly 500 articles published in the journal *Management Science* before and after the introduction of a new Data and Code Disclosure policy in 2019. When considering only articles for which data accessibility and hardware and software requirements were not an obstacle for reviewers, the results of more than 95% of articles under the new disclosure policy could be fully or largely computationally reproduced. However, for 29% of articles, at least part of the data set was not accessible to the reviewer. Considering all articles in our sample reduces the share of reproduced articles to 68%. These figures represent a significant increase compared with the period before the introduction of the disclosure policy, where only 12% of articles voluntarily provided replication materials, of which 55% could be (largely) reproduced. Substantial heterogeneity in reproducibility rates across different fields is mainly driven by differences in data set accessibility. Other reasons for unsuccessful reproduction attempts include missing code, unresolvable code errors, weak or missing documentation, and software and hardware requirements and code complexity. Our findings highlight the importance of journal code and data disclosure policies and suggest potential avenues for enhancing their effectiveness.

History: Accepted by David Simchi-Levi, behavioral economics and decision analysis–fast track.
Supplemental Material: The online appendices and data are available at https://doi.org/10.1287/mnsc.2023.03556.

Keywords: reproducibility • replication • crowd science

1. Introduction

To be relevant and credible, scientific results have to be verifiable. The integrity of academic endeavors rests on reproducibility, wherein independent researchers obtain consistent results using the same methodology and data, and replicability, which involves the application of similar procedures to new data.

The significance of these twin principles for scientific research is commonly agreed upon. Yet, recent assessments of empirical studies in the social sciences suggest a concerning rate of non-reproducibility or non-replicability (Ioannidis 2005, Ioannidis and Doucouliagos 2013, Open Science Collaboration 2015). A replicability crisis does not only erode the confidence in individual studies, but casts a shadow over entire fields and literatures, and may potentially compromise business and policy decisions based on these findings. Assessing and addressing these issues is imperative to maintain the credibility of social science research, including management, psychology, economics, sociology, and political science, and its subsequent applications in economic policies and management strategies, guiding societal progress.

Several reasons are cited in the literature as contributing to reduced replicability, such as publication bias (De Long and Lang 1992), undisclosed analysis flexibility (Simmons et al. 2011), *p*-hacking (Brodeur et al. 2016), and plain fraud (List et al. 2001, John et al. 2012). Ensuring that published results can be reliably reproduced is a necessary foundation for addressing these issues. While tackling the underlying reasons for limited replicability may be difficult, the ability to reproduce results based on the original data and analyses can be seen as a minimum criterion for scientific credibility to be expected from all published research (Christensen and Miguel 2018, Nagel 2018, Welch 2019). Indeed, if published results cannot be reproduced because data are unavailable or code used for data or numerical analysis is missing, poorly documented, or error-ridden, then the replicability crisis is partly also a reproducibility crisis.

In this study, we directly assess the reproducibility of results reported in nearly 500 research articles published in *Management Science*, a premier general interest academic journal that comprises 14 departments covering a broad variety of areas in business and management. In 2019, the journal introduced a new Policy for Data and Code Disclosure,[1] which stipulates that "Authors of accepted papers … must provide … the data, programs, and other details of the experiment and computations sufficient to permit replication." While our focus is primarily on assessing the reproducibility of work published since the disclosure policy went into effect, we also analyze articles accepted before May 2019 for comparison.

To reproduce results in articles from a variety of subfields of the journal such as finance, accounting, marketing, operations management, organizations, strategy, and behavioral economics, we use a crowd science approach (Nosek et al. 2012, Uhlmann et al. 2019) to leverage the expertise of many researchers in these different subfields. Overall, 733 volunteers joined the *Management Science* Reproducibility Collaboration as reproducibility reviewers (see Online Appendix A for all names and affiliations), who together reportedly spent more than 6,500 hours on attempting to reproduce the results reported in the articles, using the replication materials and information provided by the article authors.

For articles subject to the 2019 disclosure policy, we find that when the reviewers obtained all necessary data (because they were included, could be accessed elsewhere, or no data were needed) and managed to meet the software and hardware requirements of the analysis, then results in the vast majority of articles (95%) were fully or largely reproduced.[2] However, in approximately 29% of the articles, data sets were unavailable either because they were proprietary or under a non-disclosure agreement (NDA) or because they originated in subscription data services to which reviewers did not have access. If we consider all assessed articles under the disclosure policy, then about 68% could be at least largely reproduced. Because data availability was by far the largest obstacle to reproducing results, the methodology used in an article is strongly correlated to its reproducibility. Namely, computational and simulation studies and online and laboratory experiments are more likely to be reproducible than field experiments, surveys, and other empirical studies. These differences in methodology and data availability are also the main drivers for substantial heterogeneity in reproducibility across the 14 departments of the journal.

Comparing these results to the period before the introduction of the mandatory disclosure policy, we observe a substantial increase in reproducibility. When code and data disclosure was voluntary, only 12% of article authors provided replication materials. Out of these selected articles, 55% could be (largely) reproduced.

The share of fully and largely reproduced results in our study appears high, in particular considering that the code and data editorial team at the journal primarily assesses the completeness of replication materials but does not attempt reproduction of the results themselves. That said, in addition to limited data availability, some replication materials suffered from insufficient documentation, missing code, or errors in the code, making reproduction impossible. For some studies, reviewers obtained different results and were not able to make out the reasons for the discrepancies. This implies that there is still room for improvement. We discuss implications for disclosure policies and procedures at *Management Science* and other journals in Section 4 of this paper.

Our results complement findings in a recent literature on reproducibility and replicability in the social sciences. The definitions of these terms vary somewhat across studies, with some overlaps in their meaning (Christensen and Miguel 2018, Welch 2019, Dreber and Johannesson 2023, Pérignon et al. 2023). "Replication" typically refers to verifying the results of a study using different data sets and different methods, thus exploring the robustness of results. The term "computational reproducibility" comes closest to the scope of our study and is defined as the extent to which results in studies can be reproduced based on the same data and analysis as the original study.[3] Other types of reproducibility may consider recreation of analysis and data or explore robustness to alternative analytical decisions (see also Dreber and Johannesson 2023, for an in-depth discussion).

Recent systematic replication attempts of published results in the social sciences yielded replication rates of 36% in psychology (Open Science Collaboration 2015, $n = 100$), 61% in laboratory experiments in economics (Camerer et al. 2016, $n = 18$), 62% in social science experiments published in *Nature* and *Science* (Camerer et al. 2018, $n = 21$), and 80% in behavioral operations management studies published in *Management Science* (Davis et al. 2023, $n = 10$).

In the field of economics, a number of studies targeting different subfields have set out to evaluate the computational reproducibility of results. The *Journal of Money, Credit and Banking* (*JMCB*) was one of the first journals to introduce a "data availability policy" and one of the first ones to be evaluated. Dewald et al. (1986) assess the first 54 studies subject to the policy. Only eight studies (14.8%) submitted materials that were deemed sufficient to attempt a reproduction, and only four of these studies could be reproduced without major issues. As the authors put it, "inadvertent errors … are a commonplace rather than a rare occurrence" (Dewald et al. 1986, p. 587). McCullough et al. (2006) examine *JMCB* articles published between 1996 and 2002 and successfully reproduce 22.6% of 62 examined works with a code and data archive, and only 7.5% considering all 186 relevant empirical articles in the journal. McCullough et al. (2008) report that for articles published between 1993 and 2003 in the *Federal Reserve Bank of St. Louis*

Review, only 9 of 125 studies (7.2%) with an archive could be successfully reproduced.

One of the top journals in economics, the *American Economic Review*, introduced a data and code availability policy in 2004, and other top journals followed. In examining this policy for studies published between 2006 and 2008, Glandon (2011) reports that, among the studies with sufficient data archives, five of nine studies (55.6%) could be reproduced without major issues. Overall, however, only 20 of 39 sampled studies (51.3%) contained a complete archive, and for 8 studies (20.5%), a reproduction was not feasible without contacting the authors.

More recently, Chang and Li (2017) attempted to reproduce articles in macroeconomics published between 2008 and 2013 across several leading journals and successfully reproduced 22 of 67 studies (32.8%). Gertler et al. (2018) examined the reproducibility of 203 empirical studies published in 2016 that did not contain proprietary or otherwise restricted data, and were able to reproduce 37% of them (but only 14% from the raw data). For 72% of the studies in the sample, code was provided but executed without errors in only 40% of the attempts. Herbert et al. (2023) ask undergraduate economics students to attempt to reproduce 303 studies published in the *American Economic Journal: Applied Economics* between 2009 and 2018. Only 162 studies contained non-confidential and non-proprietary data. For these, 68 reproduction attempts (42.0%) were successful and another 69 (42.6%) were deemed partially successful. Pérignon et al. (2023) leverage a set of 168 replication packages produced in the context of an open science multianalyst study in empirical finance (Menkveld et al. 2023). Of 1,008 hypothesis tests across all materials, 524 (52.0%) were fully reproducible, with another 114 (11.3%) yielding only small differences to the original results.

Reproducibility studies in other related fields show similarly limited reproducibility. For a sample of 24 studies subject to the *Quarterly Journal of Political Science*'s data and code review, Eubank (2016) finds that only 4 (16.7%) did not require any modification in order to reproduce the results. In genetics, Ioannidis et al. (2009) report that only 8 of 18 microarray gene expression analyses (44.4%) were reproducible. An analysis of biomedical randomized controlled trials yields 14 of 37 (37.8%) successfully reproduced studies (Naudet et al. 2018). Artner et al. (2021) attempt to reproduce the main results from 46 published articles in psychology with the underlying data but no code and were successful in 163 of 232 statistical tests (70.3%). Xiong and Cribben (2023) examine the reproducibility of 93 articles using functional magnetic resonance imaging (fMRI) published in prominent statistics journals between 2010 and 2021, of which only 23 (24.7%) included the actual data set and 14 (15.1%) could be fully reproduced.

A comparison of reproducibility rates across different studies is difficult. Different studies often apply different definitions and standards of reproducibility, and reasons for non-reproducibility may differ between different journals due to different policies and enforcement procedures and different methods and data availability conditions in their fields. For example, our share of 95% of (largely) reproduced articles (conditional on data being available to the reviewer and hardware and software requirements being met) appears to be in a similar ballpark as the 85% of at least partially successful reproductions at the *AEJ: Applied Economics*. However, although both journals have similar disclosure policies, in the respective time periods, replication materials of articles at *AEJ:AE* only underwent a cursory review, whereas the code and data editorial team at *Management Science* checked all replication packages for completeness.

In recent years, there have been significant developments in the institutional arrangements for reproducibility of journal articles. For economics, Vlaeminck (2021) report that in a sample of 327 journals, 59% have data availability policies, a significant increase compared with 21% in the year 2014. Similar developments are present in the fields of business and management. For example, several other journals published by INFORMS have adopted similar code and data disclosure policies after *Management Science* took the lead in 2019. At the time of writing this paper, 20 of the 24 journals used for the University of Texas Dallas Business School rankings have a code/data disclosure policy, but only 10 made code/data sharing compulsory, and only 2 have a code and data editor enforcing the policy.[4] Colliard et al. (2023) discuss journals' incentives with respect to reproducibility, and Höffler (2017) provides evidence that, in economics, journals with disclosure policies are more often cited than journals without such policies.

The ability to reproduce results reported in published articles by executing the code on the data, both provided by the authors, does not, by itself, guarantee that results are replicable. However, it does provide a useful baseline. It increases confidence that reported results could, in principle, be replicated. Allowing access to original code and data also makes it possible for independent research teams to scrutinize robustness, conduct their own analysis including meta-analytical work spanning multiple studies and data sets, reuse code in other research, and either build on the results or design studies to show the limitations of original results. The ability to do this promotes scientific discourse, and importantly, also decreases incentives for academic fraud and data falsification.

2. Study Design and Procedures

2.1. Procedures

Prior to 2019, *Management Science* encouraged but did not require the disclosure of data for submitted/accepted manuscripts. In June 2019, a new policy was established, which applied to all newly submitted manuscripts and is still in effect at the time of this writing. The policy requires that all code and data associated with accepted manuscripts at *Management Science* have to be provided before the manuscript goes into production, but it also allows some exceptions, in particular licensed data (Compustat, Center for Research in Security Prices (CRSP), Factset, Wharton Research Data Services (WRDS), etc.), proprietary data, or confidential data under an NDA. In these cases, detailed descriptions of data provenance and data set creation are expected. The journal established the position of a code and data editor (CDE) and consequently positions of code and data associate editors (CDAEs), who review all replication packages for completeness before an article goes into production. However, the CDE and CDAEs are volunteer positions, so there are limits to a complete check of the packages of all accepted articles for reproduction.[5]

Our study, preregistered at the Open Science Framework,[6] attempts to assess the reproducibility of articles published in *Management Science* before and after the introduction of the 2019 policy, based on the materials provided by the authors. For the period after the policy change, our initial sample consists of 447 articles[7] that fell under the disclosure policy introduced in June 2019, had been reviewed by the CDE team through January 2023, and were published (with their compulsory replication package) on the journal's website. As a comparison sample we chose all 334 articles that were accepted at the journal between January 2018 and April 2019 and would have fallen under the disclosure policy (i.e., include code or data) but were accepted before the announcement of the policy and were thus not subject to the policy (which only applied to articles initially submitted after June 1, 2019).[8] Of those 334 articles, for 42 the authors had voluntarily provided a replication package, which entered our project reviews. Thus, the size of our initial sample of replication packages to be reproduced is 489.

On January 12, 2023, the editor-in-chief of *Management Science* wrote an email to all 9,762 reviewers who provided a review to the journal in the past five years, introducing the project and inviting them to serve as reproducibility reviewers (see Online Appendix E.1). In addition, the invitation to participate in the project was sent via professional mailing lists (e.g., Behavioral Economics, Finance, Marketing). In total, 927 researchers completed an initial reviewer survey asking for their research fields (namely, to which *Management Science* departments they would typically submit their manuscripts) and their familiarity with different analysis software/frameworks and databases (see Online Appendix E.2).

The assignment of articles to reviewers proceeded over two main assignment rounds and a consecutive third round. In the first assignment round at the beginning of February 2023, we attempted to find a reviewer for each of the 489 packages out of the 927 reviewers. We applied the Hungarian method (Kuhn 1955) that tries to maximize the match with penalties for mismatches in department, software skills, and database access, and random resolution of ties (see Hornik 2005, for the R implementation). These matches were then manually assessed for potential conflicts of interest (e.g., reviewer and author in the same department), in which case article and reviewer were removed from the match and reentered the "pools" of articles and reviewers. Once the match was completed, all reviewers received an email informing them of their assignment, with links to the article, the supplementary materials page, and to guidelines for reviewers. Reviewers were also asked to either confirm their assignment or to contact us to indicate any conflicts of interests or other reasons that they could not provide a report for the assigned article. These cases were also added back to the pool.

After two weeks, we ran a second assignment round. For articles, the sample consisted of previously unmatched articles (which received priority) and a second set of all articles (to find a second reviewer for many of them). All reviewers with no assignment yet entered the match. We once again used the Hungarian method with moderate penalties for department and software mismatches and prohibitive penalties for assignments of the same article or previous assignments, and random resolution of ties. The resulting match was screened for conflicts of interests. As before, reviewers received their assignment by email, and any reported mismatches or conflicts were tracked. A few dropouts of reviewers were recorded; otherwise, articles and reviewers reentered the "pool." Reviewers who did not confirm their assignment in the first or second round received a reminder email at the end of February.

The third round of assignments, from the beginning of March 2023, was run continuously in several waves and mostly manually. Once a sufficient mass of articles (rejections of assignments, leftover articles who have not received their second assignment yet) and reviewers (unmatched reviewers or reviewers available for another report) was reached, for each article, a list of all possible compatible reviewer matches was compiled, and out of this, one reviewer was assigned. As before, reviewers were informed about their match and asked to confirm their assignment.

Reviewers were asked to make an honest attempt to a reproduction of the article's main results (figures, tables, and other results in the main manuscript) solely based on the provided replication materials (and not to contact the original authors of the articles; see McCullough et al. 2006, for similar approaches) and to provide their report within about five weeks (although we also accepted late entries). Reviewers submitted their report through a structured survey implemented in Qualtrics (see Online Appendix E.3). They also received detailed guidelines (see Online Appendix E.4), providing definitions for different reproducibility assessment outcomes and explanations for all survey fields. The survey asked for an overall assessment, information about the content of the replication package (readme, data, code, etc.) and their quality, individual reproducibility assessment of all results tables and figures and other results reported in the manuscript, as well as assessments of time spent, of their own expertise in research field and analysis methods, and of their expectation of the replicability (as opposed to reproducibility) of the article. Reviewers were also asked to provide evidence of their reproduction attempts in the form of log files or screenshots.

During the whole review period, we answered any questions from reviewers by email. Once a significant number of reviews had been collected, we checked them for completeness and consistency. Where necessary, we followed up with reviewers to clarify questions and resolve inconsistencies.[9] All in all, we followed up on about 13% of all reports.

In late September 2023, we wrote emails to all corresponding authors of the articles for which we obtained reports and provided them with the reports (redacted for anonymity). Authors could submit a short comment of up to 2,000 characters on each report, which was then included in our data set.[10] A total of 115 authors or author teams made use of this possibility and submitted comments.

2.2. Final Sample

In total, we received 753 reports from 675 reviewers and reviewer teams who spent in total more than 6,500 hours on this project.[11] We allowed reviewers to enlist the help of a colleague as a secondary reviewer, so for 61 reports, reviewers are actually a team of two persons. While 599 reviewers provided one report each, 74 reviewers provided reports for two different articles, and 2 reviewers provided reports for three articles.

Table 1 shows that a majority of reviewers are at an intermediate stage in their academic career, at the associate professor, assistant professor, or postdoc level. About one in seven reviewers is a full professor and about the same number are PhD students. In addition, there are reviewers working in other roles at research and professional institutions. Across these career levels, reviewers differ in their frequency of enlisting a secondary reviewer (with full or associate professors being more likely to do so, whereas almost all PhD students worked alone) and the time spent (differences there are mainly driven by whether it was a team or not). However, they do not differ much in their self-assessed expertise in the method or topic of the article. In our analysis below, we also do not find any systematic differences across reviewer characteristics in terms of assessment outcomes or other report characteristics.

Table 2 gives an overview of our final sample of assessed articles. Of the 781 articles, 292 from before the introduction of the 2019 policy had no replication package and are therefore not assessed. For 30 articles with replication packages, we could not find a suitable reviewer and thus cannot report any reproducibility results.[12]

In Table 3, we list the *Management Science* departments where the articles in our final sample appeared.[13] This distribution is representative for articles in the journal, with Finance, Behavioral Economics and Decision Analysis, Accounting, and Operations Management being the largest fields. To facilitate the matching of reviewers and articles, upon registration, we asked reviewers to which department(s) they would most likely send one of their articles. Table 3 shows the distribution of the first-named department. This distribution follows largely the distribution of articles, with the exception that researchers from Behavioral Economics and Decision Analysis contribute disproportionately.[14] During code and data review, the CDE team usually classifies articles into one of five categories according to their main methods. While about one-fifth of the articles in the sample mainly use simulations or computations (and thus often do not rely on data), almost 60% of the articles in our sample are based on empirical data (primary or secondary data sets that do not originate from experiments or surveys), with the remaining articles discussing laboratory or online experiments (15%), field experimental data (4%), or data from surveys (3%).

2.3. Reviewer Consistency and Aggregation

To obtain information on potential variability in reproducibility assessments, we aimed to get not just one but two reports for as many articles/replication packages as possible. We succeeded in obtaining two reproducibility reports for 294 articles. For 59% of these articles, both reviewers chose the exact same overall assessment. When only considering whether a reviewer classified an

Table 1. Reviewer Characteristics

				Average expertise	
N = 675	Share	Enlisted second reviewer	Average hours spent	Method (0–100)	Topic (0–100)
Professor	14%	21%	13.1	84.3	60.8
Associate professor	26%	11%	8.3	83.2	61.5
Assistant professor/postdoc	40%	6%	8.4	84.1	58.7
PhD student	16%	1%	9.0	83.8	59.2
Other	4%	3%	6.1	82.8	52.7

Table 2. Initial and Final Sample of Articles and Reports

	Before 2019 policy	After 2019 policy	Total
Initial sample of articles	334	447	781
Articles with replication package available	42	447	489
Articles with package and report(s)	40	419	459
One report	16	149	165
Two reports	24	270	294

article as at least largely reproducible, or not, then the agreement rate is 86%. For the overall assessment of reproducibility, reviewers seem to mostly differ on whether some minor issues are worth mentioning (in generally reproducible studies), and whether a few results that can be recovered are sufficient to deem a study "Largely reproduced" rather than "Not reproduced." Otherwise, differences may result from whether reviewers obtained access to data sets, managed to run the code in the appropriate software environment, or how much effort they put into the reproduction.[15]

In our analysis presented in the next section, we aggregate assessments at the article level. Specifically, if both reviewers chose the same overall assessment, we select one report randomly. If we have two reports for an article, we select the report with the higher reproducibility assessment. This is based on the expected error structure in assessments. When one reviewer could obtain the data or run the software, but the other reviewer could not, then the former's more informed reproducibility judgement should be at least as positive as the latter's. Similarly, while random reviewer errors in assessing the results may lead to a lower reproducibility classification, it is unlikely that those errors yielded exactly the results also obtained by the original authors. Because reviewers had to document their reproducibility efforts and upload log files or screenshots, it seems unlikely that they would have incentives to overstate an assessment result.

Our approach in using the higher assessment of multiple reviews is in line with other reproducibility studies (e.g., Herbert et al. 2023). At the end of the next section, we discuss the robustness of our results to using other aggregation rules or analyzing the data at the level of individual figures and tables, with detailed results included in Online Appendix C.

3. Results

3.1. Main Results

In addition to individual reproducibility assessments of tables, figures, and other results, we asked reviewers for an overall assessment of their reproduction attempt. The guidelines given to reviewers stated the following assessment classifications:

- An assessment of "Fully reproduced" means that the output of the reproduction analysis shows the exact same results as reported in the article, for all results reported in the main manuscript.
- "Largely reproduced, with minor issues" means that there may be small differences in the reproduction output compared with the results in the original article, but the article's conclusions and learnings stay the same.
- "Largely not reproduced, with major issues" means that there are major differences in the output compared with the results in the article, such that the reproduction results could not be used to support the conclusions of the original article.
- An assessment of "Not reproduced" means that the results from the reproduction cannot support the conclusions drawn in the paper, either because the output is different, or because the results cannot be produced at all because of missing data or non-recoverable code.

Table 3. Fields of Assessed Articles and Reviewers

Management Science department	Abbreviation	Share of articles (N = 489)	Share of reviewers (N = 675)
Finance	FIN	27.4%	24.3%
Behavioral Economics and Decision Analysis	BDE	18.4%	30.1%
Accounting	ACC	12.5%	8.2%
Operations Management	OPM	9.2%	7.1%
Marketing	MKG	5.7%	6.5%
Revenue Management and Market Analytics	RMA	4.7%	0.7%
Information Systems	INS	4.3%	4.0%
Business Strategy	BST	3.3%	4.6%
Healthcare Management	HCM	3.3%	1.9%
Big Data Analytics/Data Science	BDA	3.1%	3.4%
Organizations	ORG	3.1%	3.6%
Entrepreneurship and Innovation	ENI	2.3%	4.0%
Optimization	OPT	1.4%	1.2%
Stochastic Models and Simulations	SMS	1.4%	0.4%

However, equipped with these guidelines, the eventual categorization of the article remains subjective to the reviewer. For all overall assessments of "Largely not reproduced" and "Not reproduced," we reviewed the individual reports to distill the main reasons for limited reproducibility. Consequently, cases where the reviewer was not able to get access to a required data set or could not meet the software and hardware requirements of the analysis were labeled "Not verifiable" and "Largely not verifiable" rather than "Not reproduced" and "Largely not reproduced," respectively.[16]

Based on these classifications, Figure 1 presents our main outcomes. The upper two panels show reproducibility assessments for articles that were subject to the disclosure policy introduced in 2019, whereas the lower two panels pertain to articles that were accepted before that policy. The first panel shows the distribution of assessments conditional on reproducibility being verifiable. Among these articles, 95.3% could be classified as fully reproduced or largely reproduced. However, for 29% of assessed articles, reviewers could not obtain the data set, and in 1% the hardware and software requirements could not be met (e.g., software could not be installed, or the code would run for an untenable amount of time). Also in these cases, reviewers were not able to reproduce the results. The second panel in Figure 1 includes these cases, displaying results for all assessed articles. The share of articles that our reviewers were able to fully or largely reproduce is 67.5%.

The third panel of Figure 1 shows the overall assessments for the 40 articles from the time before the 2019 disclosure policy was introduced, for which replication materials were available. Our reviewers could reproduce or largely reproduce the results of 55% of these articles.[17] In the fourth panel of Figure 1, we include all 332 articles from our sample of articles accepted before the 2019 disclosure policy. Considering those articles that do not voluntarily provide replication materials as not reproducible reduces the share of at least largely reproduced articles to 6.6%.[18]

Results from linear probability models, displayed in Table 4, lend statistical support to the positive change since the introduction of the data and code disclosure policy. In Model 1, we regress whether an article could be at least largely reproduced or not on the policy dummy for all articles in our sample (i.e., we are comparing the second and the fourth panels in Figure 1), indicating that after the introduction of the policy, a randomly chosen article is 61% more likely to be reproduced. In Model 2 we restrict our attention to the sample of articles for which a replication package was provided (i.e., comparing the second and the third panel in Figure 1). In this regression, the coefficient for the policy is positive but statistically not significant ($p = 0.109$). Finally, Model 3 focuses on all articles that are considered verifiable (i.e., comparing the second and the third panel in Figure 1 but without the non-verifiable articles). The policy coefficient indicates that conditional on data being available and hardware and software requirements being met, articles are 19% more likely to be reproducible after the introduction of the disclosure policy.[19]

Figure 1. (Color online) Overall Article Reproducibility Assessments by Policy

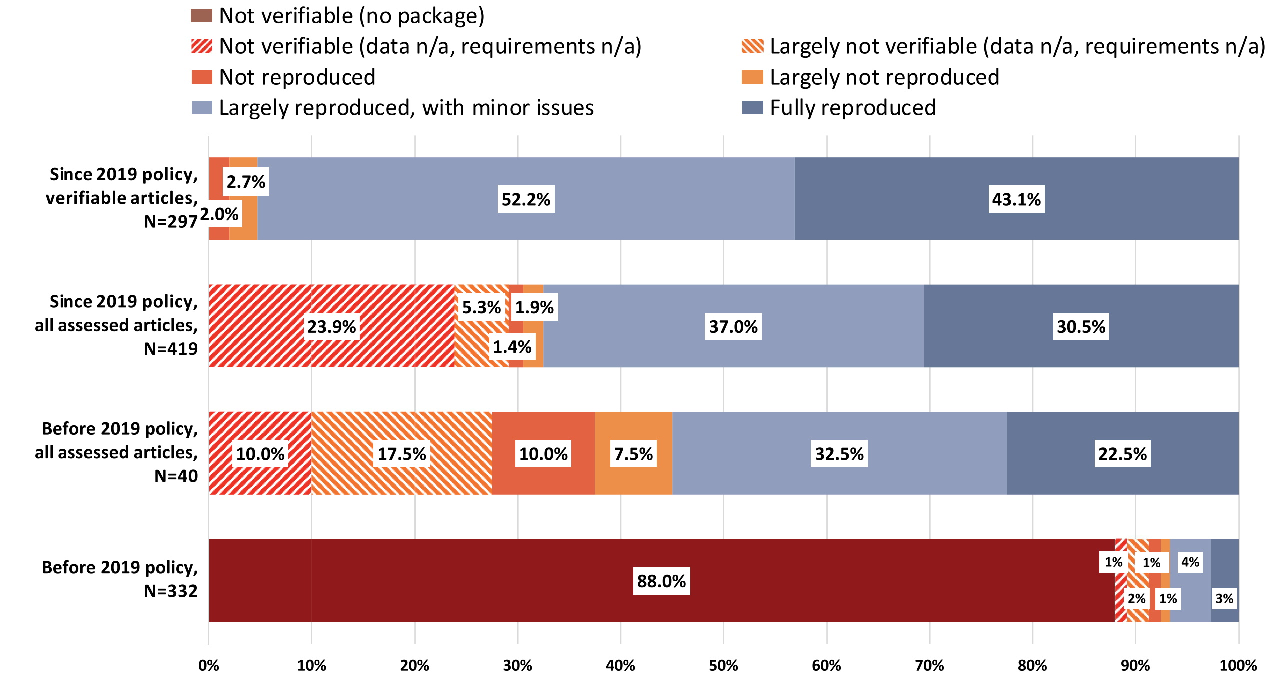


Table 4. Regressing Reproducibility on Disclosure Policy Existence

	(1) All articles, including without package		(2) All articles with package		(3) All verifiable articles	
Sample of articles	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Constant	0.066***	(0.021)	0.550***	(0.075)	0.759***	(0.045)
Disclosure policy	0.609***	(0.028)	0.125	(0.078)	0.194***	(0.047)
Observations	751		459		326	
R^2	0.379		0.006		0.051	

Note. The dependent variable is a binary indicator whether the article was classified as "fully reproduced" or "largely reproduced", or not. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

The unavailability of data is one of the major impediments for reviewers to reproduce an article. A data set may be unavailable, for example, because the reviewer does not have a subscription to the commercial provider, because the data set was collected under NDA with the involved company, or because the data set contains sensitive information (e.g., on personal health or illegal activity). For the sample of 136 reviewed articles falling under the disclosure policy that were classified as either "Not reproduced" or "Largely not reproduced," Figure 2 displays the main reasons we identified for the reviewers' failure to reproduce.[20]

Limited access to the data set was a reproducibility barrier for 88% of non-reproducible articles, and the time needed to run the code, complexity of the code, or issues with installing the software environment were the reasons for non-reproducibility of another 3%. Other reasons included the non-availability of code or functions (13%), insufficient or missing documentation (7%), or unresolvable errors when executing the code (5%). For 4% of the non-reproducible or largely not reproducible articles, the main reason for this assessment was that the reproduction yielded partly different results than reported in the article.[21]

Because many authors cannot include the original data in their replication packages for various reasons, in such cases, the code and data editor at the journal started to encourage the provision of log files that can show that the analysis code works and produces the desired results. Correspondingly, about 52% of the articles classified as "Not verifiable" or "Largely not verifiable" included log files for all results in the replication package, and a further 24% included log files for at least some results. Consequently, 60% of (largely) not verifiable articles were assessed as "Not reproduced but consistent with log files" (84% of those that provided all log files, and 66% of those that provided at least some logs).

3.2. Variation in Reproducibility

Our data allow us to break down the reproducibility of articles published under the disclosure policy to the level of research fields and types of research. Figure 3 shows the reproducibility assessments across the 14 *Management Science* departments. We observe considerable heterogeneity in the share of reproduced or largely reproduced articles across the different fields, ranging from 42% to 100%. However, there are substantial differences in the number of published articles across departments. Also, data availability may vary drastically between different fields.

While many studies in the department Behavioral Economics and Decision Analysis (BDE) rely on primary data from experiments, other fields often use proprietary data from subscription databases (e.g., Compustat,

Figure 2. (Color online) Reasons for Non-reproducibility for Articles Since 2019 Policy

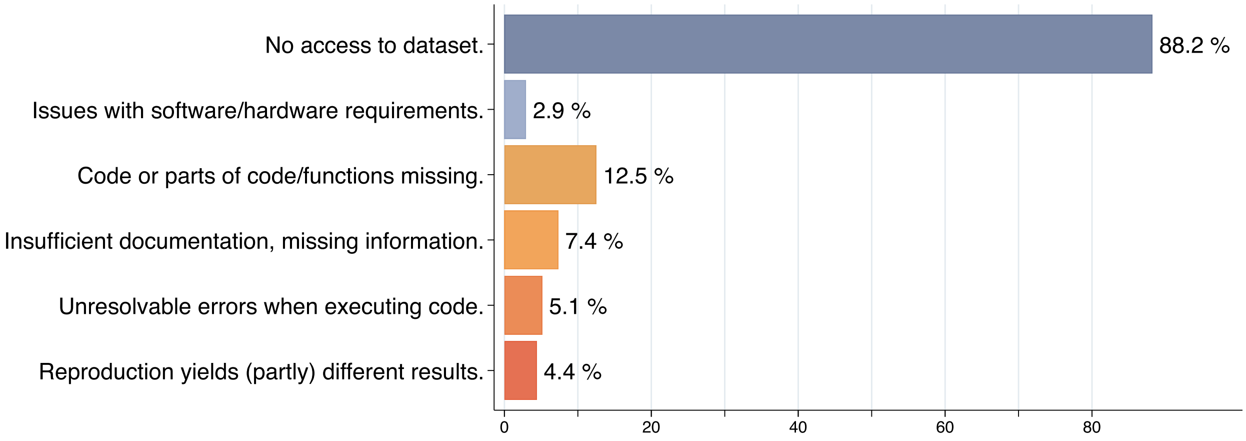
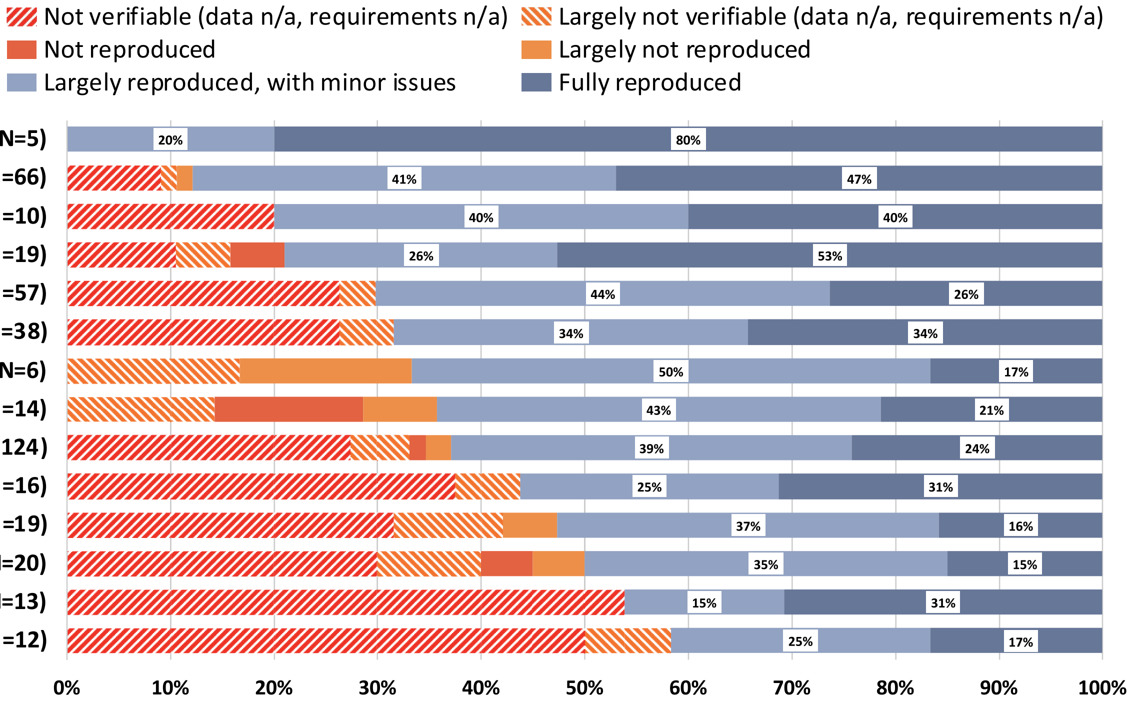


Figure 3. (Color online) Overall Reproducibility Assessments by Journal Department

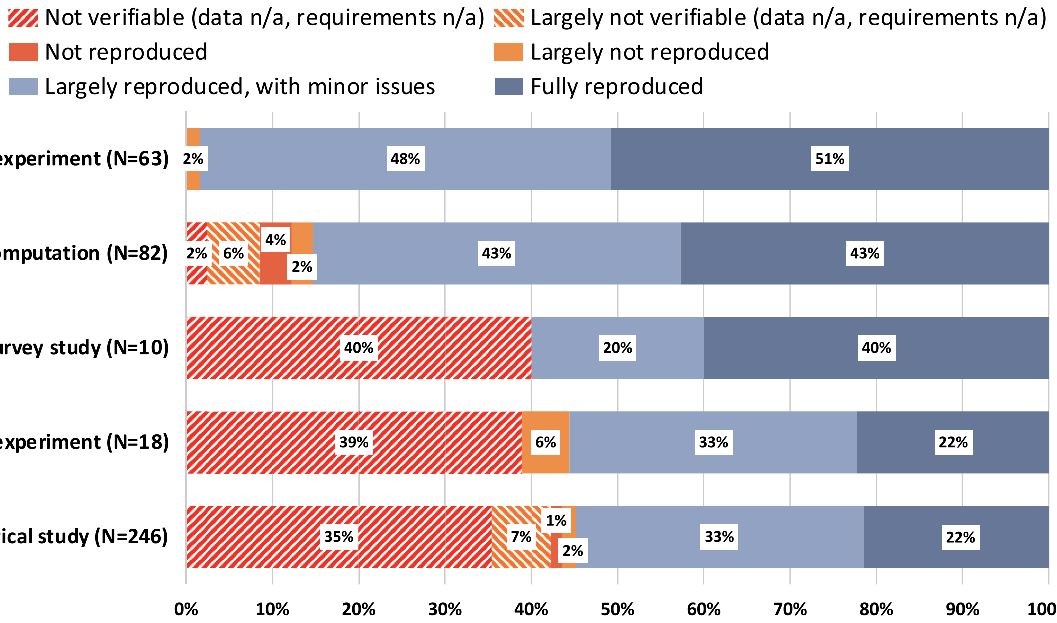


Note. Department acronyms are as follows: SMS, Stochastic Models and Simulations; BDE, Behavioral Economics and Decision Analysis; ENI, Entrepreneurship and Innovation; RMA, Revenue Management and Market Analytics; ACC, Accounting; OPM, Operations Management; OPT, Optimization; BDA, Big Data Analytics/Data Science; FIN, Finance; HCM, Healthcare Management; INS, Information Systems; MKG, Marketing; ORG, Organizations; BST, Business Strategy.

CRSP, WRDS) or confidential and sensitive data that cannot be shared with other researchers (e.g., field experiments with companies, healthcare data, or sensitive surveys). In Figure 4, we distinguish reproducibility outcomes by the primary type/method of the article, as classified during the journal's code and data review. We indeed observe significant differences in the reproducibility outcomes across articles using different methods. All studies reporting on laboratory and online experiments include their data set, making them highly reproducible. Most studies running simulations or other computations, mostly embedded in theoretical articles, do not rely on data sets, making them highly reproducible. Conversely, many empirical studies with primary or secondary data sets rely on proprietary or subscription data, making them less reproducible if reviewers have no access to these data sets. Field experiments in business fields often run under NDAs, and survey studies may include sensitive data that cannot be shared (sometimes even ethics committees restrict the publication of data sets).[22]

In Table 5, we report three linear probability models in which we assess this heterogeneity statistically. The outcome variable in all three models is a dummy indicating whether an article is classified as fully or largely reproduced, or not. In Model 1, we regress reproducibility on department fixed effects, with the baseline being the Finance department (FIN), with a sizable sample size and close to the average reproducibility level. We observe that the SMS and BDE departments have significantly higher reproducibility rates than the Finance department, while the other departments do not differ significantly from Finance. In Model 2, we regress the same outcome on article type fixed effects, with articles based on surveys as the baseline. We find that while field experiments and empirical studies (other than experiments or surveys) do not differ from survey studies in their reproducibility, laboratory/online experiments and articles featuring simulation/computation are significantly more likely to be reproducible. Finally, in Model 3, we include both department and article type fixed effects. The coefficients for article type are not much affected by including department fixed effects, whereas, vice versa, there are some sizable changes. Once accounting for the article type/method used, articles in departments SMS and BDE are not significantly more reproducible anymore compared to other departments, namely Finance. On the other hand, controlling for methods, articles in the Accounting (ACC) department are significantly more reproducible than articles in Finance (more often including the data set), and articles in the field of Big Data Analytics (BDA) are less reproducible (as data sets are often not included or accessible).

Figure 4. (Color online) Overall Reproducibility Assessments by Article Type/Method



3.3. Robustness

In the previous analysis we only considered reproducibility assessments at the article level, taking the higher assessment if two reports were available for an article. To examine the robustness of our results, we also examine the reproducibility for different aggregation rules, at the level of individual reports and at the level of tables, figures, and other results.

In Online Appendix C, Table C.1 reports distributions of overall assessments when choosing the report with

Table 5. Regressing Reproducibility on Journal Department and Article Type

	(1)		(2)		(3)	
	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Constant	0.629***	(0.041)	0.600***	(0.138)	0.630***	(0.146)
SMS	0.371*	(0.209)			0.034	(0.207)
BDE	0.250***	(0.070)			0.019	(0.087)
ENI	0.171	(0.151)			0.215	(0.143)
RMA	0.160	(0.113)			−0.110	(0.118)
ACC	0.073	(0.073)			0.128*	(0.070)
OPM	0.055	(0.085)			−0.049	(0.083)
OPT	0.038	(0.192)			−0.299	(0.191)
BDA	0.014	(0.129)			−0.323**	(0.137)
HCM	−0.067	(0.122)			−0.059	(0.115)
INS	−0.103	(0.113)			−0.073	(0.108)
MKG	−0.129	(0.111)			−0.118	(0.106)
ORG	−0.167	(0.134)			−0.120	(0.127)
BST	−0.212	(0.139)			−0.188	(0.134)
Laboratory/online experiments			0.384**	(0.149)	0.336**	(0.153)
Simulation/computation			0.254*	(0.146)	0.336**	(0.155)
Field experiment			−0.044	(0.172)	−0.009	(0.173)
Empirical study			−0.051	(0.141)	−0.087	(0.143)
Observations	419		419		419	
R^2	0.072		0.140		0.180	

Notes. The dependent variable is a binary indicator whether the article was classified as “fully reproduced” or “largely reproduced” or not. Baseline is the Finance department and survey studies. Department acronyms are as follows: SMS, Stochastic Models and Simulations; BDE, Behavioral Economics and Decision Analysis; ENI, Entrepreneurship and Innovation; RMA, Revenue Management and Market Analytics; ACC, Accounting; OPM, Operations Management; OPT, Optimization; BDA, Big Data Analytics/Data Science; FIN, Finance; HCM, Healthcare Management; INS, Information Systems; MKG, Marketing; ORG, Organizations; BST, Business Strategy.

*, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

the lower assessment whenever there are multiple reports for an article and when randomly selecting one of two reports (with 10,000 repetitions). Since in our previous aggregation we selected the report with the higher reproducibility assessment, these data show somewhat lower reproducibility levels. However, the differences are rather small. For example, compared with the 95.3% (largely or fully) reproduced results for verifiable articles reported earlier, we observe 91.4% when taking the lower assessment of multiple reports, and 93.8% when randomizing which of two assessments is considered.

The regressions reported in Table C.2 are based on all reports rather than just one report per article, clustering standard errors at the article level. Their results mirror the results on policy effects reported in Table 4. Overall, the same reproducibility patterns emerge: The main reason for non-reproducibility is data access, departments differ widely in their reproduction rates, but that is to a large extent driven by different methods being used across departments.

Online Appendix C also reports and discusses the assessment results for individual tables, figures, and other results (e.g., statistical tests reported in the manuscript texts). As to be expected, these individual results are highly correlated with the overall assessments. For example, in reports that reached an overall assessment of “Fully reproduced,” 99.1% of individual tables and 99.7% of individual figures were classified as largely or fully reproduced. When the overall assessment was “Not reproduced,” only 2.7% of tables and 7.5% of figures could be reproduced, on average.

4. Discussion and Conclusion

In this study we undertake a comprehensive assessment of the reproducibility of results in *Management Science*. With the collaborative efforts of more than 700 reviewers, we examine nearly 500 articles to assess the computational reproducibility of their results. For articles published since the introduction of the 2019 disclosure policy, the good news is that more than 95% of articles could be fully or largely computationally reproduced, when data accessibility and hardware/software requirements were not obstacles for reviewers. This appears commendable. However, reviewers faced data accessibility challenges for approximately 29% of the articles in our sample, and the overall rate of successful reproduction is reduced to 68% when considering such articles as non-reproducible. Relatedly, differences in methods and data set accessibility also drive heterogeneity in reproducibility rates across different fields.

This makes data availability a central issue in reproducibility. To improve the credibility of research within business and management, efforts should be directed toward facilitating data access and sharing. Strictly restricting a journal in the area of business, economics, and management to only articles that can freely share their data seems unrealistic and would exclude valuable research from being published. Instead, other arrangements may need to be found for such cases. Approaches could include, among others,

- The inclusion of de-identified data in the replication package, only useful for reproduction but not for new original research;
- Agreements with subscription databases for access for reproduction purposes via the journal;
- Providing access to data sets through special infrastructure that limits use to specific purposes (similar to platforms used by government agencies to provide micro data); or
- Sharing data only with a journal’s code and data editor or with a third-party agency which then certifies reproducibility.

In addition, human subjects ethics committees may need to be sensitized to also consider the ethics of research transparency in their deliberations, to find compromises that at the same time ensure human participant privacy and allow for the full reproduction of research results. Data access limitations also touch on important questions of fairness and bias: With proprietary, non-open data sets, certain research results may only be obtained by privileged researchers, with the data provider serving as a gatekeeper with potential conflicts of interest.

Our study underscores the value of large-scale reproducibility assessment projects. We provide an assessment of the current state of affairs in the field of business and management, and thus contribute to drawing a realistic picture of the overall credibility of research in the field. Repeating such assessments will serve as a form of quality control for newly developed journal policies and procedures. The project showcases best practices and may help developing standards for replication materials but also identifies major gaps and weaknesses in current policies that need to be addressed. Our results can influence journal and funding agency policy decisions. The active participation of more than 700 reviewers who invested significant time and effort in reproducing results highlights the commitment in the community to improving scientific rigor. In an ex post survey, quite a few of our reviewers reported that their participation was a great learning experience, in particular with respect to preparing their own future replication packages. Informed about the assessments of their articles, most authors appreciated the reviewers’ comments, and many voluntarily provided improved versions of their replication packages that address the reviewer comments. Thus, this project also raised awareness of reproducibility issues, furthering a culture of open science, and potentially also the quality of (existing and future) replication materials.

That said, our study also sheds light on the significance of journal code and data review procedures. We observe that the introduction of the 2019 disclosure policy is associated with a significant increase in the reproducibility of articles in *Management Science*. When code and data disclosure was voluntary, only 12% of authors submitted replication materials (out of which 55% could be at least largely reproduced). This suggests that the policy's effect is largely driven by increasing the mere *verifiability* of articles. However, there is still room for significant improvement. Smaller-scale changes could be targeted toward improving the current process, such as increasing incentives for authors to provide proper replication packages right away by making the acceptance decision conditional on replication package approval, or integrating the code and data review process into the manuscript handling system to make it more efficient and transparent.

A more comprehensive reevaluation of code and data review procedures, however, may foster the pivotal role that code and data review plays in ensuring research reproducibility more effectively. In particular, large-scale reproducibility projects such as the present study may become obsolete if the journal puts resources and processes into verifying reproducibility already upon publication of an article. In the current institutional setup, the code and data editor at *Management Science* and his team of associate editors are volunteers with naturally limited capacity to conduct comprehensive reproduction. To that end, different institutional arrangements may be advisable:

- Similar to the institutional setup at the American Economic Association (Vilhuber 2019), code and data review could be professionalized by introducing the position of a (half- or full-time) paid code and data editor, with appropriate budget for assistance and software and data access.
- Code and data review and reproducibility certification could be delegated to a third-party agency that conducts these activities for a fee (such as, for example, the Odum Institute used by the *American Journal of Political Science*, or CASCaD; Pérignon et al. 2019).
- The fact that more than 700 reviewers participated in this project indicates that there is sufficient expertise in the community to integrate the code and data review into the peer review cycle of a manuscript, with low direct costs. For example, in a last minor revision round, one reviewer could be assigned by the department or associate editor to review the replication materials and certify reproducibility. However, while the willingness to participate in this project may have been driven by its novelty, one might have to consider other incentives for reviewers when establishing such reproducibility assessments as a regular procedure.

The scope of code and data policies extends beyond just enabling computational reproduction; their broader aim is to facilitate the replication of research results to assert their robustness and generalizability. Reproducibility does not imply replicability. There may be instances where a study is reproducible but not replicable (e.g., the results can be obtained with the same data set but not with a new data set generated in a different context). Conversely, a study might not be reproducible but replicable (e.g., the original data set may be unavailable so the code cannot be applied, but results with data collected from a different source show the same effects).

We contend, however, that reproducibility serves as a vital foundation for evaluating replicability. A reproducible study boosts confidence in its results, making it meaningful to further examine its robustness and generalizability. The provision of data sets allows for the detection of anomalies and fraud. Materials provided for the reproduction of a study often facilitate its replication as well, by allowing researchers to better understand the structure of data and to apply the same analysis code to new data sets. In addition, to support replication studies, materials required to be provided under most code and data policies extend beyond those purely needed for reproduction. Even if data sets are not available and reproducibility thus not achievable, the packages nevertheless contain detailed descriptions of data provenance and variable dictionaries, aiding replication researchers in gathering new data. For surveys, materials include complete questionnaires or their software implementations, while for experimental studies, they encompass experiment instructions, software code, and other resources critical for running a replication study.

In conclusion, our study illuminates the critical importance of reproducibility in maintaining the integrity and credibility of scientific research in *Management Science* and related fields. By addressing data availability challenges and refining journal code and data review procedures, the academic community can work collaboratively to improve reproducibility. These efforts are essential to ensuring that robust research findings continue to guide decision making and contribute to the advancement of knowledge.

Acknowledgments

The authors thank the members of the *Management Science* Reproducibility Collaboration for their contributions, Matthew D. Houston, Gerlinde Rattasits, and Lucas Unterweger for research support, and Anna Dreber, Susann Fiedler, Lars Vilhuber, two anonymous reviewers, and the associate and department editor for helpful comments. A complete list of the members of the *Management Science* Reproducibility Collaboration is included in Online Appendix A.

Endnotes

[1] Retrieved on August 22, 2023, from https://pubsonline.informs.org/page/mnsc/datapolicy.

[2] We use the term "largely reproduced" when only minor issues were found and the conclusions from the analysis were not affected.

[3] Other scholars refer to computational reproduction also as verification (Clemens 2017), verifiability (Freese and Peterson 2017), or pure replication (Hamermesh 2007, Ankel-Peters et al. 2023).

[4] For comparison, out of the top 25 journals in the 2022 Scimago ranking in Economics and Econometrics, 23 have code/data policies, 17 require that code/data are shared, and 6 have code/data editors. There is some overlap of this set of journals with the University of Texas Dallas list.

[5] If code and data are included, the CDE team also attempts to run the code, but without verifying outputs. As a contrasting example, the American Economic Association (AEA) uses a different model with a paid data editor position including a budget for administrative and research assistants, where all replication packages for all AEA journals are fully reproduced before a final acceptance decision is made.

[6] The preregistration can be found at https://osf.io/mjqg5. Unless otherwise noted, we followed our preregistered procedures.

[7] In our preregistration, we mention 450 articles, but during the review phase we noted that 3 of these articles did not fall under the disclosure policy, reducing the initial sample to 447.

[8] We thus deliberately did not include articles in our study that were accepted after the introduction of the 2019 policy but were not subject to it because they were originally submitted before the introduction. For these articles, their authors could have falsely assumed that the new disclosure policy applies, whereas it did not, thus biasing our assessment of the effect of the policy.

[9] For example, a reviewer may indicate that log files are provided but did not verify whether they are consistent with the results. In other cases, the overall assessment of a replication package may not have been consistent with the individual assessments of tables and figures. Some reviewers could initially not find the replication package because the respective link was missing on the journal's web page, and we provided them with the correct links.

[10] In addition, the journal allows authors to submit an improved replication package, which will replace the previous (reviewed) replication package on the journal's replication server. However, our analysis is only based on the original replication materials.

[11] Two reviewers entered unrealistically high numbers of more than 160 hours (four working weeks); we set these observations to "missing" in our data set. The median reviewer spent four hours.

[12] These 30 articles are not part of the analysis. We observe little evidence of selection issues. Table B.1 in Online Appendix B compares the software requirements of the 30 articles without a report and the 459 articles with at least one report. It seems that articles where we could not find a suitable reviewer were less likely to use the most common software Stata and more likely to use one of the less often used software. However, these differences are statistically not significant at the 5% level (Fisher exact test, two-sided, on the frequency of Stata and frequency of "Other" software).

[13] There have been some changes in the structure of departments at the journal over the past years. In case departments were changed or merged, we classified articles by the current (successor) department.

[14] One reason for this might be a higher awareness for the issues of reproducibility and replicability in this field. Another reason could be that most of the primary authors of this reproducibility study come from this research area.

[15] In Online Appendix D, we provide more details on variability in reviewer assessments.

[16] This qualification of assessments was not yet anticipated in our preregistration.

[17] However, these 40 of 332 articles are heavily selected: authors voluntarily provided a replication package while being encouraged but not required by the journal. More than 50% of these articles were published in the BDE department, and none of them belonged to the finance department, indicating selection also on availability of data.

[18] One may argue that when replication materials are not voluntarily provided to the journal, they may still be hosted on authors' personal websites or in other archives. For a random sample of 50 of 292 articles without replication package, we searched all author websites and repositories for replication materials, and we found none.

[19] We obtain the same conclusions using corresponding probit/logit models or Fisher exact tests. Strictly speaking, our data does not allow to imply a causal effect of the disclosure policy. Authors' attitudes toward making their research reproducible may have independently changed over time, just as the intensity of policy enforcement at the journal may have varied. Older replication packages may be less reproducible due to software changes. The introduction of the policy does not have features of a natural experiment, and our sample only spans a relatively short (and interrupted, see Endnote 8) time period.

[20] Multiple issues may apply to the same article.

[21] In Table B.2 in Online Appendix B we contrast these numbers with the reasons for non-reproducibility for articles that voluntarily provided replication packages before the 2019 disclosure policy took effect. Although the sample size for this period is low ($N = 18$), it appears that reasons for non-reproducibility of voluntarily provided packages are less likely to be missing data and more likely to be issues with missing or non-working code. Reproducibility for older materials may also be affected by limited backward compatibility of statistical software, sometimes producing different results. The reviewers in our study did not report such issues, but they may be more relevant when comparing more distant time frames.

[22] Table B.3 in Online Appendix B demonstrates the variation of paper types/methods across the different departments of the journal. In the table, we ordered departments and methods by their reproducibility to highlight the correlation.

References

Ankel-Peters J, Fiala N, Neubauer F (2023) Do economists replicate? *J. Econom. Behav. Organ.* 212:219–232.

Artner R, Verliefde T, Steegen S, Gomes S, Traets F, Tuerlinckx F, Vanpaemel W (2021) The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psych. Methods* 26(5):527–546.

Brodeur A, Lé M, Sangnier M, Zylberberg Y (2016) Star wars: The empirics strike back. *Amer. Econom. J. Appl. Econom.* 8(1):1–32.

Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, Kirchler M, et al. (2016) Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–1436.

Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, Kirchler M, et al. (2018) Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behav.* 2(9):637–644.

Chang AC, Li P (2017) A preanalysis plan to replicate sixty economics research papers that worked half of the time. *Amer. Econom. Rev.* 107(5):60–64.

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

Clemens MA (2017) The meaning of failed replications: A review and proposal. *J. Econom. Survey* 31(1):326–342.

Colliard J-E, Hurlin C, Pérignon C (2023) The economics of computational reproducibility. Research Paper No. FIN-2019-1345, HEC Paris, Paris.

Davis AM, Flicker B, Hyndman KB, Katok E, Keppler S, Leider S, Long X, et al. (2023) A replication study of operations management experiments in *Management Science*. *Management Sci.* 69(9):4977–4991.

De Long JB, Lang K (1992) Are all economic hypotheses false? *J. Political Econom.* 100(6):1257–1272.

Dewald WG, Thursby JG, Anderson RG (1986) Replication in empirical economics: The journal of money, credit and banking project. *Amer. Econom. Rev.* 76(4):587–603.

Dreber A, Johannesson M (2023) A framework for evaluating reproducibility and replicability in economics. Working paper, Stockholm School of Economics, Stockholm.

Eubank N (2016) Lessons from a decade of replications at the quarterly journal of political science. *PS Political Sci. Politics* 49(2):273–276.

Freese J, Peterson D (2017) Replication in social science. *Annu. Rev. Sociol.* 43:147–165.

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

Glandon PJ (2011) Appendix to the report of the editor: Report on the american economic review data availability compliance project. *Amer. Econom. Rev.* 101(3):695–699.

Hamermesh DS (2007) Replication in economics. *Canadian J. Econom.* 40(3):715–733.

Herbert S, Kingi H, Stanchi F, Vilhuber L (2023) The reproducibility of economics research: A case study. Working paper, Banque de France, Paris.

Höffler JH (2017) Replication and economics journal policies. *Amer. Econom. Rev.* 107(5):52–55.

Hornik K (2005) A clue for cluster ensembles. *J. Statist. Software* 14:1–25.

Ioannidis JP (2005) Why most published research findings are false. *PLoS Medicine* 2(8):e124.

Ioannidis JP, Doucouliagos C (2013) What's to know about the credibility of empirical economics? *J. Econom. Survey* 27(5):997–1004.

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, et al. (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics* 41(2):149–155.

John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psych. Sci.* 23(5):524–532.

Kuhn HW (1955) The Hungarian method for the assignment problem. *Naval Res. Logist. Quart.* 2:83–97.

List JA, Bailey CD, Euzent PJ, Martin TL (2001) Academic economists behaving badly? A survey on three areas of unethical behavior. *Econom. Inquiry* 39(1):162–170.

McCullough BD, McGeary KA, Harrison TD (2006) Lessons from the JMCB archive. *J. Money Credit Bank.* 38(4):1093–1107.

McCullough BD, McGeary KA, Harrison TD (2008) Do economics journal archives promote replicable research? *Canadian J. Econom.* 41(4):1406–1420.

Menkveld AJ, Dreber A, Holzmeister F, Huber J, Johannesson M, Kirchler M, Neusüss S, et al. (2023) Non-standard errors. *J. Finance.* Forthcoming.

Nagel S (2018) Code-sharing policy: Update. *Journal of Finance Editor Blog* (March 6), https://voices.uchicago.edu/jfeditor/2018/03/06/code-sharing-policy-update.

Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, Ioannidis JP (2018) Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in the BMJ and PLOS Medicine. *BMJ* 218:360.

Nosek BA, Spies JR, Motyl M (2012) Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspective Psych. Sci.* 7(6):615–631.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716.

Pérignon C, Gadouche K, Hurlin C, Silberman R, Debonnel E (2019) Certify reproducibility with confidential data. *Science* 365(6449):127–128.

Pérignon C, Akmansoy O, Hurlin C, Dreber A, Holzmeister F, Huber J, Johannesson M, et al. (2023) Computational reproducibility in finance: Evidence from 1,000 tests. Working paper, HEC Paris Research Paper, Paris.

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psych. Sci.* 22(11):1359–1366.

Uhlmann EL, Ebersole CR, Chartier CR, Errington TM, Kidwell MC, Lai CK, McCarthy RJ, et al. (2019) Scientific utopia III: Crowdsourcing science. *Perspective Psych. Sci.* 14(5):711–733.

Vilhuber L (2019) Report by the AEA data editor. *Amer. Econom. Rev.* 109:718–729.

Vlaeminck S (2021) Dawning of a new age? Economics journals' data policies on the test bench. *LIBER Quart. J. Assoc. Eur. Res. Libraries* 31(1):1–29.

Welch I (2019) Reproducing, extending, updating, replicating, reexamining, and reconciling. *Critical Finance Rev.* 8(1–2):301–304.

Xiong X, Cribben I (2023) The state of play of reproducibility in statistics: An empirical analysis. *Amer. Statist.* 77(2):115–126.